Geometric Correction of Historical Arabic Documents

ALI ERHOUMA DULLA

School of Computing, Science and Engineering

University of Salford, Salford, UK

Submitted in Partial Fulfilment of the Requirements of the

Degree of Doctor of Philosophy

2019

TABLE OF CONTENTS

LIST OF FIGURES

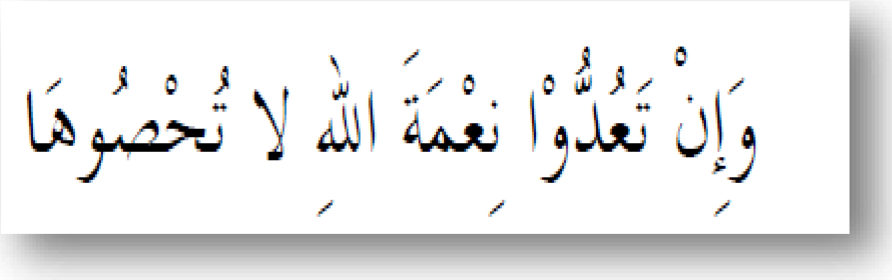
LIST OF TABLES

LIST OF ABBREVIATIONS

ACR	Arabic Character Recognition
CT	Computed tomography
CNN	Convolutional neural network
DT	Delaunay triangulation
GCS	General cylindrical surface
GT	Ground truth
OCR	Optical character recognition
PAWs	Piece of Arabic word
RMSE	Root Mean Square Error
HT	Hough Transform
HAH	Historical Arabic handwritten
TCs	Touching components
VD	Voronoi Diagram
PCA	Principal Components Analysis

ACKNOWLEDGEMENTS

All the praises and thanks to Allah for his guidance and blessings and for granting me knowledge, patience and perseverance to accomplish this work successfully.



وَإِنْ تَعُدُّوْا نِعْمَةَ اللهِ لاَ تُحْصُوهَا

And if you would count the graces of Allah, never could you be able to count them.

Dedicated to the memory of my mother Afea, who always believed in my ability to be successful in the academic arena. You are gone but your belief in me has made this journey possible.

This thesis could not have done without assistance and guidance of my main supervisor Prof. Apostolos Antonacopoulos, I thank him for his continuous support during the PhD programme. I would like to extend a huge, warm thanks to my friends, Dr. Jabril Ramadan, Dr. Youssef Boulid Engr. Abdullahi Ahmadu and I would like to acknowledge and thank everyone who helped me with good advice during my research work.

Conclusively, my deepest appreciation and very profound gratitude go to my father, my wife, my daughter, my sons, my brothers, sisters and my friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

LIST OF PUBLICATIONS

External Publications

Dulla, A. (2018), 'Data Gathering and Analysis for Arabic Historical Documents', World Academy of Science, Engineering and Technology, International Science Index, Computer and Information Engineering, 12(5), 2763.

DULLA, A. 2018. A dataset of Warped Historical Arabic Documents 9th International Conference on Pattern Recognition Systems ICPRS-2018 22-24 May 2018, Valparaiso, Chile.

Dulla, A. and Antonacopoulos, A., 2019. A Novel Baseline Estimation Technique for Geometric Correction of Historical Arabic Documents Based on Voronoi Diagrams, 10th International Conference on Pattern Recognition Systems ICPRS-2019 08-10 July2019, Tours, France.

Internal Publications

Dulla, A. and Antonacopoulos, A. (2017). Geometric Correction of Historical Arabic Documents. Second Annual Postgraduate Research Symposium of the School of Computing, Science and Engineering (CSE-PGR Sym 2017), Salford University, UK.

ABSTRACT

Geometric deformations in historical documents significantly influence the success of both Optical Character Recognition (OCR) techniques and human readability. These deformations may have been introduced at any time during the life cycle of a document, from when it was first printed to the time it was digitised by an imaging device. This thesis focuses on the challenging domain of geometric correction of Arabic historical documents, where background research has highlighted that existing approaches for geometric correction of Latin-script historical documents are not sensitive to the characteristics of text in Arabic documents and therefore cannot be applied successfully. Text line segmentation and baseline detection algorithms have been investigated to propose a new more suitable one for warped Arabic historical document images. Advanced ideas for performing dewarping and geometric restoration on historical Arabic documents, as dictated by the specific characteristics of the problem have been implemented.

In addition to developing an algorithm to detect accurate baselines of historical printed Arabic documents, the research also contributes a new dataset consisting of historical Arabic documents with different degrees of warping severity.

Overall, a new dewarping system, the first for historical Arabic documents, has been developed taking into account both global and local features of the text image and the patterns of the smooth distortion between text lines. By using the results of the proposed line segmentation and baseline detection methods, it can cope with a variety of distortions, such as page curl, arbitrary warping and fold.

Chapter 1 - Introduction

1.1 Overview

This research focuses on the rectification of arbitrary geometric artefacts in Arabic historical manuscripts that appear through the life period of a document, from when it was first printed to the time it is digitised by an imaging instrument.

1.2 Motivation of the study

Along with the growth of digital information, the requirement for digitising historical paper manuscripts has become necessary for most archives and museums; however, there is increasing anxiety over the execution of OCR techniques, which relies heavily on the quality of the manuscript images that are frequently determined by different geometrical distortions (Yang et al., 2011). The value of digitised documents is directly related to the quality of the obtained text. Main archives of historical documents are being collected around the world, demanding an exact reproduction of the text for the automated production of describing metadata, full-text searching, and information extraction (Lund, 2014). In the course of this digitisation, undesired deformation may exist in the final images, such as page curl, arbitrary warping and folds. All these deformations will be problematic for commercial OCR engines that are designed for flat pages with straight text lines. Also, even modest warping can be the reason that most modern OCR techniques fail (Rahnemoonfar, 2010).Arabic Character Recognition (ACR) applications available in the market are still far from being typical (Kanungo et al., 1999).There are numerous reasons for this insufficiency, among them are the lack of standardization, i.e., the availability of sufficient Arabic databases, electronic lexicons, language corpus, and natural language processing (NLP) as well as the unavailability of well-established benchmark test transactions.

One of the standard aims of dewarping is the improvement of the result of OCR methods on distorted document pages (Pugliese et al., 2014). The primary motivation for this research is to present a robust, accurate and adaptable dewarping technique, which focuses on the challenging domain of Arabic printed historical document images containing sheet curl, arbitrary warping and fold.

1.3 Problem of study

There are many challenges in geometric correction method for historical Arabic documents. The tasks of offline systems are considered harder than online ones, where the sequences of points and writing traces are measured. According to (Jamal, 2015), the offline recognition systems can be classified into three classes: printed, historical and handwritten. Printed-related methods have achieved great accuracy while a number of the historical documents also show a good performance. The difficulties involved in historical-related systems are mainly based on one main issue, that being the different types of distortions that can occur at several stages of a document life cycle, namely printing, storage and usage. One of these distortions is arbitrary warping, which is indicated by wavy curves in the document image. Warping is an essential process that makes manuscript images tough to be recognized, and it is crucial to restore warped manuscript images before recognition. However, the systems that deal with handwritten documents like recognition system and word spotting are more challenging because of the writing styles' variety and there is a need to perform many pre-processing tasks to improve the accuracy of such systems. Arabic document analysis is a multi-phase procedure comprising of document image clean up, content segmentation, recognition, and correction. However, there are many potential problems during the stages of this process when trying to use available OCR software. In order to assess the warping trend, it is important to get baselines in a document image and use line segmentation results to fit the baseline of each line (Zhang et al., 2008), but the cursive nature of Arabic writing, a scarcity of Arabic text databases and dictionaries, further demonstrating the sort of challenges facing Arabic document analysis and recognition. In addition, there may be errors made by the recognition system such as incorrectly segmented characters, which causes mis-ranking of characters, and in turn may lead to wrong results (Alginahi, 2013). One solution to these issues is the use of baselines. The baseline is defined as an imaginary line, and characters of cursive or semi cursive script's languages touch the line, and the baseline detection is the most controlling step in pre-processing that directly impacts the efficiency and reliability of the subsequent phases. Line segmentation and baseline detection are therefore used for realising a better performance of cursive script's OCR system (Naz et al., 2013).Despite helping the process, there may also be a problem in accuracy of the detected baseline if there are long Arabic printed words and if there is warping inside the word, which creates difficulties in determining the accurate baseline, unlike Latin texts where it is easy to separate the characters.

1.4 Aim and objectives

The aim of this study is to address the problem of geometric correction of Arabic script documents by focusing on the challenging field of the Arabic historical document images. The detailed objectives of this study are:

1. To study and evaluate the approaches for geometric correction of historical documents to find out the latest updates and developments that have been achieved by previous research in this study area.
2. To investigate text line segmentation algorithms and applying the selected text line segmentation algorithms and baseline detection to determine the most suitable one for Arabic printed historical document images.
3. To develop the proposed technique and make it independent from hardware infrastructure using a variety of scanners and cameras.
4. To implement and test new and advanced ideas for performing dewarping and geometric restoration on historical non-Latin documents, as dictated by the specific characteristics of the problem.
5. To evaluate the robustness of the approach and the developed system by using historical documents that have extremely complex layouts.

1.5 Research Questions

The main question of the study is: How to address the problem of the geometric corrections of the Arabic documents by focusing on the field of a challenge to the documented historical Arab images and to what extent this is possible?

Through the main question of the study, a number of sub-questions can be derived:

1. What are the most important approaches that can be used for geometric correction of historical documents to find out the latest updates and developments in such studies?
2. How to investigate text line segmentation algorithms and baseline detection to determine the most suitable one for Arabic printed historical document images.

3. What techniques are proposed and how they can be made independent of the hardware infrastructure using a variety of scanners and cameras?
4. What are the new and advanced ideas that can be used for dewarping and geometric restoration on historical Arabic documents, as dictated by the specific characteristics of the problem?
5. How to evaluate the robustness of the approach and the system developed by using historical documents, which are extremely difficult due to the complex layouts. For instance, the text could be very intensive. Moreover, in historical documents various font sizes often propagate; also, some ornamental characters, drop cap letters or ornamental borders are present?

1.6 Method of study

The conclusive goal of this research is to find a solution to the research problem described earlier, which can be summarized in different types of distortions that can occur at several stages of a document lifecycle namely printing, storage and usage. The research methodology will include four main stages which are: a theoretical study, measuring the performance of the Geometric Correction methods, and building a system for dewarping as well as comparison, and analysis.

1.6.1 Theoretical Study

The performance of correction of arbitrary warping and page curl methods involves text line segmentation and baseline detection as one of its important pre-processing steps to show the role and importance of the dewarping process in OCR.

1.6.2 Measuring the performance of the geometric correction methods

Choosing an effective Geometric Correction Method for historical Arabic documents is a crucial step to take when looking at dewarping systems due to its influence on the system's output. When the designed or selected Geometric Correction algorithm is ineffective, the system will be negatively affected. Therefore, the challenges that need to be considered when designing or selecting geometric Correction methods for historical

Arabic documents, in particular, are extensively studied in Chapter3. As well as, procedures for measuring the performance of the standards are proposed and carried out.

1.6.3 Performance comparison and discussion

Supervised evaluation techniques have been selected as benchmarking methods to compare with the proposed method. The selected evaluation technique procedures and reasons for selection are presented in Chapter 6.

The research will conclude with the result verification process where testing and evaluation are carried out to evaluate the de-warping system. The testing has been conducted on a variety of texts including printed and handwritten Arabic script with different fonts and sizes. The comparison is done based on the criteria that are shown in Chapter 7.

1.6.4 The proposed system for dewarping historical Arabic documents

In this study, the research will focus on developing a dewarping system. Generally, the proposed system for improving dewarping documents with arbitrary distortions involves five stages which are: pre-processing, text line segmentation, baseline detection, dewarping and evaluation. The robust, flexible and accurate dewarping method which can deal with the variety of geometric distortions including page curl, arbitrary warping and folds in historical Arabic documents are presented in chapter 5. The goal of this chapter is to design a dewarping system with the aim of fulfilling the objectives laid out in chapter 1 and the desirable characteristic of dewarping method described in chapter 2.

1.7 Contributions of the study

The overall contribution of this research is a detailed analysis of the problem of geometric correction of historical Arabic script documents and the proposal of a new method for the dewarping of historical Arabic printed documents.

The primary contribution includes:

1. A newly created dataset consisting of historical Arabic documents with different amount of warping. One of the important advantages of this dataset is the significant amount of detailed and high-quality ground truth available and the

scope for its use. By warping text lines close to the ones above and text lines, as well as overlap area.

2. An algorithm to detect accurate base lines in historical printed Arabic documents and the results of this algorithm will then form the basis for dewarping the document.

The secondary contribution includes:

1. An evaluation of existing dewarping algorithms for non-Latin historical documents.
2. Identifying the best text line segmentation algorithm that can be used for historical Arabic printed document.
3. Identifying the best effective grid-based method presented to geometrically model and arbitrarily correct warped historical documents with relatively complex layout (multi column with graphics).
4. Improving on the existing text methods to dewarp historical Arabic printed documents.

1.8 Outline of the Thesis

The proposed study consists of eight chapters; this is based on the geometric correction of documents and their contribution in the corrections of the historical Arabic documents. Where this chapter (introduction), aims to present the main themes in this study, contextualise the study and introduces the study problem, main aim, objectives and structure of the thesis.

- **Chapter 2: Background**

This chapter looks to elaborate on the major topics that are linked to this study that includes history and development of the Arabic script and describes document image analysis. In addition, this chapter will discuss geometric distortion in historical documents and characteristics of Arabic and Latin texts is discussed. Finally, issue features, image form retrieval and geometric transformation algorithms are presented.

- **Chapter 3: Literature Review**

This chapter illustrates Image Warping and De-warping. Then, it reviews Geometric rectification techniques in the document image, excluding its types and applications. It also examines previous studies for Text Line Segmentation Methods. It lists the best methods, challenges and limitations. Finally, this chapter surveys Baseline detecting approaches, including its techniques used and applications. It focuses difficulties and constraints in terms of a discussion of Baseline detecting techniques.

- **Chapter 4: Research Methodology and Tools**

In this chapter the methodology of the research presented. It introduces and describes the contents, sources, and purposes of the dataset used in this research. Furthermore, it presents and demonstrates the methodology states that were followed to perform this research.

- **Chapter 5: The System for dewarping Historical Arabic Documents**

This chapter reviews and clarifies the proposed dewarping system, including pre-processing methods, the text line segmentation methods and baseline detection, and it present the results of each step.

- **Chapter 6: Evaluation methodology**

This chapter deals with an evaluation methodology for dewarping in the case of historical Arabic documents. It explains the theory of the suggested strategy in terms of its steps and the results of each step.

- **Chapter 7: Experimental Results and Discussion**

This chapter aims to show the experimental results of the proposed techniques. It presents an evaluation and comparative analysis of the performances of the proposed methods and state of the art and commercial software.

- **Chapter 8: Conclusion and future work**

This chapter grants the summary of this theses. It examines the works that have been carried out. Moreover, it is reassessing the aims and includes some proposed suggestion for future works.

1.9 Summary

This chapter focused on the main aim of the study “to develop a robust dewarping technique” which is capable to deal with the combination of geometric distortions including page curl, arbitrary warping and folds in historical Arabic documents. In this chapter, some main topics that are pertinent to the study have been explained. In addition, this chapter focused in brief on the general models that the researcher will adopt in this study to suit the needs of the current research.

Chapter 2 - Background on Geometric Restoration

2.1 Overview

The previous chapter presented the background of the Thesis. This chapter will begin by describing the effect of geometric correction in historical Arabic documents and the challenges associated with historical documents. In the next part of this chapter, the general approach of page shape restoration will be reviewed.

2.2 History and development of the Arabic Script.

Generally, there are two schools of thought concerning the root of the existing Arabic writing. One considers it is Nabataean (of Petra in Jordan), while the other assert is Syriac. The Syriac, similar the Nabataean, is another branch of the Aramaean script, which developed from the Phoenician. The Nabataean kingdom, on the other hand, which flourished between 320 b.c. and a.d. 106, was a powerful Arab trade nation that ruled over present-day Jordan, southern Syria, the Negev, and the Sinai. Before the appearance of Islam, Arabic script was in use during the sixth century, within the Arab kingdom of the Syria-Mesopotamia region (mainly in Al-Hira) yet as in Mecca. These sources recommend that the Arab script technique was “obtained from the Syriac and had produced in Iraq or Midian (Mousa, 2001). Arab historians believe that the primary step in the evolution of Arabic hand was the antique Egyptian hand (see Figure 2-1). Then, Finiqi was produced from Egyptian (Abuhaiba, 2003)

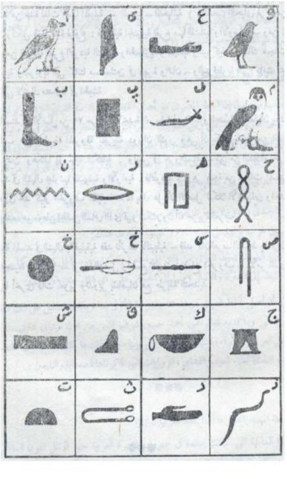


Figure 2-1 Symbols of ancient Egyptian script

Arabic script was without dots and diacritics and when the Holy Quran was written, in Arabic, for the first time, it was without dots and diacritics. Also, it was allowed to continue the same word on the following line. This regulation is permitted in new Latin languages but not in modern Arabic writing (Abuhaiba (2003)**.**

Although Arabic writing existing before the Prophet received Allah's words, it was surely the growth of Islam that served as the catalyst for the increased use of Arabic script. The messenger (peace be upon him) assigned some Muslims to write the Quran directly after the inspiration in Mecca and later in Medina. During the Othman caliphate, these words were collected into the Quran, or holy book of Islam. This meant that with the revelation of Islam, Arabic became an inspirational force. Despite this, during centuries of Arabic prominence, many older writing systems did not simply disappear. Their use, however, did somewhat diminish and although many have survived, they have had to adapt to Arabic in some way. Thus, the Arabic script, since it developed in the seventh century A.D., is now employed in the writing of around fifteen entirely different languages worldwide (Abulhab, 2007). As aforementioned, the Holy Quran has certainly contributed to the proliferation of Arabic script and has established a brand-new era in the previously Arabian unsettled society.

During that time, there have been two forms of scripts: the regular scripts that had been utilized in recording people's everyday life wants within the new Islamic state in Medina; and also the one utilised in writing the Prophet's letters to the kings and emperors of the Roman and Persian kingdoms (see Figure 2-2) (Alshahrani, 2008).

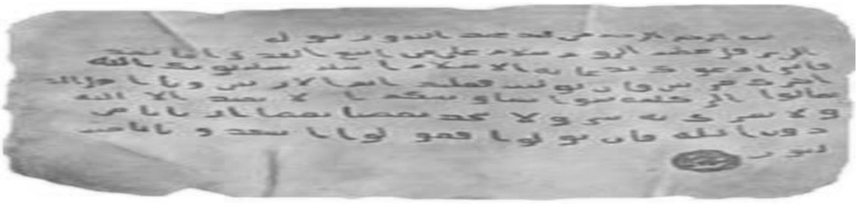


Figure 2-2: Prophet Mohammed's letter to the Roman emperor

Various Arabic calligraphic styles developed in varied Arabian cities, with completely different writing techniques and writing tools. For the most known Arabic calligraphic styles, see Figure 2-3.

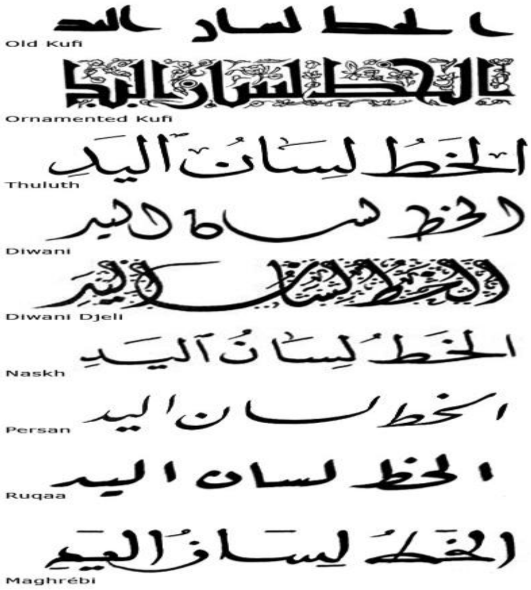


Figure 2-3: The most known scripts of Arabic calligraphy

Today, most of the text typefaces in the market area unit support the Naskh or the Thuluth vogue. The other designs like the Kufi, Diwani and Maghrébi area unit found in show typefaces.

During the industrial revolution in Europe and subsequently the invention of moveable type, many Arabic typefaces were created in France, Italy, England, Germany, Spain and Holland between the 16th century and 18th centuries.

The first Turkish journalism using Arabic printing was found in 1727 in Istanbul, and the initial Middle-Eastern Arabic press was built into a Christian cloister in Mount Asian nation in 1733 where the very first Arabic book was printed in 1735. The four developing layers of the Arabic script are shown in Figure 2-4.

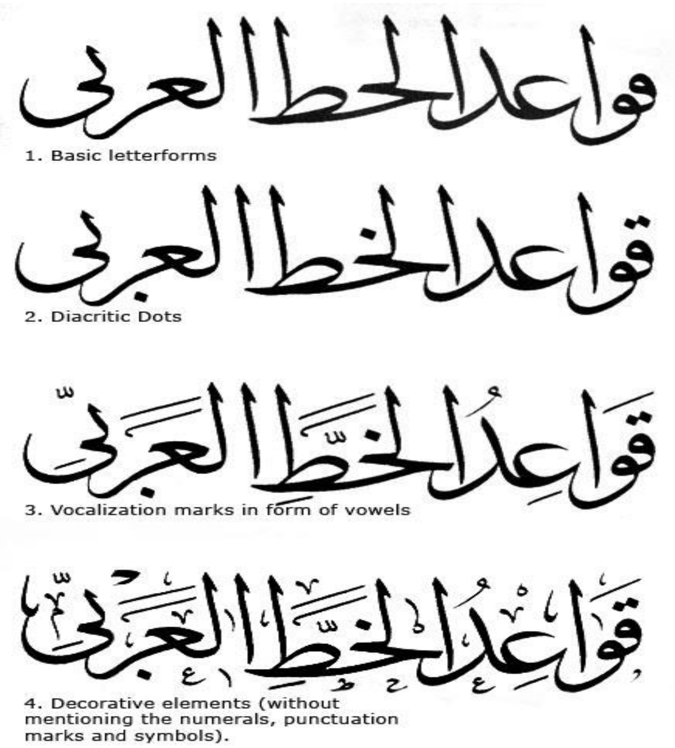

قواعد الخط العربي
1. Basic letterforms
2. Diacritic Dots
3. Vocalization marks in form of vowels
4. Decorative elements (without mentioning the numerals, punctuation marks and symbols).

Figure 2-4: The four typographic elements of the Arabic script

The Arabic language is now universal, and it is a formal language for 25 countries, with a population of 300 million. Arabic text is written from right to left, and it has 28 basic letters, 16 of which have dots. Furthermore, many Arabic letters are also used in various languages such as Urdu, Farsi, Chawi, Kardi, but still the first attempt to recognize the Arabic language was not until 1975 (Atallah and Omar (2008).

The first printing agency was founded in Cairo in the 19th century and at that time Arabic was written without the use of a keyboard; instead four or additional cases with metal glyphs were used (see Figure 2-5). The total quantity of glyphs would reach to about 500 as well as most handwriting ligatures, vowels, Qur'anic punctuation and allographs of letters.

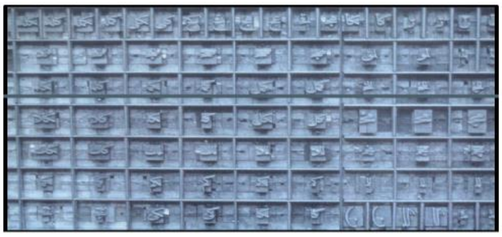


Figure 2-5: Movable type case

In 1945, the Cairo Academy for the Arabic language launched a worldwide competition to modify Arabic as a result of the necessity for a simplified script to suit the new kinds of techniques (i.e. the typewriter). The Academy also improved its own project of typographic simplification of Arabic., as it aimed to scale back the number of different types and use solely accessible glyphs (Azmi and Alsaiari, 2010).

Despite these projects, many Arabic characters are still used in different languages. This may be because the development of Arabic OCR systems has not received enough attention from researchers, as compared to Latin, Chinese and Japanese character recognition systems. It has also been argued that recognition of Arabic characters is more difficult than others, such as Latin and Chinese because the text is written cursively in addition to the complexity of the text characteristics (Al-Shatnawi, 2010).

2.3 Document image analysis

Document Image Analysis is the technique by which the text and graphic elements in the image are classified together with the spatial relations between them. The analysis of these information elements and the identification of the logical relations between them are the responsibility of a subsequent phase, namely, Document Image Understanding (Antonacopoulos, 1995). The aim of document analysis is to divert the data presented on paper and addressed to human comprehension into a computer-revisable form (Baird et al., 2012). The model process of the document image analysis stage is shown in Figure 2.6 (Rahnemoonfar, 2010). In the digitisation phase, the digital image can be acquired by a hardware machine such as a scanner or digital copier, while the image enhancement stage includes digitisation, geometric restoration and noise removal. Another form of analysis adopted is physical layout analysis, which includes three procedures: page segmentation, classification, and physical layout structure extraction. Page segmentation is concerned with the identification of areas of interest in the image of the document page. Page classification is the determination of the kind of contents of each area of importance in the document image. The description of the physical layout is created from the information extracted from the document image while page segmentation and classification. At the ending of the document analysis stage, there is a description of the physical layout of the page in terms of the inter-relations and the properties of the printed regions. This information can be used as a map by applications to navigate within the document image to achieve further processing(Antonacopoulos, 1995).

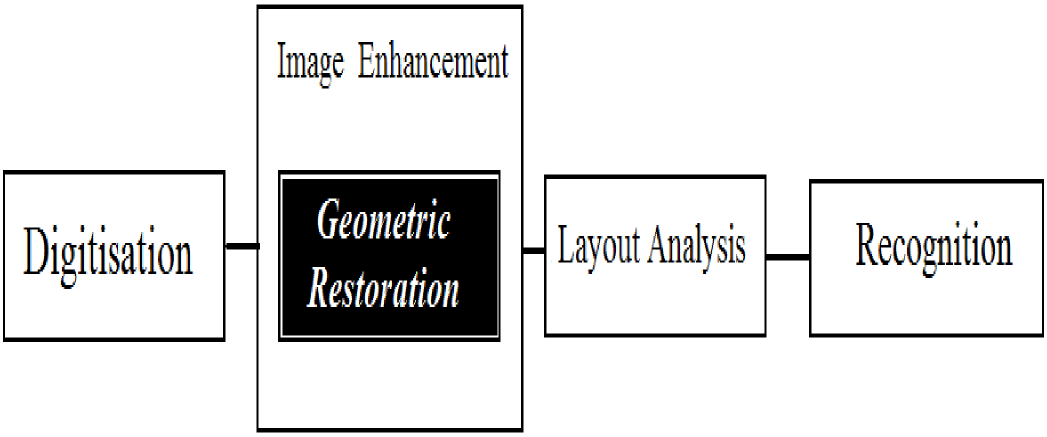


Figure 2-6 Document image analysis steps

2.4 Geometric distortion in historical documents

The issues with historical documents can be divided into three main groups: document-related, acquisition-related (scanner, cameras), and compression-related problems. The existence of geometrical deformities may be found at several steps while the life cycle of a manuscript, from when it is first written to the time it is digitised by an imaging instrument. The distortions involve arbitrary warping occurring from the printing process or storage circumstances (Rahnemoonfar and Antonacopoulos, 2011).The diagram in Figure2.7 illustrates the life cycle of an historical document (Rahnemoonfar, 2010).

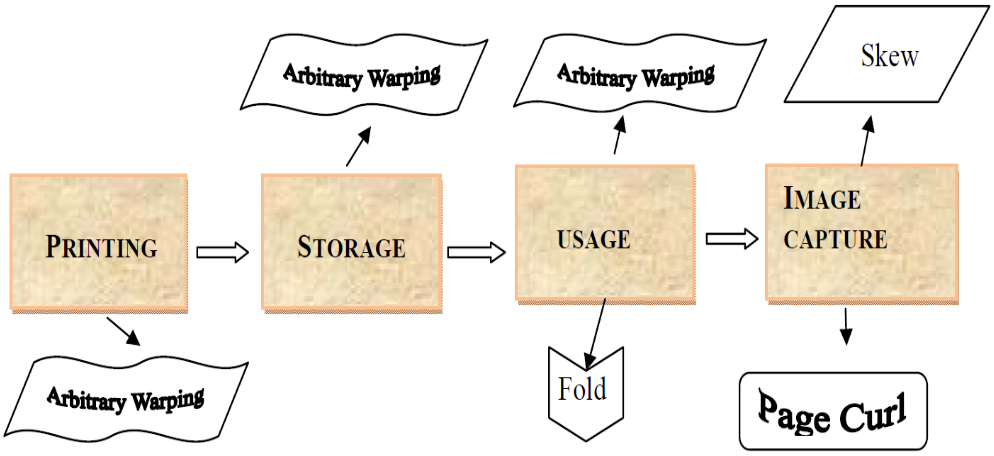


Figure 2-7 the life cycle of an historical document

2.5 Characteristics of Arabic and Latin texts

According to (Slimane et al., 2010), Arabic is now spoken by almost 300 million people across the world and it is understood as a cultural emblem for a number of people. However, in comparison to the printed Latin script, Arabic has certain variations, as the letters change according to their position in the word. In Table 2.1, different shapes of 28 Arabic letters are provided in accordance with their word positions. In the Arabic alphabets, text lines are composed of a chain of characters with different heights and locations relative to one another. Arabic script is cursive, and each letter may have up to four various shapes based on its position in a word. In each Arabic text line the following four fictitious lines as Latin text line can be imagined figure 2.8 Sarkis (21 March 2012).

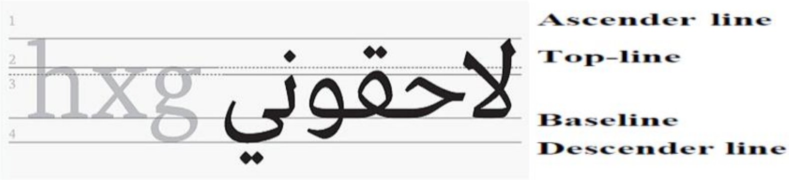


Figure 2-8: Text line definition

Baseline: The line, which connects the lower portion of the letter bodies with no descender in a text line

Top line: The line, which links the upper segment of the letter bodies with no ascender in a text line

Ascender line: The line, which links the peak of ascenders

Descender line: The line, which links the lowest of descenders

Arabic character strokes generally go slanting down whilst those written in Latin go slanting up. Additionally, machine-printed Arabic words are characterized using horizontal ligatures, more or less depending on the font used. On the contrary, machine-printed Latin words are composed of sequential characters without any ligature between them. As a result, horizontal strokes would be more frequent in Arabic words than in Latin words (Saidani et al., 2015). It should be noted, however, that there are some common aspects between Arabic scripts and Latin despite the aforementioned differences (Saïdani et al. (2013). These similarities and differences will now be explained.

Table 2-1: Arabic characters

No	Letter label	Isolated	Begin	Middle	End
1	Alif	ا		ـا	
2	Baa	ب	بـ	ـبـ	ـب
3	Taaa	ت	تـ	ـتـ	ـت
4	Thaa	ث	ثـ	ـثـ	ـث
5	Jiim	ج	جـ	ـجـ	ـج
6	Haaa	ح	حـ	ـحـ	ـح
7	Xaa	خ	خـ	ـخـ	ـخ
8	Daal	د		ـد	
9	Thaal	ذ		ـذ	
10	Raa	ر		ـر	
11	Zaay	ز		ـز	
12	Siin	س	سـ	ـسـ	ـس
13	Shiin	ش	شـ	ـشـ	ـش
14	Saad	ص	صـ	ـصـ	ـص
15	Daad	ض	ضـ	ـضـ	ـض
16	Thaaa	ط	طـ	ـطـ	ـط
17	Taa	ظ	ظـ	ـظـ	ـظ
18	Ayn	ع	عـ	ـعـ	ـع
19	Ghayn	غ	غـ	ـغـ	ـغ
20	Faa	ف	فـ	ـفـ	ـف
21	Gaaf	ق	قـ	ـقـ	ـق
22	Kaaf	ك	كـ	ـكـ	ـك
23	Laam	ل	لـ	ـلـ	ـل
24	Miim	م	مـ	ـمـ	ـم
25	Nuun	ن	نـ	ـنـ	ـن
26	Haa	ه	هـ	ـهـ	ـه
27	Waaw	و		ـو	
28	Yaa	ي	يـ	ـيـ	ـي

2.5.1 Similarities of Arabic and Latin scripts

- **Presence of writing** lines and the same writing orientation: horizontal. Arabic letters are normally joined on an imaginary line named baseline (See Fig. 2-9).
- **Basic band:** It is regularly the most leading in terms of information intensity in pixels (See Fig. 2-10).
- **Regularities and singularities:** The script appears the singularities on either side of the basic band and the regularities inside it (See Fig. 2-11). Singularities describe the start, the end, or the transition to another letter. Regularities include the information in demand for joining a letter to the next letter.
- **Inter–writer variability:** Several people have various styles of writing, which results in irregular shapes for the same characters (See Fig.2-12).

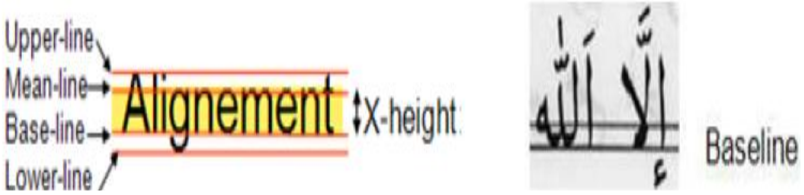


Figure 2-9 : Writing Lines

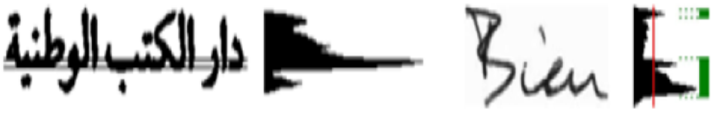


Figure 2-10 : Horizontal Projection.

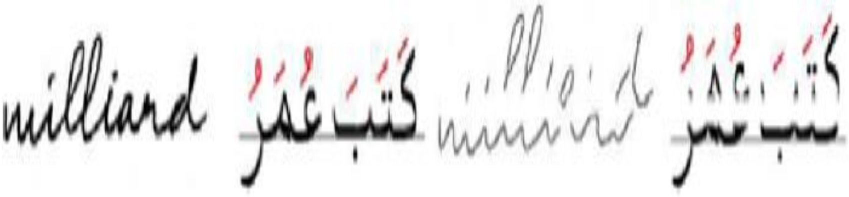


Figure 2-11: Singularities vs. regularities.

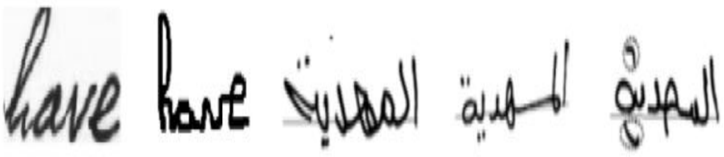


Figure 2-12: the same term was written by diverse writers.

2.5.2 Differences between Arabic and Latin scripts

There are number of aspects which differ between Arabic scripts and Latin:

- **Writing direction:** Arabic is written right to left while Latin is written the left to right.
- **Horizontal projection:** The horizontal projection profiles of Arabic scripts have a single summit around the middle of the text-line (see Fig.2-10). In contrast, projections of Latin texts have two major peaks.
- **Word:** Arabic characters are series jointly to form words in one way only. An Arabic word usually includes two or more letters whereas Latin words sometimes consist of one or more characters.
- **Alphabet:** The Arabic alphabet comprises of 28 characters without observing the difference of their shapes according to the position, the collation elements and the acoustic context. There are 15 out of 28 basic Arabic letters have from one to three dots, and these dots distinguish one letter from another of the same form. Those characters are (ب , ت , ث , ج , خ , ذ , ز, ش , ض , ظ , غ , ف , ق , ن , ي).
- **Letter shape:** All the Arabic Letters are not the same, nor are there capital and lowercase letters. The character forms, however, are variable in form, reliant on the position of the letter at the beginning, middle or at the end of the word.
- **Diacritics:** Arabic script is very wealthy in diacritic symbols. The existence or the obscurity of these diacritics distinguishes the same main shape of Arabic letters.
- **Semi cursive writing:** The Arabic word is a string of connected components named PAWs (Piece of Arabic Word).
- **Elongations:** The same Arabic term does not have a constant length since several elongation numbers could appear between characters (See Figure.2.13).

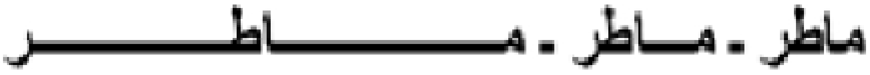


Figure 2-13**:** Writing elongations.

2.6 Problem Characteristics

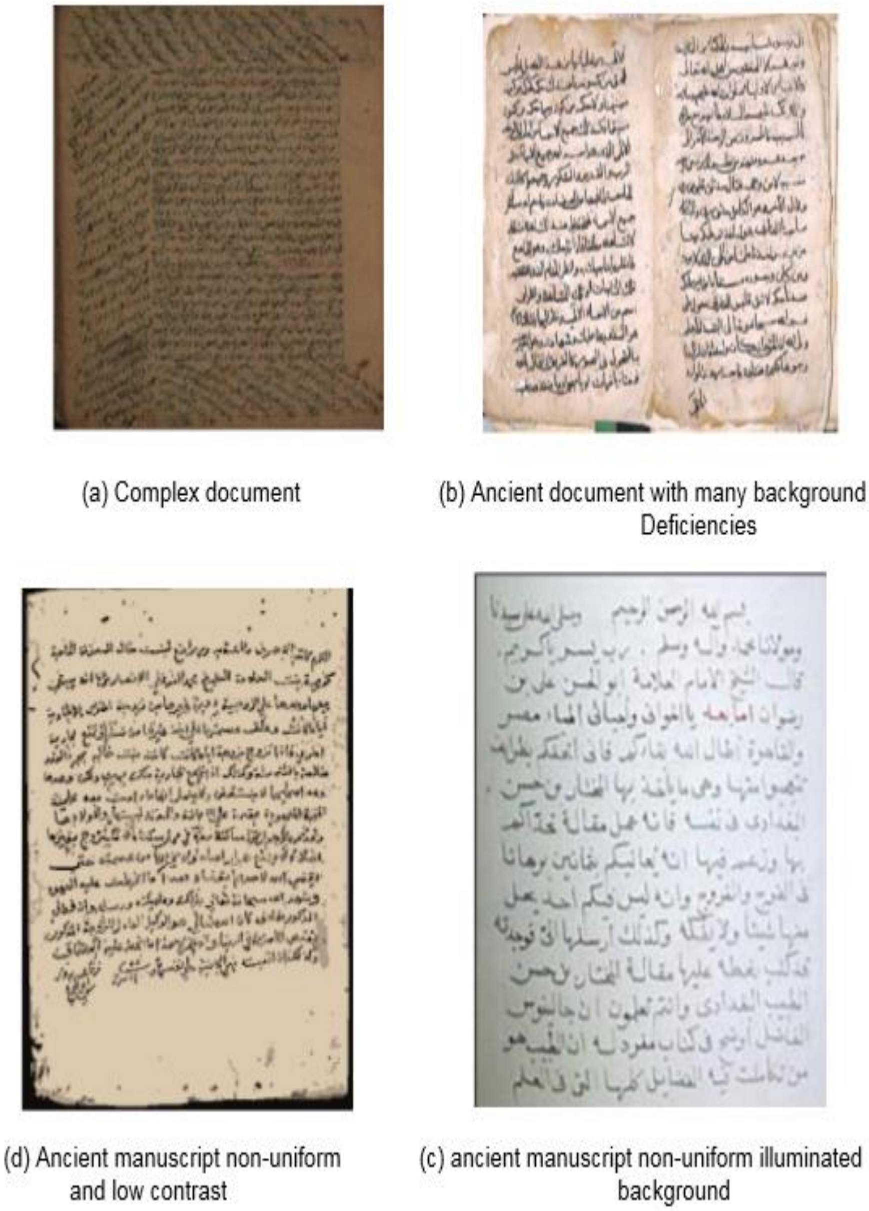
The correction of geometric artefacts is an extremely significant factor of the document recognition process. The significance of correcting geometric artefacts lies in three distinct fields:

1. Excellent recognition outcomes from OCR
2. Extra distinct layout analysis
3. A high- quality reproduction for print on request

Geometric deformation will affect commercial OCRs that are designed for flat pages with straight text lines, and even direct contortion can cause most current OCR systems to crash. However, most region and text line detection techniques act on flat pages, and consequently, methods trying to identify regions and text lines in documents containing such distortions will not work and it is important to remove geometric distortions in initial steps. Such, geometric deformities, relying on their intensity, can have detrimental impacts to on Arabic Character Recognition (ACR) and readability. Finally, there is a viable business for book reprints as well as prints on request, and it is highly beneficial for libraries to have book pages without any geometric deformity. The difficulty of correctly removing different varieties of geometric distortions, for example, page curl, arbitrary warping and fold, is a testing one which has not yet been settled, and towards which a lot of research is as yet being performed. This part outlines some of the fundamental qualities in document pages which make the process of geometric restoration in historical documents a challenging one.

2.6.1 Different geometric distortions

Geometric distortion varies in historical documents from smooth page curl to complicated multi-contortions (see Figure 2.14). Geometric distortions generally, in modern books occur only during catch of images from open books with a flatbed scanner or digital camera. In historical documents there are numerous other factors of distortion in addition to the normal page curl that has been occurred in the document life cycle and is defined in Chapter 1, and which involve differences like arbitrary warping and fold. Moreover, in the state of historical documents bookbinding is often tightly extended, creating a significantly more articulated curl in document images (Figure 2.15) when contrasted with current books.



(a) Complex document

(b) Ancient document with many background Deficiencies

(d) Ancient manuscript non-uniform and low contrast

(c) ancient manuscript non-uniform illuminated background

Figure 2-14 Samples of distortion in historical Arabic documents

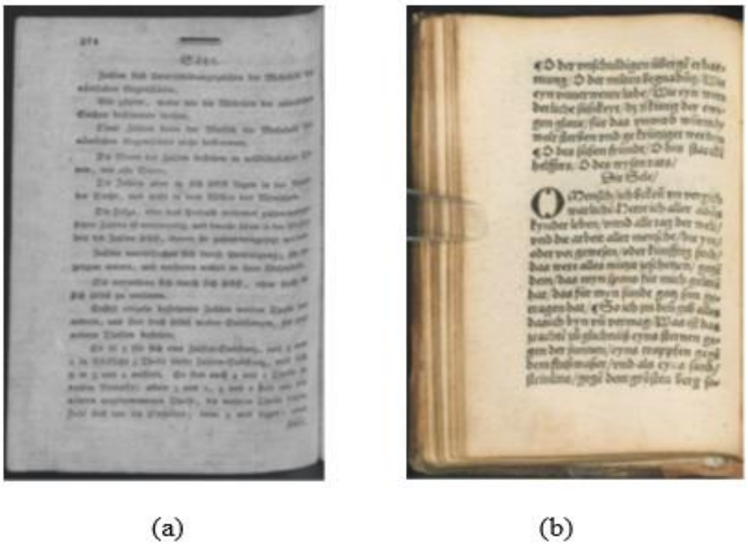


Figure 2-15 Page curl: (a) smooth, and (b) distinct (Copyright: Bavarian State Library, IMPACT project)

Given the previously mentioned deformities in historical documents, different deformities may happen together and make geometric restoration significantly more complicated. Based on these difficulties, the suggested geometric restoration system for historical documents should be robust concerning all of these distortions. Any presumption about the sort of the distortion, for example, the same rate of curving at both the top and the bottom of the sheet or cylindrical surface, will be unsuccessful in the state of historical documents with different distortions. It is not proper to apply a pattern to the whole document, since one single model is not appropriate for application to various pages with assorted distortions. Consequently, geometric restoration in historical documents is harder and is a less developed field at present.

2.6.2 General observations concerning the geometric restoration

The former subsection introduced to a number of distinct geometric distortions which have a significant impact on geometric restoration. Although, there are likewise various non-geometric attributes which are seen in historical document images which fundamentally impact the steps of the document recognition system and subsequently the performance of dewarping approach. These are depicted below.

2.6.2.1 The rate of noise in historical documents

Noise is a common feature in image processing areas. It is degradation by some arbitrary blunders, and it is viewed as an unwanted by-product of image capture. The noise is perceived as any feature for the designated image in document imaging, which does not actually survive in the original document but is brought about by the electronic noise in the scanner or digital camera sensor used for obtaining the document. A lot of image noise can occur through the whole life cycle of a document from printing and storage to the use and scanning of a document, each of which creates difficulties in addition to the common noises that are already in new document images. Historical documents can present a critical amount of noise, based on the age of a document, which may differ from various decades to various centuries, and the quality of preservation. The quality of ancient manuscripts rely on different characteristics: the kind of paper on which they were written, the age of the document, the inks used while printing and the circumstances in which they were stored. The quality of document on which manuscripts were written may range from different textured papers to paper with visible grain which can present significant noise in the binarisation step or even when processing the colour image. Figure 2.16 shows some samples of documents images which include the various amounts of noise (Kefali et al., 2010).

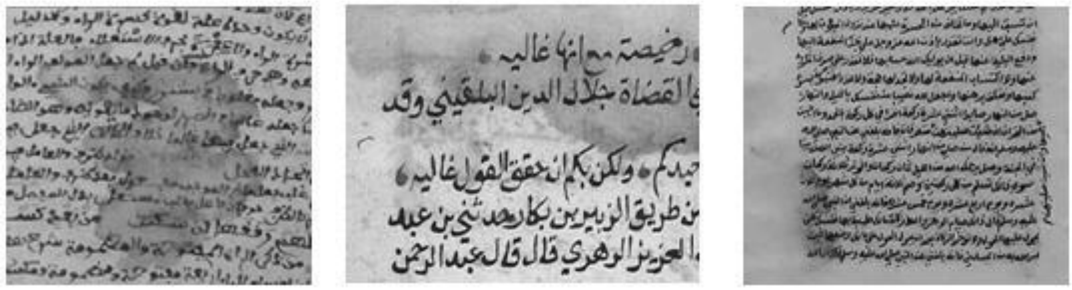


Figure 2-16 some samples of documents images which include the various amounts of noise

2.6.2.2 Fragmentation and merging of characters

In addition to noise, there is another difficulty in historical documents which is low-quality images. Faded ink can contribute several broken characters. In this instance, letters and strokes may be divided into many connected components which confuse the technique of text line detection, since broken components are no longer linked to the baseline of the text and become complex and difficult to segment into the right text line. Figure 2.17 illustrates a section of manuscript of manuscript with faded ink. As depicted in a sample image from a database of degraded Arabic historical manuscripts (Sulaiman et al., 2017). Characters like “ل”, “ص” and “ر” have fragmented into various portions due to faded ink.

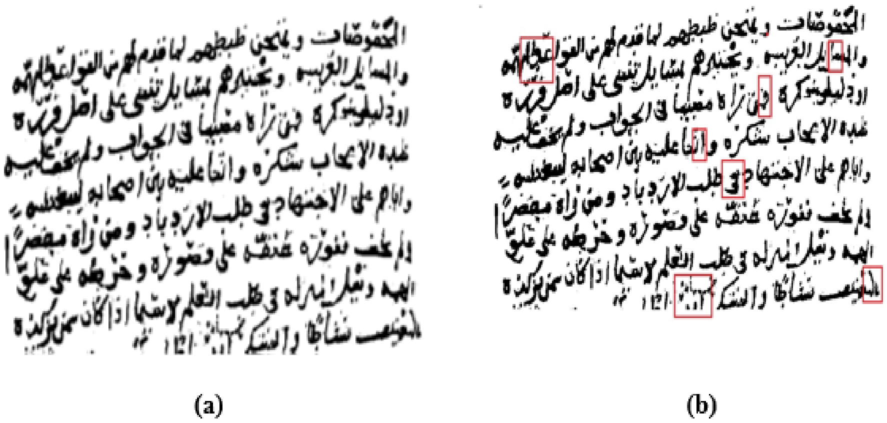


Figure 2-17 faded ink results in broken characters (b) connected components

2.6.2.3 Specific Layout in historical documents

Ancient documents pose significant research difficulties for the document analysis field. Irregular layout structure in historical documents is much larger than in the modern book. In addition to some explicit characteristics for example, headlines, subtitles, paragraph sections, sheet numbers and page separators, there are some particular features such as decorative letters, drop letters and watchword. A watchword (Figure 2.18) is a word located at the left-hand side of the end line on a book page that is intended to be binding over with other pages in the book. The word usually matches to the beginning word of the subsequent page and was put there to support the bookbinder or printer ensure that the sheets were bound in the right scheme or that the leaves were set up in the in the correct order. This method was used widely in the mid-sixteenth centenary and remained till late in the eighteenth century until the coming of industrial printing systems (Rahnemoonfar, 2010).

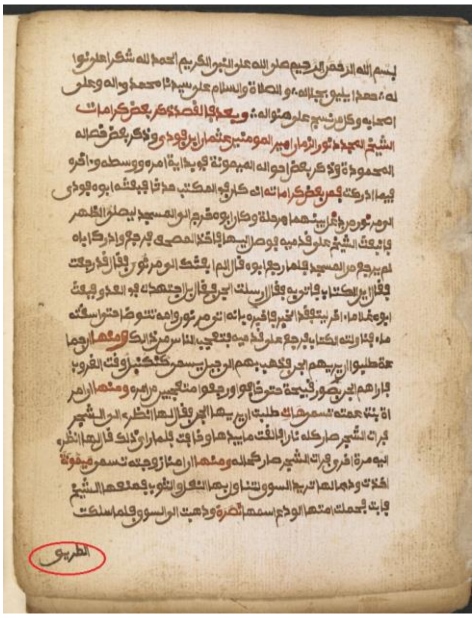


Figure 2-18 Page from the 'Meadows of Paradise' (Rawḍal-Jinān), from the British Library.

Historical documents also contain pages with narrowly spaced lines, usually with touching components (Figure 2.19). The lines are regularly spaced in new documents and consequently simple to detect. The technique of extracting text lines grows more complicated the narrower the inter-linear space becomes.

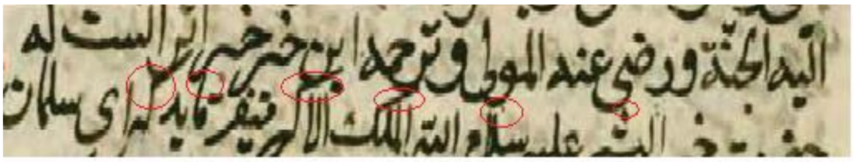


Figure 2-19 Quite dense text including touching components

The extra point of historical documents that differ from current documents is that whereas with current documents there is constant spacing between lines, in historical documents the line spacing is usually inconstant. Additionally, in historical documents, there is no ensured consistency of spacing between characters, words and paragraphs. In new documents, the spacing between lines is less than space between paragraphs, and the lines, in turn, have a wider spacing than the existing words, also more space between words than singular letters in historical documents (Figure 2.20).

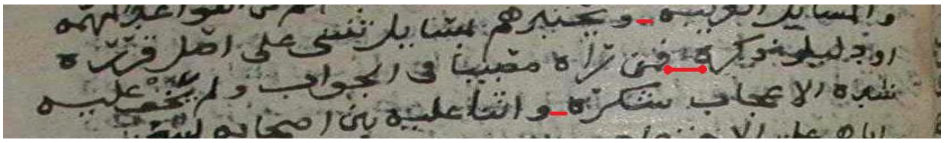


Figure 2-20 inter-word spacing greater than inter line spacing

Historical manuscripts also have unique and diverse forms, depending on the language used and the time when they were written. The languages in historical documents applied in this study Arabic from the 17th to the 20th centuries. The languages offer an extra factor of variability to the books. Given these difficulties, the suggested geometric restoration method for historical documents ought to be sturdy in regard to all of these factors. Different layouts of historical Arabic documents and the appearance of a variety of noise, cause the process of these documents very challenging. Therefore, most of the time, the geometric correction of the algorithm suggested for new documents fail in the state of historical documents.

2.7 Image Form Restoration

Image processing techniques which intend to restore or retrieve an image that has been degraded using the gained expertise regarding the nature of degradation are named image restoration. Shape restoration leads to the retrieving of geometric artefacts. Geometric distortions in images can happen in different methods due to poor imaging techniques for example camera lens deformation or perspective drawing.

2.8 Geometric Transformation Algorithms

Geometric transformation is an essential step in many patterns of image analysis. Image retrieval by techniques of geometric transformation is the product of creating a destination image from a source image according to a mapping among reference space (x, y) and destination space (x', y'). Vector function T that represents the point (x, y) to a novel location (x', y') is named a geometric transformation (Gonzalez, 2016). T is determined by the following equations (Figure 2.21):

$$x' = T_X(x, y)$$

$$y' = T_y(\ x, y) \qquad (2.1)$$

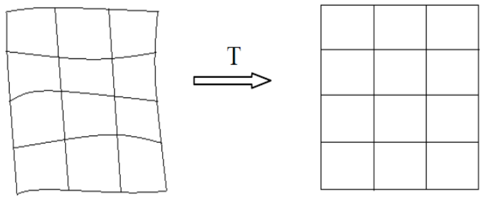


Figure 2-21 Geometric Transformation

Geometric transformation patterns can be classified into two classes: linear and nonlinear Figure 2.22 shows different Transformation types (Øye, 2015). In the next section a brief explanation of various geometric transformations and their features will be given.

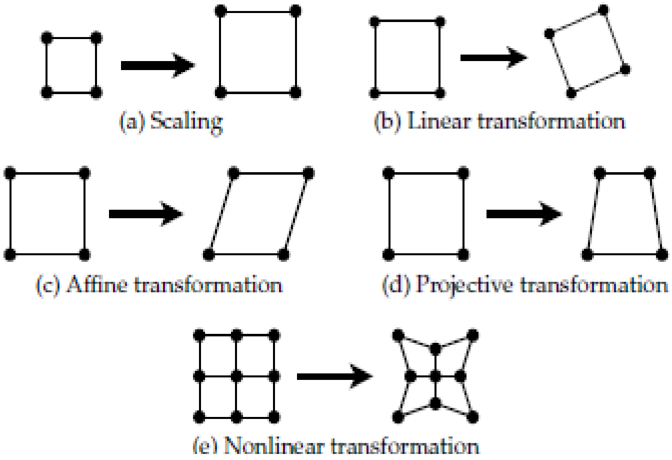

(a) Scaling
(b) Linear transformation
(c) Affine transformation
(d) Projective transformation
(e) Nonlinear transformation

Figure 2-22 Transformation types

2.8.1 Affine Transformation Algorithm

The most common linear transformation of an image is an affine transformation which allows variable contraction, rotation, expansion, transformation and shearing along rows and columns of an image while keeping up co-linearity. Specifically, after affine transformation straight lines stay straight and the rates of distances will be saved, for instance, the parallel lines remain parallel after transformation and the midpoints of a line segment will remain, but it does not necessarily maintain lengths or angles (Rahnemoonfar, 2010). Essentially, an affine transformation is a structure of translation, scaling and rotation, according to the subsequent equations:

$$x' = a_0 + a_1 x + a_2 y$$
$$y' = b_0 + b_1 x + b_2 y \qquad (2.2)$$

Here a_0, b_0 are translate factors, a_1, b_2 scale factors, and a_2, b_1 are shear factors. At least three control points in the source image to produce affine transform. As shown in figure 2.9c, forms of affine transformation.

2.8.2 Perspective Transformation Algorithm

Perspective transformation, also named as the homogeneous or projective transformation, the computations for distortion which appears when a 3D object is introduced within an optical image system onto a 2D plane. The homogeneous distortion creates objects proximate to the camera seem bigger than same-sized objects more remote far from the camera. Projective distortion can be fixed by implementing a perspective transformation. The perspective transformation maintains the straightness of lines while mapping an arbitrary quadrilateral into another arbitrary quadrilateral (Luong and Faugeras, 2001). In case the three-by-three matrix adjusts the z coordinate, at that point the transformation is projective. The perspective transformation is described by a 3×3 matrix that converts projective source coordinates $(\mathrm{x}, y, 1)$ into destination coordinates (x', y', w). Where x' and y' are divided by w, to transform back into non-homogenous coordinates.

$$\begin{vmatrix} x' \\ y' \\ w \end{vmatrix} = \begin{vmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{vmatrix} \begin{vmatrix} x \\ y \\ 1 \end{vmatrix} \tag{2.3}$$

$$x' = \frac{a_{00}x + a_{01}y + a_{02}}{a_{20}x + a_{21}y + a_{22}}$$

$$(2.4)$$

$$y' = \frac{a_{10}x + a_{11}y + a_{12}}{a_{20}x + a_{21}y + a_{22}}$$

The perspective transformation does not maintain the distance between points, but it like affine transform preserves lines at all orientations although conserving the rates of distances as shown in figure 2.9d. An extra important feature is that the inverse of a perspective transform is a perspective transform (Hartley and Zisserman, 2003).

2.8.3 Bilinear transformation Algorithm

The common quadrilateral-to-quadrilateral difficulty can be represented as a bilinear transformation. Also, this transformation described as a mapping of a square into a quadrilateral, to obtain transformation coefficients for every four pairs of corresponding points.

$$x' = a_0 + a_1x + a_2y + a_3xy$$
$$y' = b_0 + b_1x + b_2y + b_3xy \qquad (2.5)$$

In the source image, bilinear mapping conserves lines which are horizontal or vertical and maintains equal-spaced points along such lines, but it does not conserve diagonal lines. Moreover, this transformation does not promise to be invertible (Bookstein, 1997).

2.8.4 Polynomial transformation Algorithm

The universal form of transform can be approached with a polynomial. Polynomial transformations of order m are specified by the following equations(Shapiro and Stockman, 2001).

$$x' = \sum_{r=0}^{m} \sum_{k=0}^{m-r} a_{\mathrm{rk}} x^{\mathrm{r}} y^{\mathrm{k}}$$

$$(2.6)$$

$$y' = \sum_{r=0}^{m} \sum_{k=0}^{m-r} b_{\mathrm{rk}} x^{\mathrm{r}} y^{\mathrm{k}}$$

The transformation is bilinear if m =1 in the equation (2.6) and the initial three terms will produce the affine transform. Polynomial transformation specially contains cubic, bi-cubic, quadratic and bi-quadratic.Tang and Suen (1993) explain these particular transformations as a distortion pattern, and their reverse transform as a restoration type.

For instance, bi-cubic transformation can be described with the following equation, which requires at least ten control points to define twenty strange parameters.

$$x' = a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 + a_{30}x^3 + a_{21}x^2y + a_{12}xy^2 + a_{03}y^3$$
$$y' = b_{00} + b_{10}x + b_{01}y + b_{20}x^2 + b_{11}xy + b_{02}y^2 + b_{30}x^3 + b_{21}x^2y + b_{12}xy^2 + b_{03}y^3 \tag{2.7}$$

Here $a_{\mathrm{rk}}, b_{\mathrm{rk}}$ are transformation parameter, x', y', and the coordinate of a point in the output images.

2.9 Summary

This chapter has given a brief history and development of the Arabic script and describes document image analysis. Following this, geometric distortion in historical documents and characteristics of Arabic and Latin texts is discussed. Finally, problem features, image form retrieval and geometric transformation algorithms are presented. The following chapter will discuss the existing geometric correction and geometric correction approach studies that have been carried out worldwide.

.

Chapter 3 - Literature Review

This chapter provides a complete review of previous studies on geometric correction and geometric correction approach studies that have been carried out worldwide. The chapter will also provide a summary of previous approaches for text line segmentation and baseline detection.

3.1 Image Warping and De-warping

There is a growing body of literature that recognises the importance of digital image warping in the image processing field that deals with geometric transforms of digital images (LI, 2008). A geometric transformation is an action that also defines the spatial correlation between points in an image. The warping could vary from something as simple as a translation, rotation, or scale, to something as complicated as wavy curves. The difficulty of image warping has been the topic of significant attention in numerous fields together with remote sensing, medical imaging and document image analysis for over thirty years. In all these techniques, distortion models are measured as bivariate polynomials whose coefficients are received by reducing an error function over some reference points. Similar techniques can be immediately used to medical imaging and document image analysis fields as well, such as image registration and rotation for digital radiology. Other than these, people in the field of graphics design also make usage of image warping to produce exciting visual effects. However, within the area of document image analysis, the goal is more geometric correction rather than geometric distortion. Assume we have an image of a flat manuscript page (2D texture image) which is designed to a multi-folded surface, now when we take an image of the folded document page, we will have a 2D camera image that seems warped as shown in Figure 3.1. Conversely, if we have the warped 2D camera image, how can we retrieve the original image?

This is identified as the digital image de-warping difficulty, which attempts to correct the warping distortions in the digital document image. In reference to the distortions determined in camera-based document images, a warped 2D image is supposed to have geometric distortions due to the non-planar geometric shape of the document being imaged. For instance, when imaging an opened thick bound book, the warped superficies form produces originally straight text lines to look warped particularly near to the spine region.

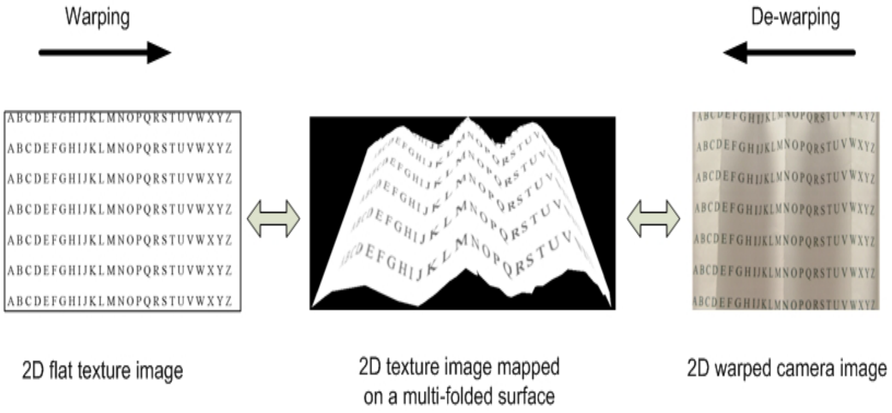


Figure 3-1: Digital image warping and de-warping.

3.2 Geometric correction methods in document image analysis - background and review

The literature on geometric correction proposes a variety of approaches. These methodologies have addressed these difficulties and can be categorized into two key groups based on two-dimensional (2D) document shape reconstruction and three-dimensional (3D) document image processing. The methods in the category above depend on 2D information from camera document images. The approaches in the latter category get 3D information about the manuscript image through special devices.

3.2.1 Restoration approaches based on 2D document image processing

This section will provide a survey about the methods, which generally use 2D information on the image.

3.2.1.1 Geometric restoration based on a priori information

Tsoi and Brown (2004) show a method that uses boundary interpolation in correcting geometric exaggerations, such as, page curl and fold, and shading artefacts present in images of artwork materials. Brown and Tsoi (2006) extend their earlier work to give additional detail to the algorithms, where the full page is framed with only four boundary arcs that represent the top, right, bottom, and left sides of the material. The major drawback of this approach is that it does not use text line information to restore the image; it may be applicable in the case of restoring only images and artwork. Subsequently, the algorithm of Meng et al. (2012) can be viewed as an extension of this technique to a more general case. This extended approach entails using a general cylindrical surface (GCS) to model the curved page shape. In comparison with the previous approach, this method is for fully rectifying the nonlinear geometric distortions in a camera-captured document image, including the distortions caused by perspective, page curl, and their coupling. Meng et al. (2014), meanwhile, can cover the most common types of geometric exaggerations that generally don't work in methods that use the cylindrical assumption from Sharma and Sharma (2016), but focus blur has been given less consideration in this technique, and there are also noise and low contrast in the camera captured. The difficulty of this technique is extraction is weak with low resolution (Felix et al., 2017).

3.2.1.2 Geometric restoration by using flow lines

Rahnemoonfar and Antonacopoulos (2011) published a paper in which they described a new technique to reveal and re-establish subjective warping and folds with page curling. The suggested method is grounded in text line segmentation, accurate standard modelling, and de-warping using transformed primary and secondary flow lines. The method requires no special hardware and prior information is capable of realising a high average accuracy close to 93.94% for a varied dataset of past document images.

On the other hand, the main shortcoming of the technique is its dependency on text lines. When the document image contains an image only, or text within an image, this method cannot be applied effectively, which is also the case for non-Latin script here (Rahnemoonfar, 2010). Another major source of uncertainty is that this technique needs to be modified depending upon morphological features of the written language of the document (Baumann et al., 2012).

3.2.1.3 Geometric restoration by word segmentation

Zhang and Tan (2001b) examined the case of scanning thick books. They found that shadow and warping occurred in the area close to the spine in the scanned book while the rest of the page was clean and straight. They assumed that only the words in the shadow area were warped. Consequently, their technique is based on the correction of word images only within the shadow region. Following this, they proposed a method Zhang and Tan (2001a) that completely performs the same procedure, but uses Hough transform to determine the orientation of words. The method is implemented in four steps: 1) defining the perimeter between the shadow area and the spotless area; 2) an adjusted Niblack(Niblack, 1986) method that removes the shadow; 3) clustering connected components to the word by the nearest neighbour and word to the text line by box hands; 4) the moving and rotating of words. Zhang and Tan (2002) exchanged the fourth step by modelling the straight line in the clean area by linear regression and polynomial regression in the shadow area based on coordinates of the centres of the highest and lowest boundaries of the constraining boxes of related components. Zhang and Tan (2003) also replaced the restoration portion within the old system to flatten curved text lines using easy connected component analysis and regression methods. However, these techniques have several obstacles, which will now be examined. Overall, the first proposition that the geometric distortion occurs only in shadow areas is not generic; oftentimes there is no shadow on whole images or even in a distorted part of an image even in the case of the page curl, and there is no guarantee that the rest of the line is straight in historical documents. Another shortcoming of the method is in connecting words to each other by means of box-hands. Another weakness of the first two methods that Zhang and Tan (2001a) ignore is whether there is any deformation within a word, or cases where between a clean area and a shadow area due to restoration is based on words which are assigned only one transformation and one rotation for each whole word. This is developed in their later method (Zhang and Tan, 2003) of restoring components instead of words. Gatos et al. (2007) suggested the de-warping method based on the segmentation of words and text lines. In the first step, the image is binarised with an adaptive algorithm (Gatos et al., 2006). The major drawback of this approach is when the curl is not moderate because it binds the components that are available in diverse top and bottom lines. Stamatopoulos et al. (2008), however, obtained better segmentation results and consequently de-warped by presenting another stage. The extension of this work by Stamatopoulos et al. (2011a) includes a new technique for the arched surface projection, where the word baseline fits as well as the re-establishment of horizontal orientation.

This approach cannot recover a homogenous font size within those image parts that have considerable distortions (Meng et al., 2014), but it is distinctive in three main ways. The first one involves a content independent technique; the second is a procedure that uses a more common developable theory on the rounded page shape and this method is fully automated; it recreates a developable surface based on two structured beams.

The major drawback of this method is the need for special equipment. Meng et al. (2016) extended their previous work into new techniques for non-uniform shading improvement, which is also established basing on the 3D shape evidence derived from the geometric refinement stage.

3.2.1.4 Geometric restoration by dividing documents into quadrilateral patches

Lu and Tan (2006a) submitted a method restoring camera documents using perspective and a smoothly curved line and Lu et al. (2006), who developed the work presented by Lu et al. (2005), designed only for perspective. To detect text lines, top and bottom points are extracted based on different morphological operators. To extract tip points based on the orientation of the text, different morphological sets are defined. For instance, the applicable morphological sets in the case where the orientation of text lines is from the topmost left to the lowest right are different from a case where the orientation is from the top right to the bottom left. Lu and Tan (2006a) likewise indicated that their proposed technique would fail in the case of non-Latin script and if arbitrary warping is present, because it cannot detect vertical stroke boundaries properly.

This approach (Lu and Tan, 2006b) has two problems. Firstly, there is an issue with the document skew to select the structure the elements to calculate the of topmost and lowest points. Secondly, if the distance between the foundation and the topmost lines is smaller than the distance between the words, the grouping of neighbouring points goes wrong (Bukhari et al., 2008). This method also produces large numbers of over segmentation errors for a small length and a large number of under segmentation errors for a comparatively big length (Bukhari et al., 2013). This technique merges curled line segmentation and baseline approximation, by coupled snakes, although its calculation cost remains too high and hence unsuitable for real OCR applications (Liu et al., 2015). This method can be distributed into three portions: segmentation, estimation and restoration. Each of these portions can be developed in isolation to attain more exact restoration result. The major shortcoming of this technique dependence on the content of the document (Das et al., 2017). This technique processes the document de-warping restriction using a convolutional neural network (CNN) combined with piecewise polynomial regression.

3.2.1.5 Geometric restoration by grid modelling

Lu and Tan (2006a) proposed an approach to solve the problem in their previous papers (Lu et al., 2005), relating to when a character stretches between two rectangles by defining a grid, based on the technique of Lu et al. (2005). The baselines and top lines are extracted and cubic polynomials are adapted to each segmented line. This approach splits a warped manuscript image into several grids based on a predictable lined and character bearing thereby changing every lattice into a square and gathering them to attain a full reconstruction of the image. That technique however, appears to heavily subdivide the image and encounters some discontinuity difficulties related to reassembling each grid together (Zhang et al., 2008). In this technique, the text line and character separation utilises projections on the initially warped document leading to various segmentation errors (Shamqoli and Khosravi, 2013), as the technique projects curved lines to the 2-D rectangular area. This algorithm is, however, sensitive to large numbers of text lines. The method has been enhanced further (Shamgholi et al., 2014), highlighting the projection related to the arched line to 2-D rectangular region minus any external equipment.

Yang et al. (2011) suggests a practical grid-based technique to geometrically illustrate and rectify warped ancient documents with complex layouts. This technique can handle document images that have composite contents, including graphics, several font sizes and column outlines. Another characteristic is that the performance of this method is superior to the best start-of-the-art geometric refinement techniques on transactions with haphazard warping properties. Those difficult historical documents that have exceedingly dense and composite layouts pose a major obstacle to this method. Yang et al. (2016) enhanced the output quality, but there still were missing text regions in the middle of the document. However, most techniques in the literature highlight how to enhance restoration accuracy for certain distortion effects, and not their use within large-scale digitisation. Yang et al. (2017) proposed an effective lattice oriented geometric renovation system for large-scale twisted historical document digitisation.

3.2.1.6 Geometric restoration using active contours

Lavialle et al. (2001) introduced a method based on deformable systems to straighten arched texts using a network of active contours. Essentially, the technique works by deforming and moving an initial curve toward a region of interest. Through minimising the external and internal energy of the model the solution is acquired. This energy is specified based on the presumption that the line spacing is constant throughout the whole page. In this approach the resolution gained using the Bezier curves still contains distortions, and thus not suitable in most cases. To cope with this flaw and to improve the flexibility of the model, in the subsequent approach the Bezier curve fitting was replaced by a cubic B-spline fitting method. Other obstacles of the method are for modern documents since the uniformity of the text lines, such as equal distances between the text lines, might not be maintained between two paragraphs, and also, in historical documents even the interlinear distance in a single paragraph is not necessarily uniform. In (Bukhari et al., 2009) a novel method is presented based on curled text lines information, which was extracted using ridges based on a modified active contour model (coupled snakes). The operation starts by enhancing text lines using Gaussian smoothing, then edge detection techniques are utilised to find the central lines of each text line. The gradient vector flow (GVF) that was introduced by Xu and Prince (1998) for appreciating the external energy in active contours is calculated. The snake models based on energy minimisation algorithms do not recognise descenders, and cannot distinguish between descenders and non-descender characters, which belong to the baseline and should be preserved (Rahnemoonfar, 2010). In Salvi et al. (2015) a novel 2D distortion grid estimation technique was suggested to rectify distorted images based on white space lines that are present between text lines. This method achieves better results compared with the state-of-the art method.

3.2.2 Restoration approaches based on three-dimensional modelling

Pilu (2001) suggested a technique for rectifying the deformation caused by page curl through representing the suited surface as a multilateral mesh. A mesh of appropriate dimension and with the recognised gap between nodes is adjusted to the bare 3D data points. Next, a repeated improvement method adapts the points so that the starting distances between them are reinstated, thus obtaining plane isometry. After this iterative technique, the lattice regresses to an appropriate state that is similar to its original shape. The restriction of this technique is that it is computationally slow, and it assumes a significant number of duplications (>1000) for the system to converge. However, the main weakness is the extrapolation features of the pertinent mesh, as in the case of having a gap in the data a considerable amount of distortion will remain in the image.

A broader perspective has been adopted by (Brown and Seales, 2001, Brown and Seales, 2004) who suggested an image restoration method acquired from a document 3D geometry with a regular lighting system. This technique uses a 3D particle system that has point masses. The particles are lucidly connected through Hookian springs to create a sheet. Their technique can be implemented to haphazardly bent and stretch paper, which was common in crisp historical copies as experienced in their project. However, this method needs a particular setup in addition to high cost. In (Brown and Pisula, 2005, Brown et al., 2007) they calculated a corrective map between the original image and stored representation through using the same input, which is a triangulated 3D mesh of the non-planar document's surface acquired using a structured light scanner and a 2D image of the surface and parameterising the 3D surface to a 2D plane with the conformal constraint. This algorithm is reported to be faster than the mass-spring model. To flatten the 3D model to a planar shape Chua et al. (2005) used a similar mass-spring model, as suggested by (Brown and Seales, 2004) but they modelled the lattice edge as a stick rather than as springs. The stick is simulated by impelling the two particles to maintain a static distance between them thus eliminating the trouble of defining the appropriate spring constant within the mass-spring model. To get quicker simulation using the Verlet integrator (Dummer, 2004), in contrast to previous work submitted by Brown and Seales (2004), this technique is capable of handling sporadic triangular interlaces, which better characterise compound shapes, and also decreases the simulation time significantly. Furthermore, the potential instability complications caused by the physically attuned parameters within the mass-spring model are disregarded using a stick model.

In (Zhang and Tan, 2008) the 3D model and the 2D structure are obtained using a laser range scanner. The geometric rectification technique submitted in this article uses the same concept of stick constraint presented in (Chua et al., 2005). They suggested drag forces on every particle within horizontal and vertical directions to suppress the particles while spreading out the lattice during flattening. Contrasted to the usual digital cameras that simply capture high-resolution images, 2D images caught using a laser range scanner is usually low resolution. Laser scanners have several drawbacks in addition to being costly. There is also a limitation for capturing certain very shiny surfaces caused by specular reflection. However, if such chunks are as small as black text, they will be captured with no challenges at all.

The restoration method is based on a shape defined by a shading technique that recovers the 3D shape of a book surface based on its scanned image, by using the features of the imaging device proposed by (Zhang et al., 2004a, Zhang et al., 2004b, Zhang et al., 2005, Tan et al., 2006). A 3D geometric and optical model constructed based on the structure of the flatbed scanner. To determine the geometric and optical model, two assumptions are introduced. First, the book surface is assumed to be generalized as a cylinder. Second, the book surface is supposed to be lambertian, meaning an even brightness of the mirrored light to all directions with no specular reflections.

The shading information can be used to define the shape using these two assumptions. The shape based on the shading algorithm retrieves the shape associated with the shading information, and therefore when the disparity of the input image is too high or too low, especially close to the book spine, this technique cannot give the correct shape and will have a large deviance from the actual shape, especially for a low contrast image. Furthermore, when the height changes along the spine this algorithm assumes a generalised cylindrical surface and may fail.

In (Zhang et al., 2007, Zhang et al., 2009) the comparable shape based on the shading technique (Tan et al., 2006) is applied to model the shape of the distorted document, and the essential parameters of the camera, which have been used in the shape model are calculated from camera calibration (Zhang, 1999). However, the major obstacle of this technique is that the dimension of the forming element to attain morphological transformations should be carefully fine-tuned (Meng et al., 2013). The authors suggest effective method for revising two common types of radiance artefacts, dark perimeter noises and non-uniform scanning shading, based on a manuscript image scanned from an open book. Furthermore, the submitted method does not have any parameters and hence it does not require any settings and can be fully automated. Moreover, Zhang et al. (2009) work on the assumption that the light source is at immensity or at the optical centre of the camera. However, according to Shen et al. (2015), this assumption is not realistic for endoscopic applications, and shape from shading (SFS) approaches are implemented while deriving image depth information as captured by the bronchoscope compared to the 3D information derived from computed tomography (CT). In (Esteban-Lansaque et al., 2016), one of the drawbacks of such techniques is that SFS consumes huge amounts of time implying that it is unsuitable for real time systems.

3.3 Text Line Segmentation Methods

Text line segmentation is among the essential constituents of document image analysis, and the text line segmentation processes trying to isolate the text lines from images. The segmentation problem is the greatest challenge and main issue in Arabic OCR, and difficulties in text line segmentation may happen due to numerous causes as shown in Fig3-2 (1): When the guideline of the initial word is at a plane that is higher than the subsequent word's guideline. (2) and (3): Undesirable contact involving components from diverse lines. (4) Two extreme lines meeting each other. These properties make the text line segmentation more complicated (Shakoori, 2014), and they directly impact the feature extraction and classification procedure (Safabakhsh and Adibi, 2005, Zeki, 2005).

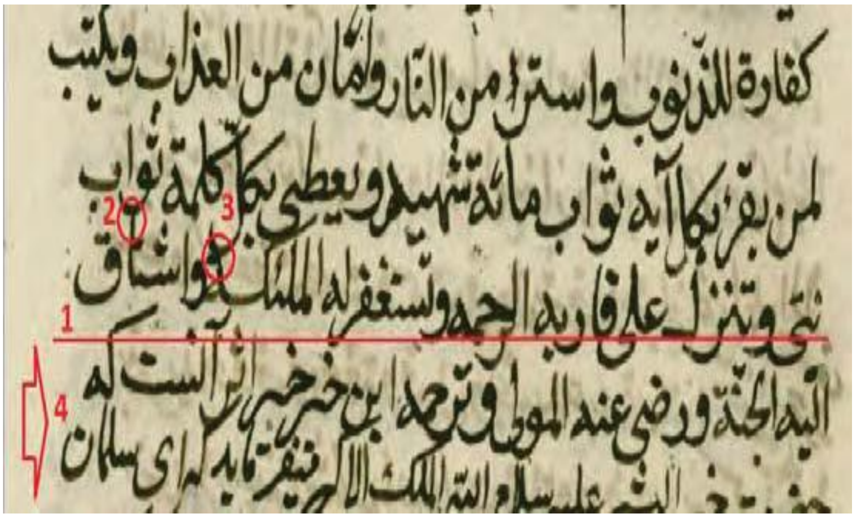


Figure 3-2: Difficulties in text line segmentation

Diverse text line segmentation approaches have been suggested in the literature. In the following sections, these methods will be explained in more detail.

3.3.1 Projection-based methods

Projection profiles, sometimes called projection histograms, have been used in many page segmentation techniques. A projection profile is a histogram of foreground pixel amounts along parallel lines whose indication is perpendicular to that of the profile (Rahnemoonfar, 2010). Let I (w, h) be a grey scale image with width w and height h. Vertical projection profile: the sum of intensity value of a pixel in horizontal lines, lateral to the y axis which is represented using vector I_{VP} of size h and defined by:

$$I_{VP}\ [i]=\sum_{j=1}^{w} I(i,j) \qquad (3.1)$$

Horizontal projection profile: the sum of intensity value of a pixel in vertical lines, lateral to the x axis, which is represented using vector I_{hP} of size w and defined by:

$$I_{hp}\ [j]=\sum_{i=1}^{h} I(i.j) \qquad (3.2)$$

Figure 2.2 shows the vertical and horizontal prediction contour of the image acquired from the above definition.

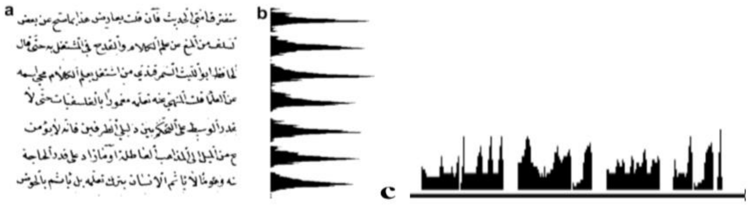


Figure 3-3 (a) binary image, (b) vertical projection profile of the image, (c) horizontal projection

Al Aghbari and Brook (2009) presented a novel holistic method for classifying and retrieving historical Arabic handwritten (HAH) manuscripts and utilised the horizontal histogram method on Naskh documents. The procedure transversally casts a manuscript on a binary image to produce a binary histogram and identifies the apex of the histogram, thereafter computing the mid-points between every two sequential tips and traces the mid-points as line boundaries. This method works completely for reproduced text that has text lines that are relatively straight (Parvez and Mahmoud, 2013). (Zahour et al., 2007, Zahour et al., 2001) utilized bitwise prediction based procedures. In the bitwise parallel projection mechanism, an image disintegrated into vertical streaks and the parallel prediction of each line is calculated. Yet, these methods have a few drawbacks, as they produce too many potential separating lines and the parameter of stripe width is predefined. On the other hand, drawing a complete detach line is difficult, if it lacks a suitable bitwise line among the first and last lines.

3.3.2 Hough transform-based methods

Hough (1962) designed the original Hough transform to detect straight lines and curves, which includes defining lines after transforming each particular pixel in the image with Cartesian space into its analogous parametric space. Likforman-Sulem et al. (1995) and Pu and Shi (1998) used the Hough transform in document image analysis to locate lines of text. In these methods, by starting from some primary points such as the centres of connected components or minima points the Hough transform extracted the lines that fit best to these points. This method is not suitable for arbitrary warping where there are various skew angles along the same text line (Rahnemoonfar, 2010). The Hough transform (HT) is known as a robust tool for graphic element extraction from images (Song and Lyu, 2005). Louloudis et al. (2008) presented an adaptation of the customary Hough transform and a block-oriented Hough transform. The block-oriented Hough transform is practical in detecting the possible text lines after dividing the connected component domain to three spatial sub-domains. This method scrutinises a complete word and a minor dot or stroke within the Hough domain to be equally important. However, the Hough transform has problems in identifying curved text lines (Shakoori, 2014).

3.3.3 Methods based on connected components grouping

A connected component is a group of neighbouring pixels, which share the same intensity values. Subsequently, for example, a single letter “ر”, would be a single connected component, whereas the letter “ز”, would be comprised of two connected components, one for the dot and the other for the body. Connected modules are among the prime elements that obscure text segmentation, due to the anomaly about handwritten texts. Belaïd and Ouwayed (2012) proposed technique divided up a connected module into parts, and the procedure starts by discovering the interchange points and the preliminary pixel of the ligature close to the base of the higher line. However, this method remains impractical with multi-touching components (Shakoori, 2014). In contrast, the method suggested by, Zahour et al. (2009) is founded on block covering and is capable of handling three kinds of connected modules. This technique is designed for separating text-lines even if they are overlapping and multi-touching. Yet, according to Alaei et al. (2011), this technique provides no rollback facility to return to an earlier step, especially if the particular kind of an input manuscript image is inaccurately determined during the initial stage of the algorithm. Clausner et al. (2012) suggested a method using a permutation of rule-based alignment involving the connected components (bottom-up) and prediction profile analysis (top-down). The proposed procedure has been successfully tested on a different dataset of historically representative documents. Khayyat et al. (2012) proposed a strong technique for handwritten text line extraction where their suggested technique used morphological dilation with a dynamic adaptive mask that accomplished suitable smearing to isolate Arabic text lines within a document. Normally, smearing fails to appropriately address overlapping and touching lines. Aouadi et al. (2013) proposed a segmentation method that separates the touching from the overlapping connected components within handwritten Arabic documents. This technique can be used on links involving text-lines and among words. However, these methods have several drawbacks in the case where there are touching components (TCs) representing connections between word letters of consecutive text-lines or those of words of the same text-line. Arvanitopoulos and Süsstrunk (2014) suggested a novel text line extraction algorithm for colour document pages without prior binarisation. Their algorithm is founded on seam carving to work out separating layers between text lines. However, free seam carving tends to generate moving seams among gaps between multiple text lines, and if these are the lowest energy regions of the neighbouring image space. Aouadi and Kacem (2016) proposed a different method for touching components (TCs) segmentation that indicates an improvement on their previous works. The proposed technique concerns segmentation of TCs among words within the identical text-line. In a follow-up study, Boulid et al. (2017) proposed a method for automatic text line detection on binary handwritten documents, based on the analysis of the linearity of neighbouring text components using the watershed transform.

3.3.4 Morphological Operations based methods

Using morphological procedures (Motawa et al., 1997), for instance, a closing followed by an opening, will divide the word into several segments. The starting of the character preserves most of the important information desired to distinguish the character. Semi-vertical or vertical strokes which might describe the start, end or a move to another letter (or subword) are determined by singularities. On the other side, regularities include the information desired for linking a letter to the next letter. Consequently, these regularities are the nominees for segmentation. Regularities are found by subtracting the singularities from the authentic image. While Singularities are detected by effecting an opening to the word image. All regularities are experimented by scanning them from right to left. Segmentation details should appear at regularities. Regularities are categorized to either long or short based on their proportional width to the word, e.g. word aspect rate. The researchers observed that segmented words written by various authors were consistent. The algorithm was fit to accurately segment words with every segment including one letter. Although, in other tests, the algorithm segmented two letters in the same segment. The completion rate reached was 81.88%. Timsari and Fahimi (1996) applied morphological hit-or-miss conversion to segment the letters.

Having the input words depicted in terms of some predefined types. The system's knowledge-base, including characterizations for all letters, are searched for possible matches. Returns a match ends in the recognition of a letter. The authors supposed that this approach demonstrated that to be fast and authoritative in practice. However, no segmentation proportion was given. Techniques based on morphological processes are not sufficient to segment the letters if not supported by other methods (Zaki, 2008).

Table 4.2 is a synopsis of definite Arabic alphabetic segmentation techniques including pattern matching processes, graph theory, undistinguishable Markov models, neural networks and connected morphological operational techniques.

Table 3-1 Summary of the main Arabic character segmentation techniques based on pattern matching, graph theory, hidden Markov models, neural networks, and morphological operations techniques

Reference	Segmentation method	Dataset Type	Accuracy	Comments
Bushofa and Spann (1995)	Pattern matching	Printed	--	Limited to printed fonts
Timsari and Fahimi (1996)	Morphological operations	Printed	81.88%	Limited to printed Documents
Motawa et al. (1997)	Morphological operations	Handwritten	98.30%	Over and Under segmentation
Bushofa and Spann (1997a)	Pattern matching	Printed	---	Limited to printed Documents
Elgammal and Ismail (2001)	Graph-based	Printed	--	Limited to printed documents
Hamid and Haraty (2001)	Neural network	Handwritten	69.72%	Unsatisfactory results
Gouda and Rashwan (2004)	HMMs	Printed	--	Limited to printed documents

3.4 Baseline detection

Often Arabic characters are connected along an imaginary defined as the baseline illustrated in Figure 3.4. The line’s thickness is similar to the pen point although thinner than the breadth of the first character (Romeo-Pakker et al., 1995). Baseline in Arabic script is defined as a virtual straight line where all characters align and connect over it in a specific part of each character (Gacek, 2009). During more than three decades an adequate amount of research has been done on Arabic baseline estimation. The start was by Parhami and Taraghi (1981) where the parallel prediction technique was used to compute the baseline, which was then enhanced by Timsari and Fahimi (1996). The baseline classified the typescripts and outline into three sets of ascenders, descenders and special marks termed diacritics like dots, shadda (zigzag) and maddah. The Arabic character shapes based on baseline are shown in Figure 2.4 (Al-Shatnawi, 2010) .

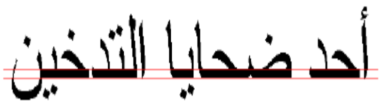


Figure 3-4 The baseline of Arabic script

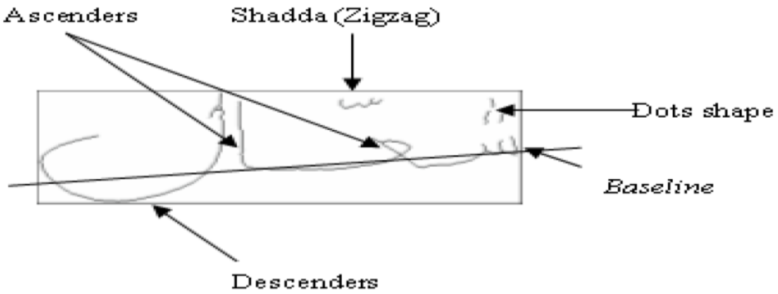


Figure 3-5 Arabic character shapes based on baseline.

Baseline sensing approaches are classified according to the procedures used. In the following, section these procedures will be described in more detail.

3.4.1 Baseline detection methods based on Horizontal Projection

The horizontal prediction based method computes the pixels on the parallel line and supposes that the full number of pixels on the horizontal line as the baseline (Naz et al., 2013). Although this method is strong and very easy to carry out, for Naskh it needs a long straight line of text. Consequently, the histogram projection is unsuccessful in guessing the accurate baseline to isolate the handwritten text from the ligatures with a major number of ascenders and descenders.

This technique has also failed for Nastaliq writing style because of the dots and overlapping, but it has, on the other hand, been very sensitive to the skew (Pechwitz and Margner, 2002). Abu-Ain et al. (2013a) joined the horizontal profile technique with directional features of the primary ligature to approximate the baseline comprehensive dots and diacritics in the handwritten Arabic text.

3.4.2 Baseline detection method based on the word skeleton

Pechwitz and Margner (2002) suggested a method to identify Arabic handwriting baseline based on the word frame. In this technique, the frame of the word is produced from form approximation after which its features can be derived by using the polygonal skeleton.
A major advantage of this method is that it befits various word handwriting styles and is suited for both offline and online baseline recognition even where there are diacritics and dots. Fig (2-5) illustrates the Arabic baseline approximation for the Arabic handwriting word based on the word frame.

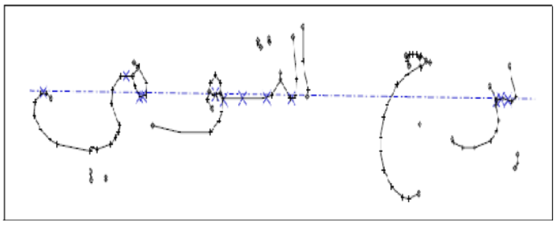


Figure 3-6 Estimation of Arabic baseline based on the word

However, there are certain drawbacks associated with the use of this method in that it consumes time exceeding additional baseline discovery methods since it is applied with a set of complex estimations (Atallah and Omar, 2009).

3.4.3 Baseline detection method based on word contour representation

A single study was prepared to discover the Arabic handwriting baseline based on the word contour illustration used by Farooq et al. (2005). This technique depended on the confined minimal points of the word contour. A major advantage of this method is that it works well with or without diacritics. Another benefit of this approach is that it is also more accommodative with either machine printed or handwritten text for various word handwriting styles. The main disadvantage of the experimental method is that the function of the technique is affected where the diacritics sizes are comparably larger than the main word.

3.4.4 Baseline detection based on the principal components analysis

Many researchers have utilised principal components analysis (PCA) to shorten and distinguish between two and the three-dimension images as well as those used in Latin. One study by Burrow (2004) detected the Arabic handwriting guideline founded on the principle components analysis. This technique is particularly based on angles discovered by principle components analysis. The benefit of this approach is that it determines the Arabic word baseline bearing basing on both the forefront or background pixels distribution. It is also used with dots and diacritics, although the outcomes are more superior without dots and diacritics (Atallah and Omar, 2009).

3.4.5 Baseline detection based on Hough transform method

Li and Xie (2003) presented a new curve detection method aimed at improving accuracy and robustness by diagramming within a parameter space within an image. This space contains accumulators with n-dimensional matrix. The maximum accumulator provides the baselines for word processing. The following equation (2.3) represented the right segment in the Hough space.

$$\rho = x \cos\theta + y \cos\theta \qquad (3.3)$$

With $\theta \in [0; 2\pi]$; $\rho \in [0; R_{max}]$ as R_{max} the diametrical of the processed image, x and y are the measurements of the concerned space. The benefit of this approach is that it can be used on several lines at a time. However, the main weakness is that it is a costly method that is also time consuming during commutation (Naz et al., 2013).

3.4.6 The entropy method

Several methods currently exist for quantifying the information of an image. The entropy is one of the most common procedures for measurement. This technique utilises the horizontal forecasting of the contours' image along the y-axis based on various inclined axes. The histogram density and associated entropy are determined for each projection. This calculated entropy is oriented according to the word that is most compact (Maddouri et al., 2008). The entropy measures the information provided by the following formula (2.4).

$$E = -\sum_{i}^{nb1} P_{i\,\log(P_i)}\ and\ P_i = \frac{N_i}{N} \qquad (3.4)$$

where $nb1$ is the sum of lines within the contour image, Ni denotes the sum of pixels within line number, N the entire number of pixels within the contour word while P_i is the probability of pixel occurrence within the line number i. A major problem with this method is that it also takes more calculation time (Naz et al., 2013).

3.4.7 Voronoi Diagram

Voronoi diagram (VD) is one of the most essential and convenient constructs depicted by asymmetrical lattices widely used within geometrical construction and most other image processing areas. Voronoi diagram is determined where those lines that split the lines between a centre point and its surrounding points are illustrated. The bisecting and connecting lines run perpendicularly and whenever this rule is used the resulting area is fully covered by contiguous polygons. It is also described as a group of geometric objects that divide the plane into cells, where each of such cells has points closer to a specific object than to any others (Eppstein, 1992). The boundaries of Voronoi regions may include line segments, half or infinite lines also referred to as Voronoi edges. A Voronoi edge perpendicularly bisects the segment that joins the two sites. Such two points are known as Voronoi neighbours. A Voronoi point (or vertex) denotes a communal point created from exactly three Voronoi edges. Figure 2.6 shows an example of a VD for a set of points (**pi**) in a plane (**P**) with **n** difference (Al-Shatnawi and Omar, 2009).

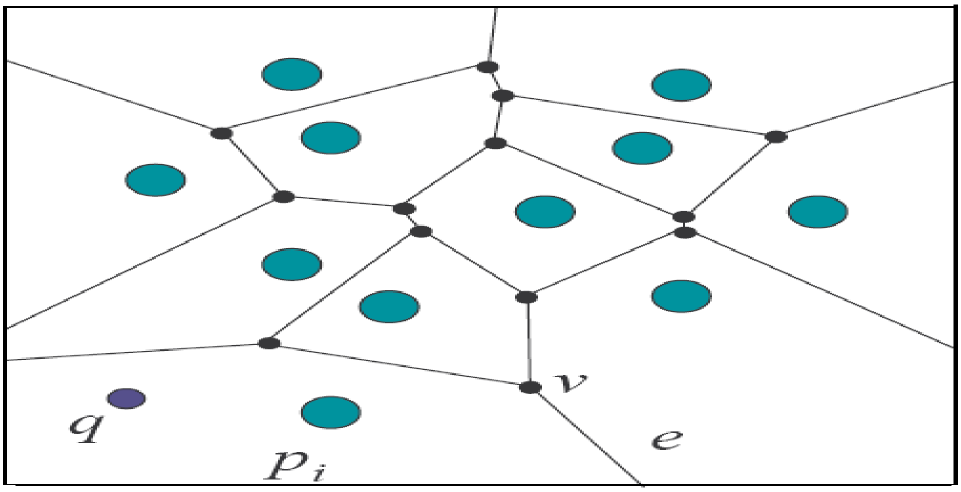


Figure 3-7 (pi) are set points, (q) is as free point, (e) is Voronoi edge and (v) is Voronoi vertex The benefit of this approach is that it is suitable in detecting the zigzag baseline and it is insensitive to diacritics, dots and skew (Al-Shatnawi and Omar, 2009).

The various methods baseline were also presented in the literature Table 3-2 presents the comparison between the Arabic baseline detection techniques (Atallah and Omar, 2009).

Table 3-2 Comparison between the Arabic baseline detection methods

The method	**Horizontal projection**	**Based on the word skeleton**	**Based on word contour representation**	**Based on the (PCA)**
The challenges				
Diacritics Effected	**YES**	**DEPEND**	**DEPEND**	**YES**
Slop Effected	**YES**	**NO**	**NO**	**YES**
Subawards Effected	**YES**	**YES**	YES	YES
Advantage	Easy to use and most common	Work good with the handwriting	Work well with the handwriting	Easy to use
Disadvantage	Work bed with the handwriting	Works with set of the complex calculations	Determine the correct local minima points	Based on the horizontal projection

3.5 Summary

A comprehensive survey of geometric rectification techniques in the document image, including its types and applications was done in this chapter. The methods were categorized under two different categories based on the techniques used. It also examines previous studies for Text Line Segmentation Methods and Baseline detecting techniques. It lists the best methods, challenges and limitations. From this survey, it is concluded that no perfect and error-free geometric correction technique dedicated to historical Arabic documents is available yet. Based on a brief comparison done between VD and different Baseline detecting techniques, a decision was made to explore the use of VD in Baseline detecting rather than others. In the next chapter, the theory and applications of VD will be introduced in more details.

Chapter 4 - Research Methodology and Tools

4.1 Overview

This chapter illustrates the methodology that has been applied in this study. It defines the overall stages in the research framework applied in this study. The explanation of the dataset applied in this research as well as outlining the framework for developing the proposed dewarping system have been stated in this chapter. It further clarifies how the both of experiment and analysis are conducted in order to investigate the performance of the proposed dewarping system.

4.2 The dataset

There is a lack of resources for historical Arabic documents datasets for training and testing purposes .In contrast to the Arabic text, the historical Latin text has had datasets publicly available for a long time, where for Arabic, unfortunately, the availability of historical datasets is scarce, and mostly private (Kassis et al., 2017). This may be because the development of Arabic OCR systems has not received enough attention from researchers, as compared to Latin, Chinese and Japanese character recognition systems. This, in fact, is one of the major challenges faced by the research. To achieve a dataset, which is representative sampling of the historical images found in the real world experiments, the images from some important libraries such as such as Qatar National Library, the British library, and the Library of Congress have been taken. The dataset contains the images with two groups of features, including features due to age or being historical and features due to layout. In consideration of these historical features, the selected images include textured paper, paper with visible grain, bleed-through and show-through, stains, smudges, pencil marks, and broken characters. The layout features in the dataset contain images with ornamental letters, drop letters, decorative borders, dense paragraphs, catchwords and documents in which different font sizes coexist as well as documents with non-constant space between text lines, words and characters. Due to limits in the kinds of distortion some synthetic images were created such as fold, while the majority of documents suffer from arbitrary warping. Figure 4.1 shows an example of document image with synthetic image page curl as a geometric distortion. Various challenging elements are observed in the image such as dense paragraph and drop letter. Table 4.1 epitomises the challenging problems in the image.

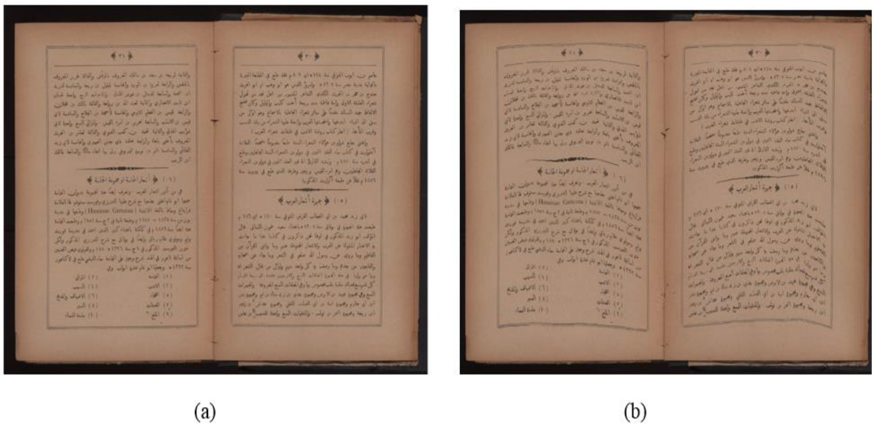


Figure 4-1 (a) the original image (b) the warped image (images with page curl)

Table 4-1 Different issues in Figure 4-1

Geometric issues	Page curl at the right-hand side of the page, Different curve at top and bottom of the page
Historical issues	Visible letters, touching components (dense paragraph)
Layout issues	Drop letters, touching components (dense paragraph)

This research presents a new dataset (and the methodology used to create it) based on a wide range of historical Document Images. There are various qualities that describe a good dataset both in terms of content and usability (Papadopoulos et al., 2013). The three main characteristics are:

Realistic: The dataset have to contain a representative breadth of real documents likely to be scanned in everyday situations.

Comprehensive: It has to encompass detailed information to allow in-depth evaluation.

Flexibly structured: It ought to be easy to browse,search and select subsets with specific conditions.

4.3 Research methodologies

The conclusive goal of this research is to find a solution to the research problem described earlier, which can be summarized in different types of distortions that can occur at several stages of a document life cycle namely printing, storage and usage. One of these distortions is arbitrary warping, which is indicated by wavy curves in the document image. The proposed dewarping system based on the baseline detection is considered as a major problem for this research and so the study is particularly concerned with definition of process/ method/ algorithm, and framework.

The research methodology was developed in four main phases, including theoretical study, measuring the performance of the Geometric Correction methods, and building a system for dewarping, comparison and analysis. The methodology followed in this research is presented in Figure 4.2.

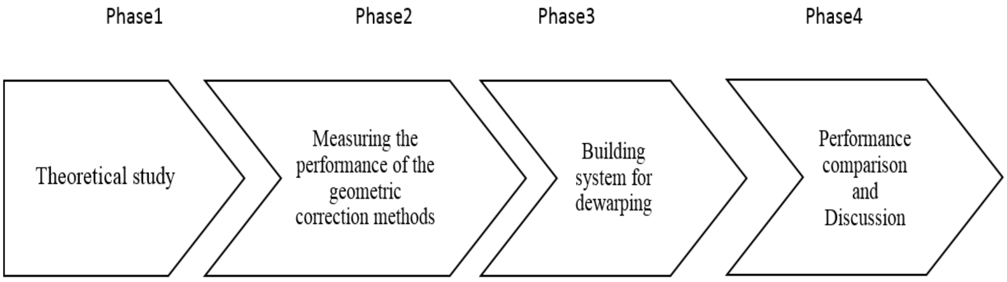

Phase1
Phase2
Phase3
Phase4
Theoretical study
Measuring the performance of the geometric correction methods
Building system for dewarping
Performance comparison and Discussion

Figure 4-2 the methodology followed in this research

4.3.1 Theoretical study

The performance of dewarping (correction of arbitrary warping and page curl) methods involves text line segmentation and baseline detection as one of its important pre-processing steps, and a review is done on text line segmentation and baseline detection methods for printed or handwritten historical Arabic documents. Focus is placed on the usage of the correction of arbitrary warping and page curl technique in this area to show the role and importance of dewarping process in the OCR. Then, an inclusive previous studies and researches look into the previous Geometric Correction Methods and classifies them into a number of categories based on the technique used; this is presented in Chapter 3.

4.3.2 Measuring the performance of the geometric correction methods

Choosing or designing an effective geometric correction method for historical Arabic documents is a crucial step to take when looking at dewarping systems due to its influence on the system's output. When the designed or selected algorithms are ineffective, the system will be negatively affected. The quality of the performance of a restoration algorithm depends on the purpose for which the algorithm is intended. Since the visual quality of the output image is important for prints on demand purpose, visual perception is still a significant factor in evaluation. Therefore, the challenges that need to be considered when designing or selecting a dewarping system for historical Arabic documents in general are extensively studied in the Chapter 2.

4.3.3 The proposed system for dewarping historical Arabic documents

In this research we focus on developing a dewarping system. Generally, the proposed system for improving dewarping documents with arbitrary distortions involves five stages, as shown in Figure 4.3.

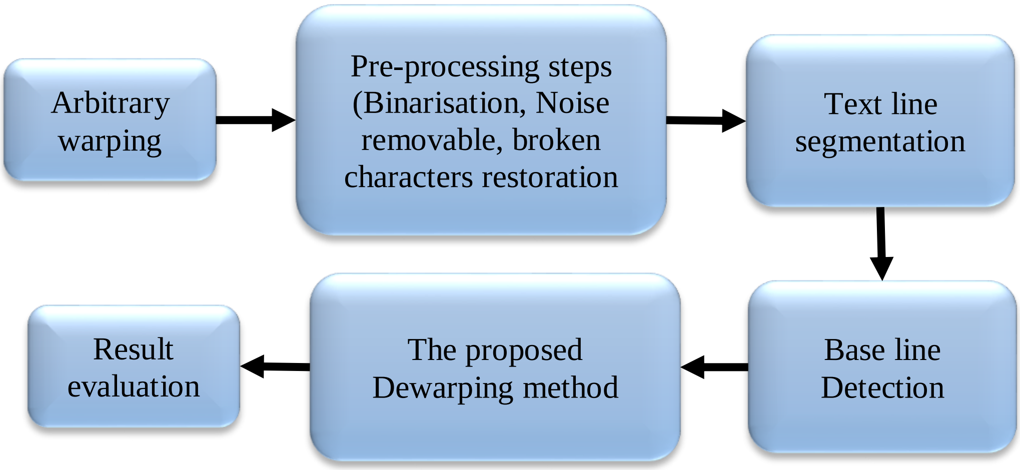

Arbitrary warping
Pre-processing steps (Binarisation, Noise removable, broken characters restoration
Text line segmentation
Base line Detection
The proposed Dewarping method
Result evaluation

Figure 4-3 Diagram showing the general process stages

4.3.3.1 Stage one – Pre-processing steps

The major challenges with historically important documents and manuscripts are the high degree of degradation of various types, and so many research studies have been carried out to solve the problems that arise as a result of degradation of old document images. A key step in all document image processing workflows binarisation, noise removal and broken characters' restoration (Nafchi et al., 2014, Sonka et al., 2014).

4.3.3.2 Stage two - Text line segmentation

Text line segmentation plays a crucial role in this research. The performance of the line segmentation engine has significant influencer on the accuracy of character segmentation and recognition (Zahour et al., 2007). Printed historical documents belong to a large period from the sixteenth to twentieth centuries (reports, ancient books, registers, card archives). Their printing may be faint, producing writing fragmentation artefacts. However, text lines are still enclosed in rectangular areas (Likforman-Sulem et al., 2007). Different text line segmentation approaches have been suggested in the literature and a number of techniques have been developed for text line segmentation. Connected components are one of the main elements complicating text segmentation because of the irregularity aspect of handwritten texts such as line warp, interline and the characters' size. The technique proposed in this stage is based on morphological procedures (Motawa et al., 1997). The performance of the proposed method has been systematically evaluated in the context of large-scale digitisation using a standardised framework. As previously declared, the majority of Arabic characters are related at the baseline and correspondingly, the application of morphological techniques (Zolait, 2013, Alginahi, 2013) in terms of closing and then opening would conclude in various segments. In the starting, the techniques regarding character conservation would contribute towards determining upon the alphabet. Semi-vertical strokes or straight up could be considered at the beginning whereas the conclusion of the transition relative to other letters or sub-words could be taken comparatively to various singularities. Table 3.1 presented a summary of definite Arabic alphabetic segmentation techniques including pattern matching processes, graph theory, undistinguishable Markov models, neural networks and connected morphological operational techniques. Hidden Markov Models (HMMs) is a method of modelling systems with discrete, time-dependent behaviour characterized by the common, short time “processes” and transitions between them. The 2D problem of text image must be converted into 1D problem in order to use HMMs for handling Arabic text. This is implemented by scanning the line of text from right to left column by column (after eliminating the secondaries) to extract features. The invariant moment is used to extract the features. HMMs are then fed with the features (Zaki, 2008).

4.3.3.3 Stage three - Baseline detection

One of the most important steps is baseline detection and straightness. The majority of the techniques failed to distinguish the right baseline when short characters and extensive diacritics exist. Baseline Detection is always necessary to fit a curve after the text line segmentation approach. The baseline in Modern Arabic printed texts can be detected ideally by using the horizontal projection histogram. Due to the arbitrary nature of large-scale digitisation of historical documents, it is not always practical to model baselines (especially in the presence of folds) with a smooth curve (Rahnemoonfar and Antonacopoulos (2011)). The various methods baseline were also presented in the literature. Table 3-2 presented the comparison between the Arabic baseline detection techniques (Atallah and Omar, 2009).

4.3.3.4 Stage four – Dewarping

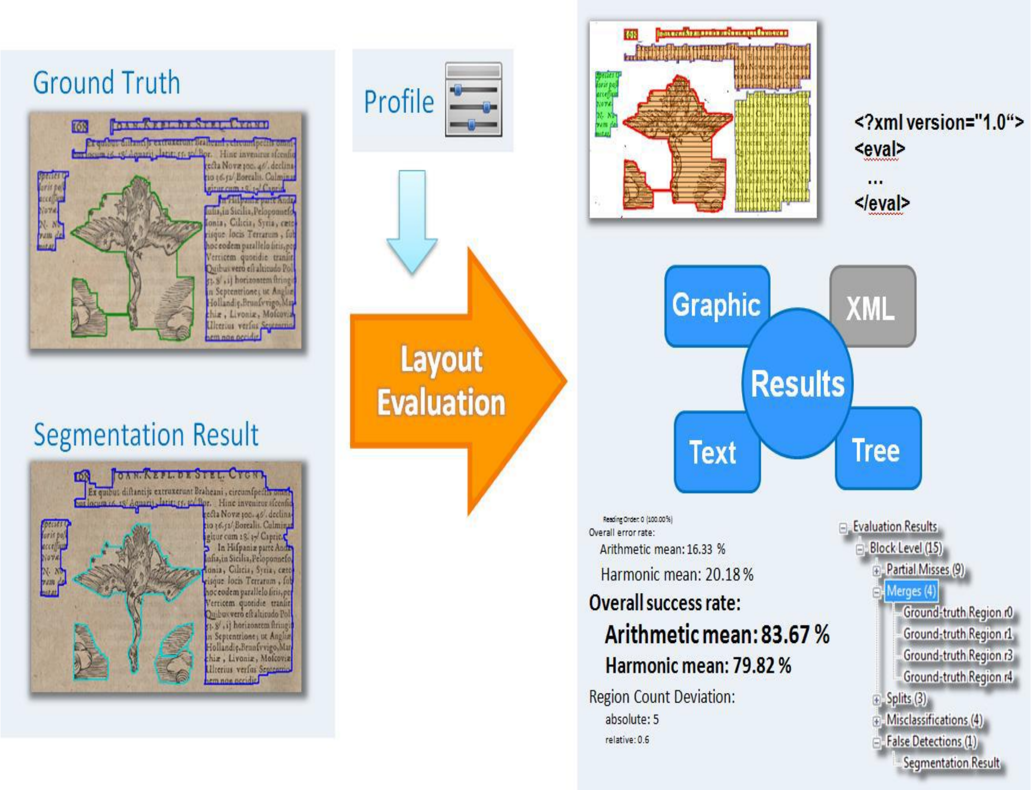
There are several methods for dewarping, and in this research we will use the proposed dewarping method that take into consideration both the global and local characteristics of the document image and models the smooth deformations between text lines. Also, we will take advantage of the proposed line segmentation and baseline detection stages, to cope with a variety of distortions such as page curl, arbitrary warping and fold, in a reliable, robust, and flexible method. The proposed method is capable of dealing with various languages, such as Arabic (both old and modern), unlike techniques that produce bi-tonal results; the proposed dewarping method is suitable for colour, grey scale and binary images. The suggested dewarping method effective mesh based geometric restoration system proposed by Yang et al. (2017) for large-scale distorted historical document digitisation.

4.3.3.5 Stage five – Evaluation

The performance of the proposed methodology will be evaluated and compared with several supervised and unsupervised methodologies for dewarping performance evaluation proposed.

To gain deeper understanding of the behaviour of a method and identify problems it is important to have a means of directly evaluating segmentation and baseline detection results. Supervised evaluation methods consequently need a perfect image to evaluate the production image is mainly based on the ground-truth, whereas unsupervised methods can estimate the performance independent of the perfect image, that mainly based on the simple idea that the final text lines should be straight. Based on precise baseline detection, the proposed evaluation method is therefore an accurate and flexible option for both historical and modern documents.

The evaluation technique that has been used for the experiments is based on line correspondence analysis proposed by Clausner et al. (2011b) . Resulting text lines are compared to those manually created by Aletheia. The ground-truthing system Aletheia (from the Greek word for ‘truth’). As input the ground truth XML file, the segmentation result XML file and the black–and –white document image are required. Figure 4-4 displays evaluation system as a general overview. Resulting text lines and baseline detection are compared to manually created, by Aletheia. Aletheia is an excellent system for accurate and yet cost-effective analysis, recognition and annotation of scanned documents. It supports the user with a number of automated and semi-automated tools which were developed and fine-tuned based on feedback from major libraries.



Reading Order: 0 (100.00%)
Overall error rate:
Arithmetic mean: 16.33 %
Harmonic mean: 20.18 %
Overall success rate:
Arithmetic mean: 83.67 %
Harmonic mean: 79.82 %
Region Count Deviation:
absolute: 5
relative: 0.6

```
Evaluation Results
  Block Level (15)
    Partial Misses (9)
    Merges (4)
      Ground-truth Region r0
      Ground-truth Region r1
      Ground-truth Region r3
      Ground-truth Region r4
    Splits (3)
    Misclassifications (4)
    False Detections (1)
      Segmentation Result
```

Figure 4-4 evaluation system (Clausner et al., 2011b)

4.3.4 Performance comparison and discussion

The state-of-the-art methods will be selected as benchmarking techniques to compare with the proposed method. In this phase we will conclude with the result verification process where testing and evaluation are carried out to evaluate the de-warping system. The testing is conducted on a variety of texts including printed and handwritten Arabic script with different fonts and sizes. Finally, the success rate of the proposed system will be estimating and compared with comparable methods presented in the previous researches.

4.4 Development environment

In our empirical function, and to achieve our research aims, we used MATLAB Image Processing Tool Kit to implement the proposed collection of algorithms. MATLAB provides a comprehensive set of functions and algorithms customised by tuning their parameters in order to fit different image processing problems. Ascender line and top line in modernistic documents are always parallel to baseline and can be obtained with a simple shift according to the height of letters; however, in historical documents due to vertical retraction of letters, this theory is not accurate in all situations.

4.5 Conclusions

In this chapter, the following methodology of this research is clarified. This methodology consists of four phases. The theoretical study phase examined the literature on the related issues and problems of this research. Then, the fundamental theories and models of the proposed techniques were proposed. In the third and fourth phases, the designs were executed and the experiments on the proposed techniques were conducted on the selected datasets consequently. Experiments outcomes evaluation and analysis were achieved in the last phase. The main novelty of the proposed approach is to utilize the knowledge of the Voronoi diagram (VD) to detect baseline. The purpose is to identify the baseline; however, in first step the text line should be segmented. This method, which is both flexible and has the desired accuracy, will serve as the basis for a major correction stage, namely the dewarping procedure.

Chapter 5 - The System for dewarping Historical Arabic Documents

5.1 Overview

In Chapter 3 a survey of existing dewarping techniques along with their features and drawbacks was presented. It was explained in Chapter 3 that there is not an efficient general approach in the literature which can be applied, with no difficulty, to historical documents with arbitrary mutilations. This chapter illustrates a sturdy, resilient and accurate dewarping technique which is capable to deal with the combination of geometric distortions including page curl, arbitrary warping and folds in historical Arabic documents. The adaptability, accuracy and strength of the suggested strategy accomplished by several procedures, to be specific, multi-step starting handling, adaptable text line segmentation, accurate baseline detection, and an inclusive dewarping technique. In this chapter, all the previously mentioned stages will be explained about in detail.

5.2 Review of the researched techniques

The scope of this chapter is to design a dewarping framework with the purpose of satisfying the aims laid out in Chapter 1 and the acceptable characteristic of dewarping strategy defined in Chapter 4 -to design the adaptable, reliable and robust dewarping system illustrated in this thesis, different techniques were designed in which some of them are appropriate for special purposes but they are not adaptable, reliable or precise in all requirements. In addition to these, there are some definitive techniques proposed in this chapter which fulfil all the coveted qualities and have been used as a part of the last dewarping method. To be distinct from any specific setup, the proposed strategy will apply the information in a single image to flatten the document. The most important characteristics in the document image are text lines. The goal is to distinguish the baseline; in the initial step, the text line ought to be segmented. One of the text line segmentation techniques which require less pre-processing the vertical projection profile technique. In the next part, the ability of the vertical projection profile will be tested, however since it does not achieve the wanted adaptability and precision, it was not used as a part of the final system.

5.3 Text line detection based on projection profile

One of the most important suitable solutions for line segmentation in historical documents which is sturdy to noise, different corruptions and needs less pre-processing is based on projection profile. Generally, this technique showed up in the beginning period of OCR Zaki (2008). It performs better with printed documents, particularly with fonts which do not form ligatures such as Arabic, for example, Arabic Transparent and Simplified Arabic. Although, for text styles like Traditional Arabic which contains numerous ligature patterns, it does not work well and even most noticeably worst with handwritten text. The main defect of this technique is its affectability to the skew of the lines in the document. Moreover, a slight skew point makes this method impossible.

The strategies of projection profile depend on the fact that the stroke between letters is always of less thickness than different parts of the word.

The horizontal projection is defined as:

$$h(i) = \sum_{j} p(i,j) \tag{5.1}$$

and the vertical projection is defined as:

$$v(j) = \sum_{i} p(i,j) \tag{5.2}$$

Where

$$p \text{ is the pixel value}$$

$$= \begin{cases} 0 \text{ white pixel (or background)} \\ 1 \text{ black pixel (or forground)} \end{cases} ,$$

$$i \text{ is the row number , and } j \text{ is the colum number.}$$

The horizontal projection is beneficial in isolating the lines and detection of text baseline, while the vertical ones are applied to segment words, subwords and letters. Figure 5.1 illustrations the horizontal and vertical projection profiles of an Arabic text.

.

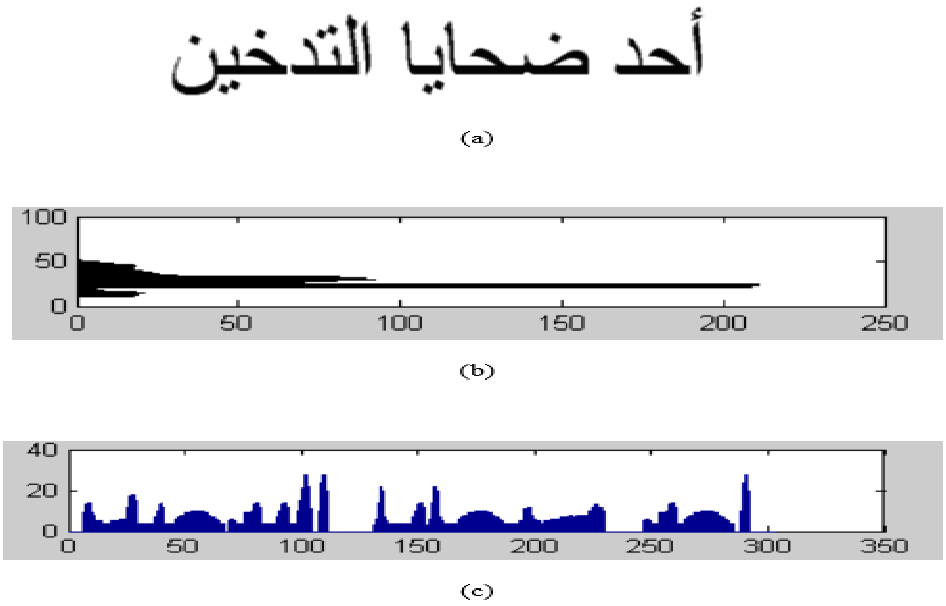


Figure 5-1: Horizontal and vertical projections: (a) An Arabic text, (b) Horizontal projection, (d) Vertical projection

Figure 5.2 illustrations the horizontal and vertical projection profiles of Persian text.

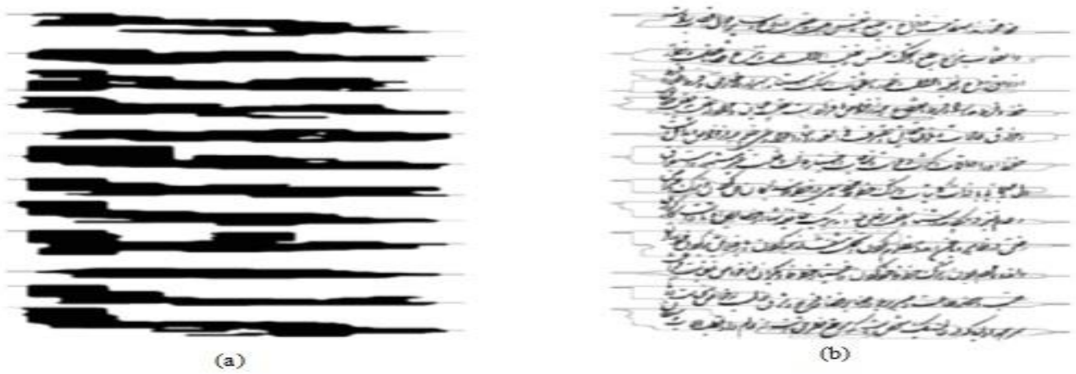


Figure 5-2: Line segmentation. (a): smoothed image (b): Final Segmentation of Individual Lines

After experimenting horizontal Projection-based methods with different document images, it can be observed from the figure 5-3. The projection-based method does not have the coveted versatility and reliability. In addition, even if the technique was able to segment text lines correctly, it is still necessary to do further processing on components' bounding boxes to detect the baseline because the segmented line is far departed from the wanted baseline.

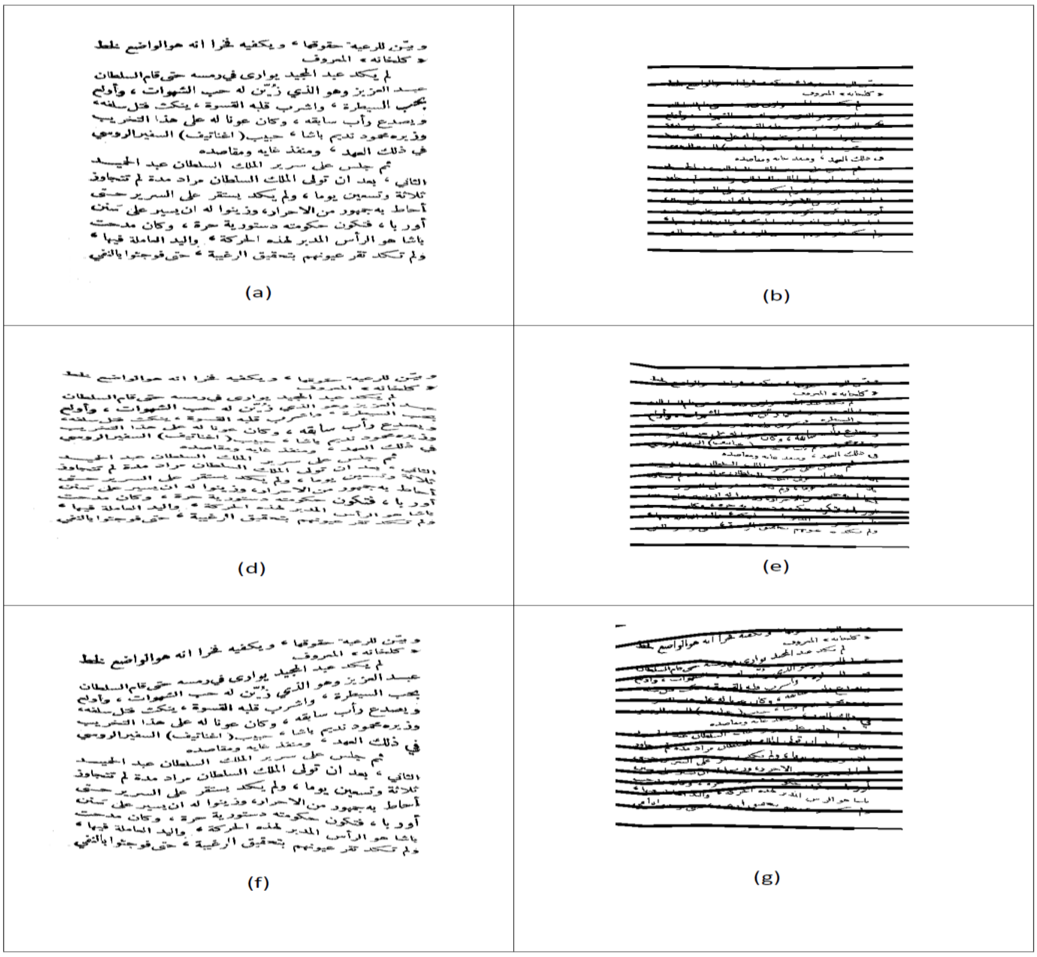


Figure 5-3:Result of Projection-based methods on different warped samples (a), (d) and (f) .segmented text lines (b), (e) and (g)

5.4 Text line segmentation based on connected components grouping

In order to identify text lines more accurately and expedite the method of line detection with connected component grouping, different primary processing stages are required.

5.4.1 Initial processing of the connected components grouping

The method is noise-independent, also it is computationally costly to make connected component analysis direct on colour images and it slows down the procedure significantly. Subsequent the noise elimination stage, the difficulty of broken characters either created by binarisation or because of some special fonts will be considered.

5.4.1.1 Binarisation

Binarisation is the key point in all document image-processing workflows, and this stage just accelerates the computation and does not have any effect on the quality of the final image. There are different binarisation techniques in the literature, which can be divided into two categories, global and local thresholding methods. In spite of the presence of all binarisation techniques, there is not any technique that can be applied effectively in all types of digital documents (Sezgin, 2004, Leedham et al., 2003). In this research, we have introduced an image binarisation method by Nafchi et al. (2014) that uses the phase information of the input image, and robust phase-based features extracted from that image are used to build a model for the binarisation of ancient manuscript. We applied the binarisation method on warped document images selected from Qatar Digital Library. Fig 4.1 shows some sample warped document images image before and after applying the binarisation method.

A collection of global and local adaptive binarization methods, namely Otsu's method (global) by Otsu (1975) and Niblack's method by Niblack (1986) were applied on sample warped document images selected from Qatar Digital Library. Fig. 5-4 the results show that this technique performs extremely well; however, it is restricted to binarising handwritten document images only.

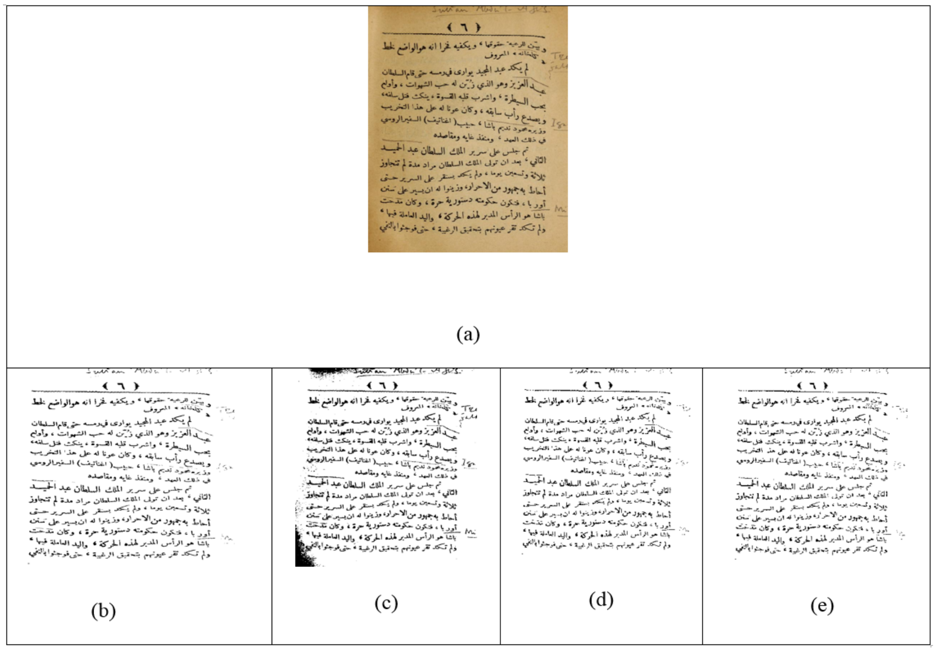


Figure 5-4 Sample warped document images image: (a) An original warped image, (b) after applying Sauvola binarisation, (c) after applying threshold (d) after applying Otsu binarisation, (e) after applying Phase-Based Binarisation

5.4.1.2 Noise removal

In this stage, we intend to remove small noisy connected components in order to simplify the following procedures of the proposed method. In the image, degradation by some random errors is called noise, and it is regarded as an undesirable. An image denoising technique suggested by Kovesi (1999) is used in this stage, which is based on the assumption that phase information is the most significant feature of images. For the output of Kovesi’s method we used Otsu’s method on the normalized denoised image, where normalized denoised image is gained by applying a linear image transform on the denoised image. This method can also eliminate noisy and degraded parts of images, because the denoising method tries to shrink the amplitude information of the noise component. The difficulty with this approach is that it misses weak strokes and sub-strokes, which means that we cannot trust its output. To resolve this problem, we combine this binarised image with an edge map acquired using the Canny operator (Canny, 1986). Researcher then compute a convex hull image of the combined image. Fig. 5-5 shows an example of this stage.

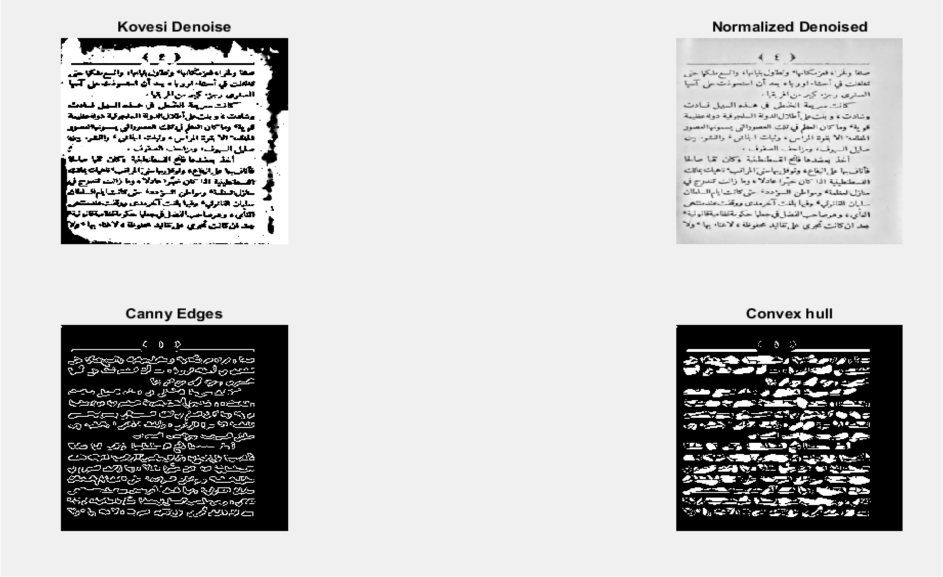


Figure 5-5 Example of the steps used in the Noise removal phase

5.4.1.3 Broken characters’ restoration

The major challenge for baseline detection in historical documents is caused by the high occurrence of broken characters. In some historical documents, certain characters are also fragmented due to the special nature of the font. In Arabic script, for instance, both printed and handwritten script is semi-cursive. Every letter is linked to the baseline with a connection point on either the right and/or left side. Faded ink can produce many broken characters. In this situation, characters and strokes may be separated into various connected components, which helps the process of text line detection since broken components are no longer linked to the baseline of the writing and become ambiguous and hard to segment into the correct text line (Rahnemoonfar, 2010). At this stage, we investigate the possibility of the existence-faded ink. If it does exist, canny operator is applied.Cohen et al. (2012) proposed technique to restoring broken Hebrew characters, using active contours with shape-prior. Researcher can apply this approach to restoring broken Arabic characters, using active contours with shape-prior.

5.4.2 Proposed text line segmentation techniques

Contrary to modern documents where the spaces between text lines are regularly while the spaces between text lines are unevenly in historical documents, most of the time, this is not the case; above all the arbitrary warping in text line usually changes these spaces. Therefore, the simple nearest neighbour joining techniques would often be unsuccessful to group complex historical document units since the adjacent neighbour of a unit often refers to another line. For instance, Figure 5-6 presents the state where ordinary nearest neighbour calculated between the bottom points of each connected component.

Due to the existence of a pencil mark, some letters will be combined together which produces a big component. However, as it can be seen from Figure 5.6, by the backing of the adjacent components, the segmented line is near to the ideal baseline.

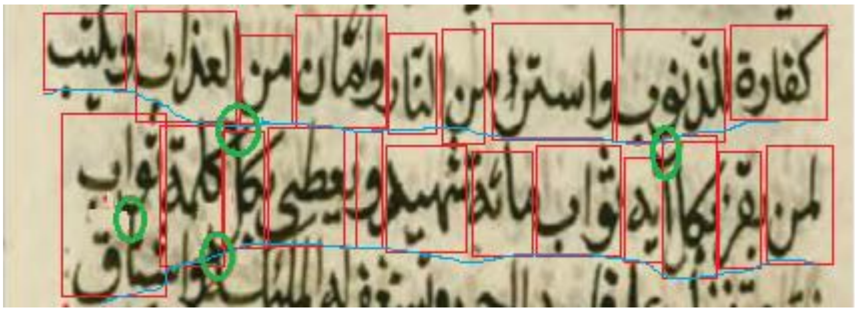


Figure 5-6: Breakdown of ordinary adjacent neighbour in segmenting text lines

Another challenging issue is the existence of touching components in the image due to dense paragraphs. However, the proposed line segmentation can resolve this problem.

Most of the suggested text line methods in the literature are based on Connected Components (CC) analysis. This type of techniques has shown some struggles dealing with low text-to-background contrast images, for example, historical documents or warped documents because of the binarisation pre-processing as discussed in (Arvanitopoulos and Süsstrunk, 2014).

The proposed process generates piece-wise linear seams that nearer to the central axis of the text lines in the document page. The method consists of two steps:

1) Medial seam computation applying a projection profile.
2) Separating seam calculation applying the adjustment of the seam carving method.

Based on this technique Daldali and Souhar (2018), that exactly performs the same procedure, to avoid difficulties, for instance, ignoring any seam not starting from the first part. The new technique suggested additional numerical and geometrical procedure, based on interline distances applying the algorithm on Fig 5-7. The technique can be used to languages written horizontally for instance Arabic (right to left) or Latin (left to right), however also to foreign languages like Japanese and Chinese manuscripts written vertically. Fig. 5-8 illustrations the results of comparing between the first technique and the new technique proposed using the same image.

Algorithm

Input:

Pieces: list of detected pieces sorted by column

Output:

Lines: list of detected lines axis approximations

```
for each piece in pieces do
        unmatched ←True
       for each line in lines do
            if line [-1]=piece[0] do
             line ←concatenate (line, piece)
           unmatched ←False
         break
     end if
     end for
    if   unmatched =True do
         lines ←insert (piece, Lines)
   end if
end for
```

Figure 5-7: Algorithm applied for gathering the piece-wise approximations to produce complete lines axis approximations baseline (Daldali and Souhar, 2018).

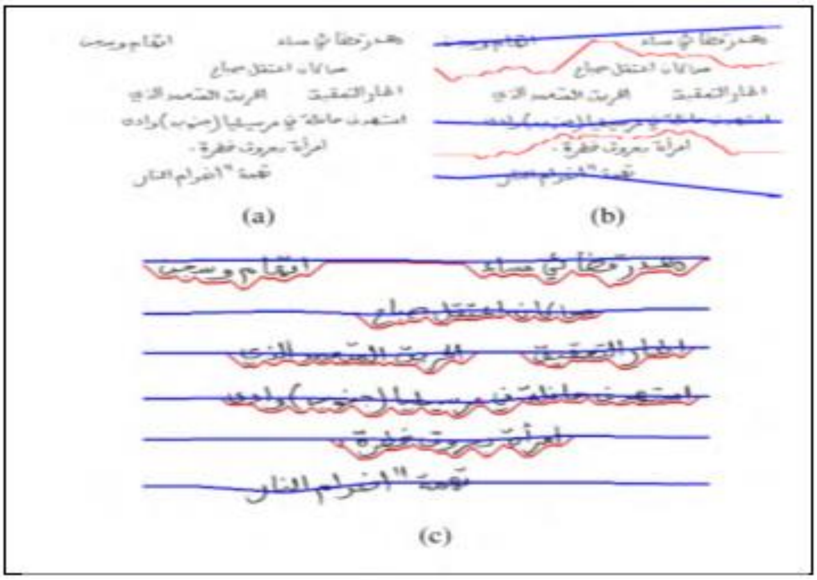


Figure 5-8:Comparison between first technique and the new technique proposed applying the same parameters: (a) the original image, (b) the resulting seams from a first method, and (c) the result applying the suggested method.

Where the seam carving technique was applied combined with a baseline technique Voronoi diagram in order to guide the seam carving technique through the inter-line space. It is very essential to find a new method for segmentation that considers the Arabic script text and connects the letter part together. That is attentive all the difficulties and techniques of handwriting Arabic calligraphy in terms of joining the letters, dots, arrangement, and others like touch characters, and overlapping. By applying the seam carving method, we were able to obtain better versatility which will offer new foundations to improve the segmentation accuracy and capability to detect more complex baseline. Seam carving, the segmentation techniques used could be considered appropriate in the context of other performance measures and to find the exact baseline (Daldali and Souhar, 2018).

5.5 Proposed Baseline Detection method

Detecting baseline is one of the major processes in the pre-processing phase of Arabic OCR system (Farooq et al., 2005). A main uniqueness of ACR is that it starts with baseline estimation. Consequently, the baseline can be applied in either Arabic text segmentation or geometric deformations extraction from the document (Al-Shatnawi, 2016). With the above-mentioned method suggested in section 5.4.2, text lines are accurately segmented, but the precise baseline is not detected and yet, one of the most important steps is baseline detection and straightness. The majority of the techniques failed to distinguish the right baseline when short characters and extensive diacritics exist. In the following section, the reliability and flexibility of a new method for base line detection will be evaluated. The proposed base line detection method not only takes into consideration unforeseen geometric distortions, but also it distinguishes specific main components of the text line.

5.5.1 Baseline detection by VD Technique

Most of the suggested baseline detection techniques in the literature have shown some struggles dealing with low text-to-background contrast images, for example, historical documents or warped documents as shown on Table 4.3. This type of techniques were unable to determine a suitable baseline when there were text lines curved and warped. To tackle this issue, a novel solution that involves assessing the native baseline, which is both flexible and has the desired accuracy, is proposed to serve as the basis for a major correction stage, namely the dewarping procedure. The suggested baseline assessment method was based on Exploited Components of Voronoi Diagrams proposed by Al-Shatnawi (2016). In this Chapter will detect baseline after VD construction process stages. Where the Figure 5-9 explain the stages of baseline estimation process stages. The pre-processing steps to be deliberated below involve document image acquiring, edge detection, labelling, the sampling of definition peaks inclusive the baseline.

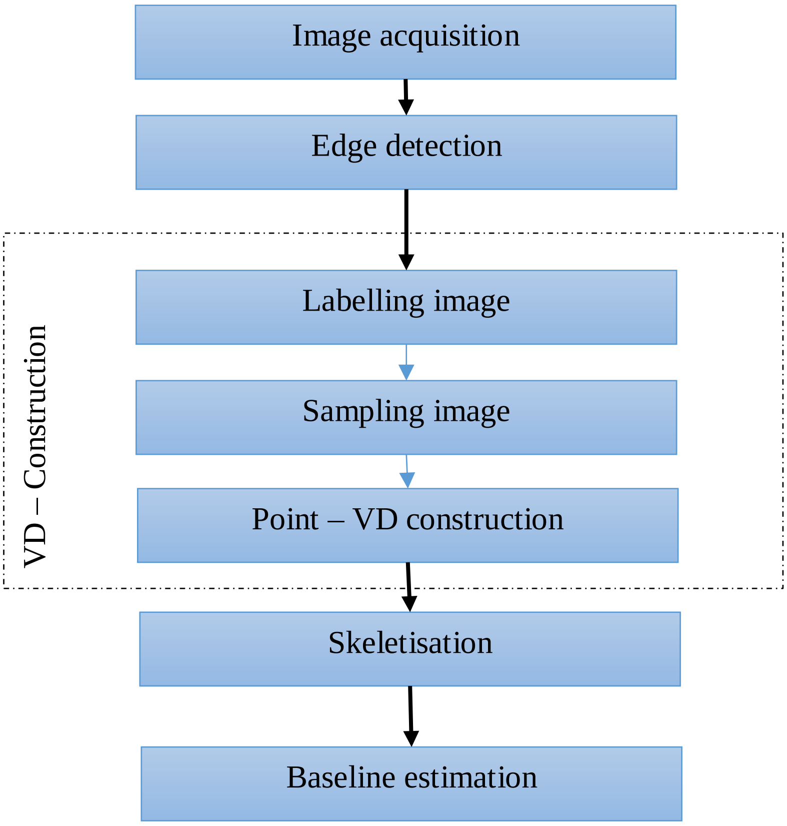
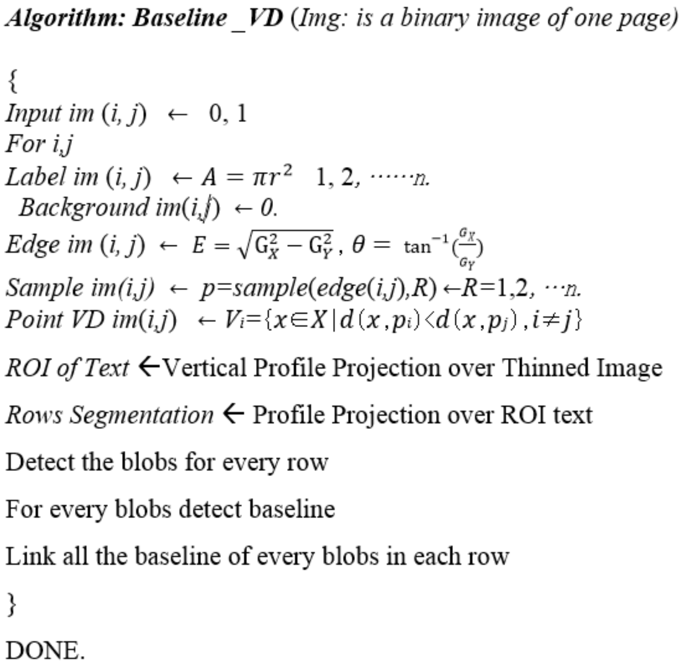


Figure 5-9 baseline estimation process stages

Figure 5.10 presents our new algorithm to detect baseline by VD

.



```
Algorithm: Baseline _VD (Img: is a binary image of one page)

{
Input im (i, j)  ←  0, 1
For i,j
Label im (i, j)  ← $A = \pi r^2$   1, 2, ⋯⋯n.
  Background im(i,j) ← 0.
Edge im (i, j) ←  $E = \sqrt{G_X^2 - G_Y^2}$, $\theta = \tan^{-1}(\frac{G_X}{G_Y})$
Sample im(i,j) ←  p=sample(edge(i,j),R) ←R=1,2, ⋯n.
Point VD im(i,j)  ← $V_i=\{x \in X \mid d(x,p_i)<d(x,p_j), i \neq j\}$

ROI of Text ←Vertical Profile Projection over Thinned Image

Rows Segmentation ← Profile Projection over ROI text

Detect the blobs for every row

For every blobs detect baseline

Link all the baseline of every blobs in each row

}

DONE.
```

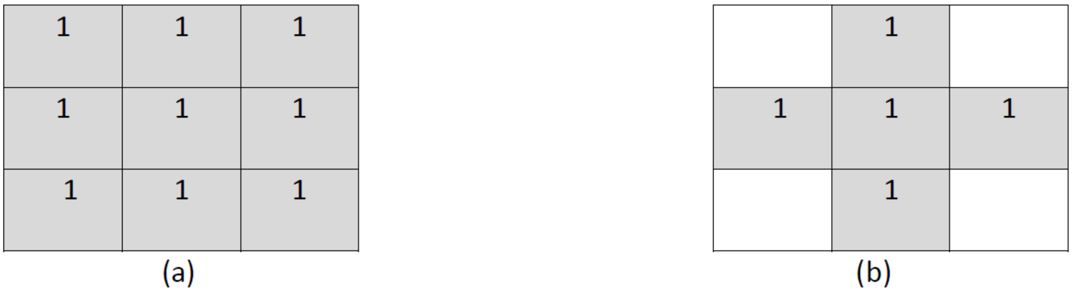
Figure 5-10: Algorithm baseline detection using VD

5.5.1.1 Document Image Acquisition

Every document image process begins with the obtaining of an image from any dataset. In this research, the images from some important libraries such as such as Qatar National Library, the British Library, and the Library of Congress have been taken. The binary format is specially applied for the processes of the document image. Edge detection, labelling, Sampling; are some instances of the binary format.

5.5.1.2 Edge Detection

The prophesied of the objects' edges of the image by the edge detection block. The pixel's positions are determined by these blocks, holding the gradient degree quite large. All image processing demands the edge detection, which that way needs factors in the adjoining regions that can be critical to the continual modifications and the areas of values in the grey (Jähne, 2002). Since sudden changes happen at the edges and only two values of pixels can be utilised (black and white), binary document images are in general processing. A 3x3 pixel mask is applied to distinguish the pixels, commonly identified as pixel neighbourhood. With regard to this method, the 8 and 4-neighbourhood masks are applied. In the above, the pixels are deemed to be adjoining when there is one combined corner at least, (north-west, northeast, or south-east), whilst in the latter, the pixels are considered to be neighbouring only when there is a combined edge (west, south, east or north). Figure 5-11 explains the 3x3 mask for 8 and the 4-neighbourhood.



1	1	1
1	1	1
1	1	1

(a)

	1	
1	1	1
	1	

(b)

Figure 5-11 The 3x3 mask personification for (a) The 4-neighbourhood and (b) The 8-neighbourhood

5.5.1.3 Image Labelling

The procedure labelling CCs was a pledge by subsequent the CCs' points for consecutive images beginning from the left side and moving across the right. Pixel connectivity for a 2D image has distinguished a linked in pixel neighbourhoods. A normal rectangular inspection model produces a limited mathematics grid {(x,y): x = 0, 1, ..., X−1; y = 0, 1, ..., Y−1} propping advanced images allows the description of two types of the neighbourhood including a pixel. The initial is a 4-neighbourhood that includes only the pixels up, under, to the left and to the right of the centric pixel (x,y). The secondary is an 8-neighbourhood that combines to the 4-neighbourhood four diagonal neighbours (Ramdan et al., 2016). Figure 5-12 exhibits 4 and 8 neighbourhoods of pixel connectivity for 2D document images.

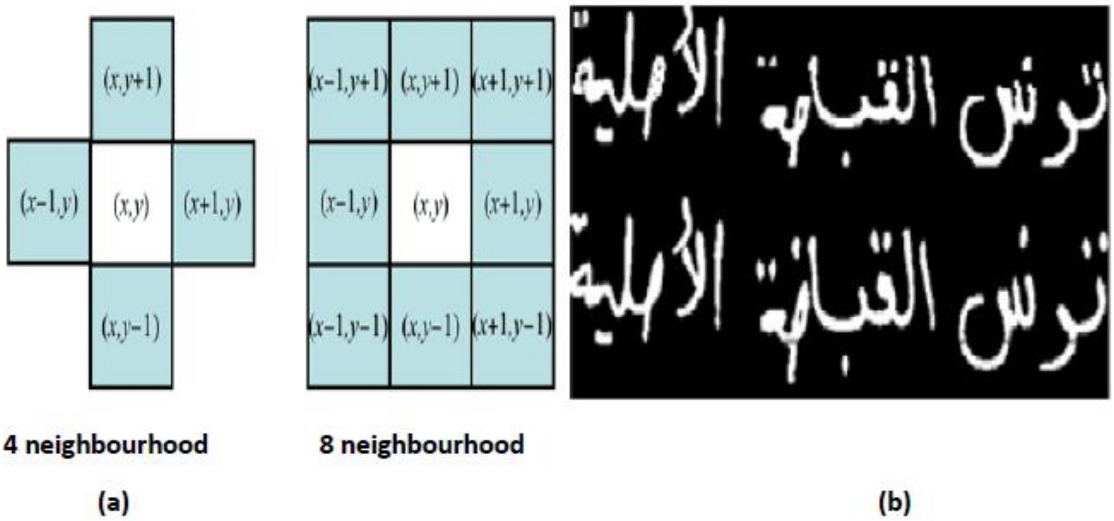


Figure 5-12 a- Presents 4 and 8 neighbourhoods' connectivity and b- label of connected component

Starting from number 1 to n, which performs the last number of the last CC for every image, by applying the 8-neighbourhood description for pixel connectivity a CC label can be produced as given in Figure 5.12(a). Figure 5.12b) illustrations CC labelling for the document image.

5.5.1.4 Sampling Points

There is a directly proportional relationship between VD building and the consuming of memory and time (Berman et al., 1999, Gold and Snoeyink, 2001). Consequently, decreasing the number of the mentioned generators can be an optimizing procedure without missing the information, which would, in the outcome, speed up the entire process. We can select points on the word contour for the improvement. Zaki (2008) proposed the technique of sampling applied in this process with regard to the description of the various mistake. We attain the text image over the external and the internal regions, later to be applied as input. With regard to the first step of the process, we perform scan column-wise from left to right of the document image. The outcome is a contour tracing in a clockwise trend, and every other R^{th} pixel is sampled.

This selection procedure is continued until the contour series ends, where R is a sampling parameter, R = 1, 2, 3… n. If R = 1, all pixels of the contour are chosen.

Figure 5.13 describes the Arabic handwritten document, prior and next to the sampling procedure. It impacts the VD framework and speed of building(Zaki, 2008).

Algorithm: Sampling of Arabic text algorithm.

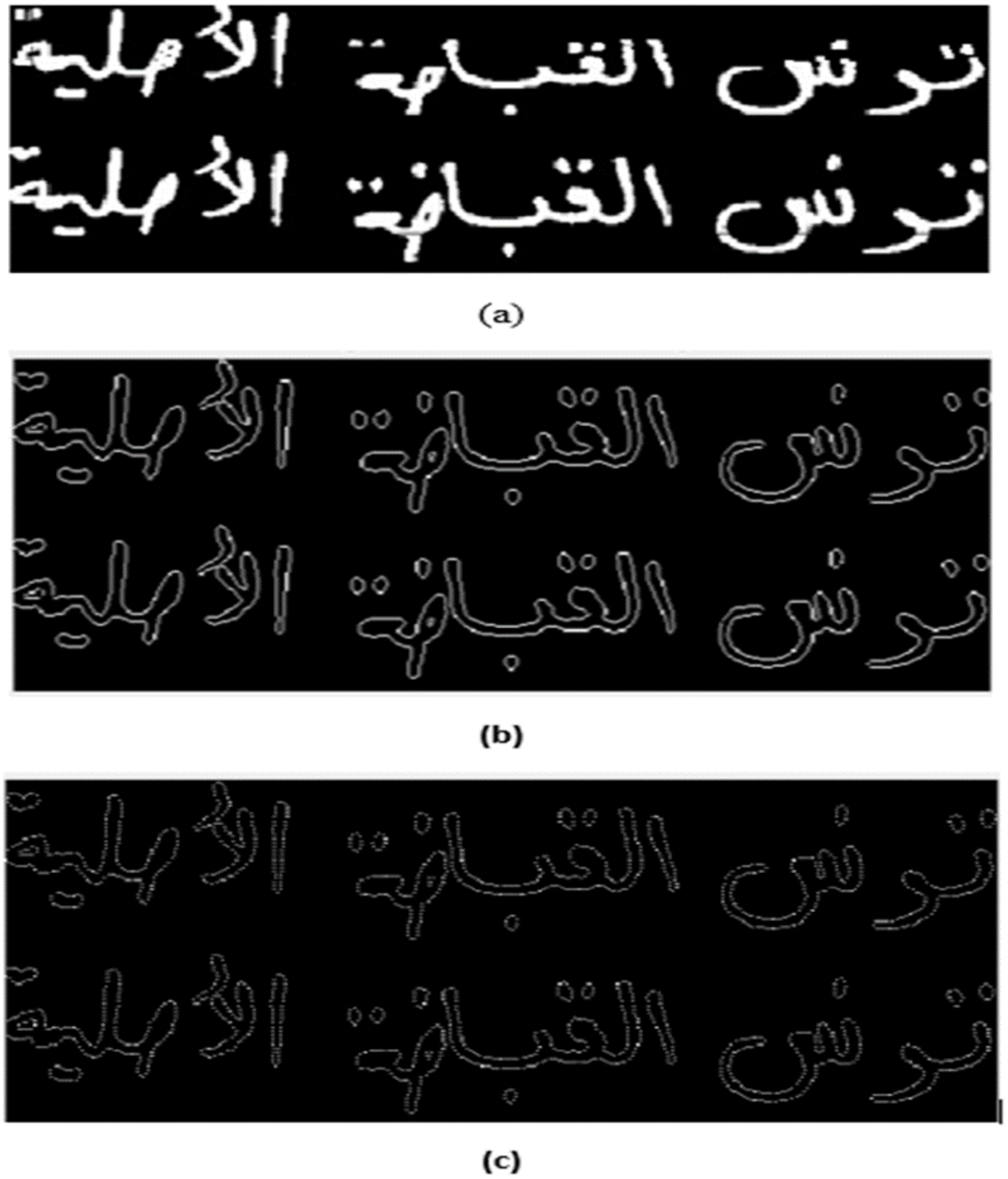
```
  for i = 1:image_width

 for j = 1:image_height

  if cell[i % x-axes] [j % y-axes] > image[i][j]


 binary_image (I,j) = 1;
          else
               binary_image (I,j) = 0;
        end if
end for
end f or}
```

Figure 5-13: Sampling of Arabic text algorithm(Zaki, 2008).



(a)

(b)

(c)

Figure 5-14 the sampling process: (a) Arabic handwriting document, (b) the document is being traced using the 8-neighbourhood contour representation, (c) the document is being sampled from the boundary using R=4

The construction of VD is time-consuming. On the average by selecting the sampling period R to the sampling procedure, the processing time is decreased and the time consumed by all boundary points are applied to generate VD.

5.5.1.5 Point-VD construction

Construction VD fundamentally rely on the count of generators. Many algorithms for VD structure have been suggested in the literature (Zaki, 2008).

This stage includes the construction of VD using Divide-and-conquer algorithm. The specifics of VD development procedure will be demonstrated later. The sampling points are applied as generator points for the improvement of VD.

The divide-and-conquer algorithm (DAC)

Due to the impact of divide-and-conquer model, the major algorithm, it is thought to be the preferable algorithm, in theory speaking. The generator points (S) are divided into two sets in this algorithm. Both the left S (L) and right S(R) are of the same size. After that, with left groups, V(L), and right groups V(R), the VD of both are computed. In the final merge stage, the set B (L, R) of all Voronoi edges of V(S) is calculated. Several consecutive steps follow that are in demand to implement this algorithm completely. These are presented below:

First Step: the x-coordinates of the points are systematised in ascending order, so it is easier to separate them into left and right divisions.

$$S = \{P_1, P_2 \dots \dots \dots . P_n\} \tag{5.3}$$

Where, S is the group of generator points and n is the number of points.

$$T = n/2 \text{ , So } S(L) = \{P_1, \dots \dots \dots . P_T\}\text{, and } S(R) = \{P_{T+1}, \dots \dots \dots . P_T\} \tag{5.4}$$

Second Step: The generator points are divided and put into two set in this step**.** After that, by enabling use of any known algorithm, the VD of every group V(L), V(R) is created. Figure 5.15 demonstrate the $V(L)$, $V(R)$ graphs.

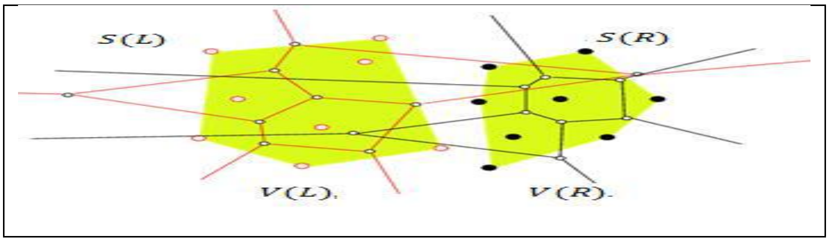


Figure 5-15 VD prior the joining step

Third step: This is one of the most important steps. In this step, the two graphs are joined to produce the final diagram. Figure 5-16 illustrations VDs merge step.

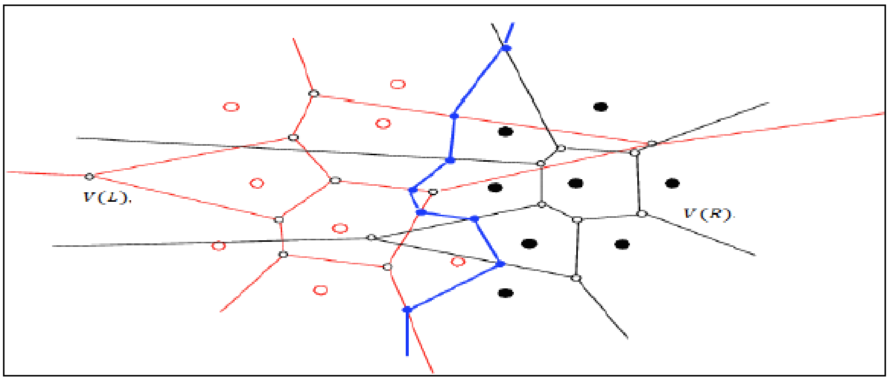


Figure 5-16: VDs next merge step

In this research, use the DAC algorithm to create area VD. Figure 5-17 depicts created area VD for left and right side for Arabic handwriting document.

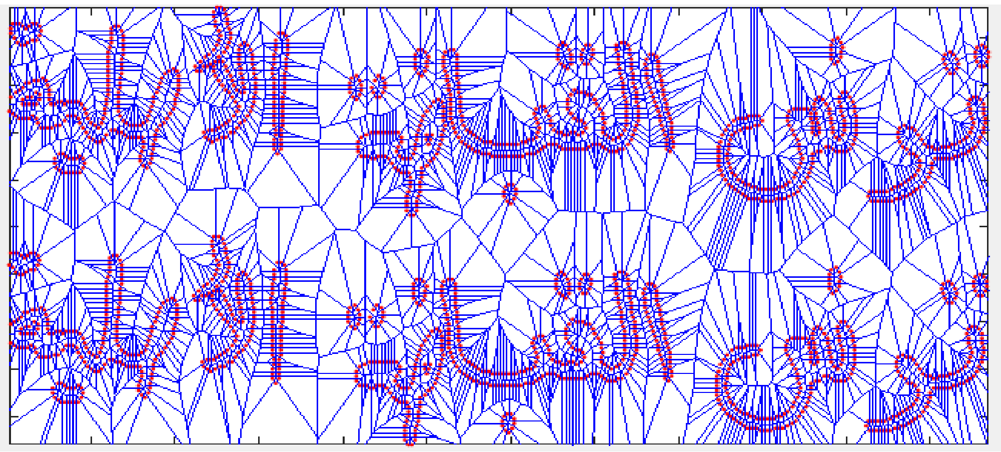


Figure 5-17: The created area VD for left and right side

5.5.1.6 Skeletonisation

The skeleton of any form supplies the fundamental information about it. Thinning- axis extraction or skeletisation indicates the procedure of extricating the skeleton of a form. Thinning "Skeletonisation" is a very important phase in the Arabic Character Recognition (ACR) technique (Al-Shatnawi et al., 2014). Abu-Ain et al. (2013b) a new skeleton algorithm is suggested solved the difficulties with previous techniques. The method contains of three steps. First two steps concerning reproduce the skeleton and the third is counted for improving the skeleton into one-pixel width. The framework of these steps directed in Flowchart as presented in Fig 5-18. Al-Shatnawi (2010) suggested the effective skeleton method on the exploited VD vertices. Algorithm thinning method based on the exploited vertices of VD offered in Fig 5-19. Figure 5-20 representation the resulting skeleton from a document image in figure 5-17.

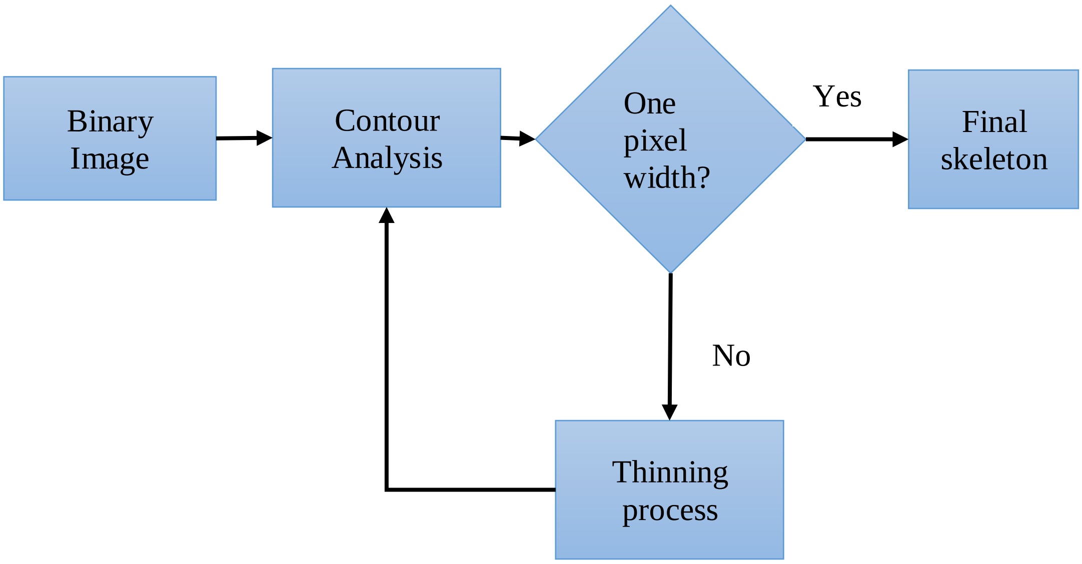
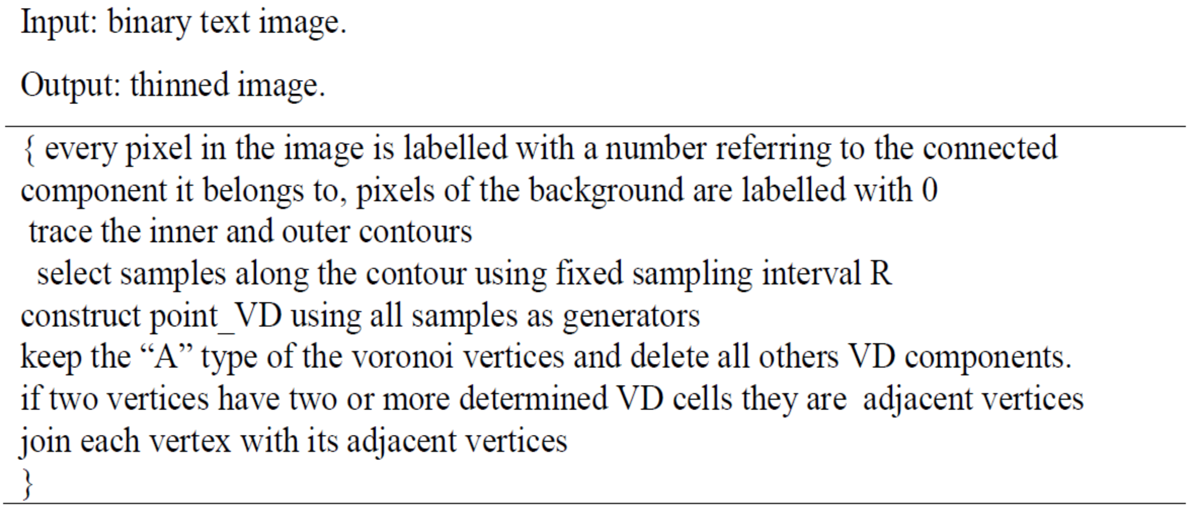


Figure 5-18 flowchart of the skeleton technique (Abu-Ain et al., 2013b).



Input: binary text image.

Output: thinned image.

```
{ every pixel in the image is labelled with a number referring to the connected
component it belongs to, pixels of the background are labelled with 0
 trace the inner and outer contours
  select samples along the contour using fixed sampling interval R
construct point_VD using all samples as generators
keep the "A" type of the voronoi vertices and delete all others VD components.
if two vertices have two or more determined VD cells they are  adjacent vertices
join each vertex with its adjacent vertices
}
```

Figure 5-19: Algorithm thinning method based on the exploited vertices of VD (Al-Shatnawi, 2015).

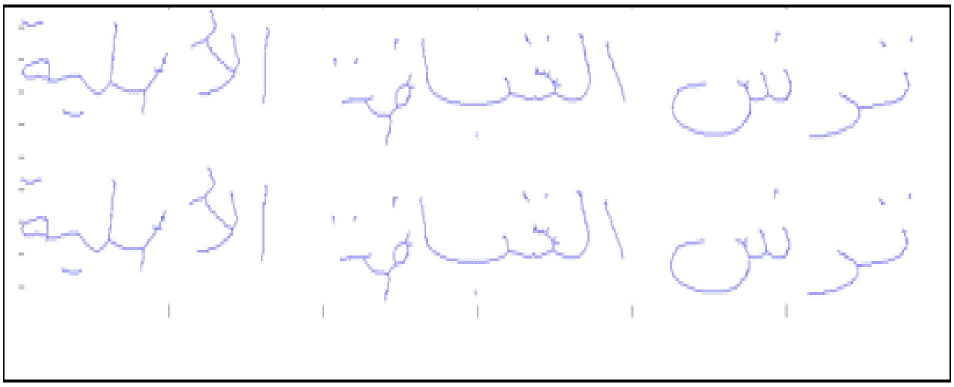


Figure 5-20 the resulting skeleton from a document image in figure 5-17.

5.5.2 Baseline Estimation

The Voronoi generators are the sampling points which are produced from the text contour tracing. The VD is built from those selected generators. The suggested baseline based on VD estimation procedure determines the baseline in three steps. After determining skeletisation- axis or thinning from Binary Original Image, dilate the binary Image. In the first step, assign each blob a different colour to visually show the distinct blobs. Then, in the second step, pseudo-coloured labels, Blobs are numbered from top to bottom, then from left to right. Extract each blob and find the bottom of it. Finally, detect the blobs for every row, for every blobs detect baseline and link all the baseline of every blobs in each row. In the third step, determine "good" rows, extend the first good row out to column. Based on middle rows we determine the baseline. In this research, where the baseline is determined to be subordinate on the joined components, it is always more than one text line. Figure 5.21 has depicted the baseline as the red line for Arabic handwriting document.

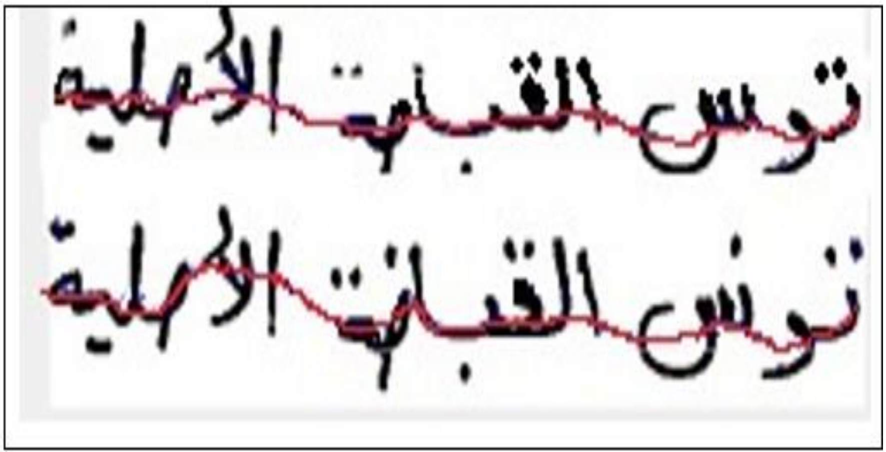


Figure 5-21 The baseline as the red line for Arabic handwritten document.

5.5.3 Delaunay Triangulation (DT)

VD has its dual tessellation called the Delaunay Triangulation (DT). Triangulation is the division of the convex hull of a given set of points S into triangles (where S points are assumed to be the Voronoi generators). In the Preliminary Stage we will use DT to convert the text into convex hull that connects the sampling points. After constructing DT on Voronoi Sites (generators), a lot of common properties for VD and DT from their association can be noted. DT has some additional properties and has finite edges unlike VD edges, which can have infinite value. So, having a finite value and also being a planar graph, the number of edges and triangles can be related to the number of sites mathematically. The formal definition is given below (Okabe et al., 2009)

Let $V(P)$ be a VD generated by a set of n distinct points :$P = \left\{p_{1,...,}p_n\right\} \subset 2(3 < n < \infty)$ that are not on the same line (non-collinearity assumption);$Q = \left\{q_{1,...,}q_n\right\}$ is the set of Voronoi vertices in V(P): $X_{iI,......,}X_{ik}$ are the position vectors of the generators whose Voronoi share vertex q_i.

After creating DT on Voronoi Sites (generators), a lot of standard features for VD and DT from their association can be seen in addition to the VD properties which were already introduced and explained at the top of this portion. DT has some extra features and has finite edges, unlike VD edges which can have infinite value. The number of edges and triangles can be related to the number of sites mathematically if there is a finite value and also being a planar graph. They are related as if n is the number of sites then DT will have $3n$-6 edges and $2n$-5 triangles. Through the other important attributes of VD and DT are:

- Every Voronoi edge matches to an edge in the DT.
- The convex hull of sites creates the border of DT.
- The minimum angle overall triangulations maximized by DT.

Figure 5.22 shows VD and DT generated of the same set.

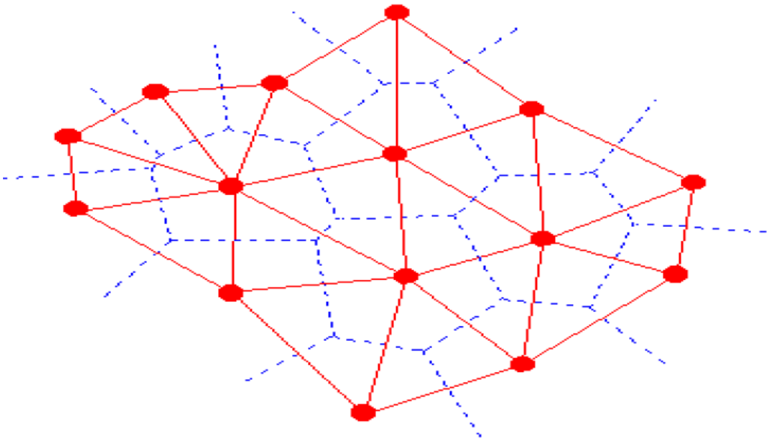


Figure 5-22 Generator set (VD and DT). DT represented by sold lines, while VD is represented by dotted lines.

.

The above introductions may result in a mistake that DT is post processing of VD and VD is compulsory. Yes, there is a big interdependence between VD and DT, but this does not indicate that DT cannot be explained or applied alone without the use of VD. DT was mandatorily used in many research for various implementations (Al-Shatnawi, 2010).

5.6 Dewarping

The methodology for the accurate baseline detection was exhibited. In this part, the suggested method for dewarping, along with the tests guided to the appropriate solution, will be debated. Determining the geometric deformation in a document by accurate baseline modelling enables us to convert the document locally to the identical flat page. In the next part, three techniques for partitioning the document image and transforming each partition to the identical flat parts are offered. The first two techniques do not have coveted flexibility. It is the last suggested method based on VD Technique which have been used in the final dewarping system.

5.6.1 Dewarping based on strip modelling

To separate the documents to some small parts and convert each partition to the identical object partition, in the first trial, the baseline is criss-cross with the lower and upper borders; the upper and lower borders are determined to be parallel to the baseline with the distance H_{max} and $-H_{max}/3$ sequentially, where H_{max} highest is the component height of letters. Now for producing each strip accurate, pieces are separated into some divisions count on the curving of the baseline as shown in Figure 5-23. In other words, despite the word segmentation, if a collection of letters lie on a text line with regular slope, they connect to the same box; but not in the case of tendency difference

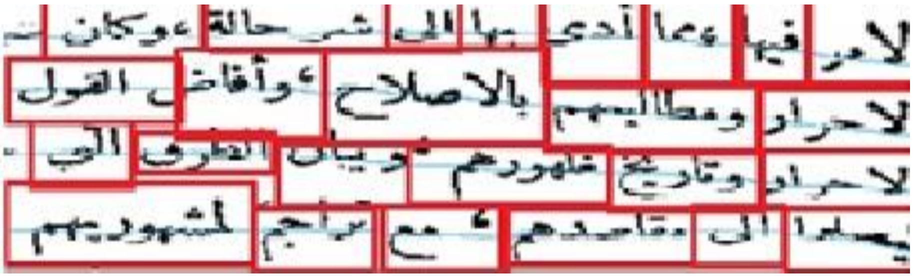


Figure 5-23 Separating document to several boxes

At the next stage, each box will be rotated by the slant of the box round the bottom left of the box and then all of the rotated boxes relevance to a particular text line will be vertically transposed in order to get the parallel alignment. This technique is doing well when text lines are parallel and there are sufficient spaces among text lines. However, in the situation where two consecutive baselines are not parallel, two pieces matching to two various lines may cross (Figure5.24). In this case, letters intersecting over two neighbouring slices will be converted with two various transformations and will appear irregular. Because this technique does not have proper flexibility and precision, it has not been applied in the final dewarping system.

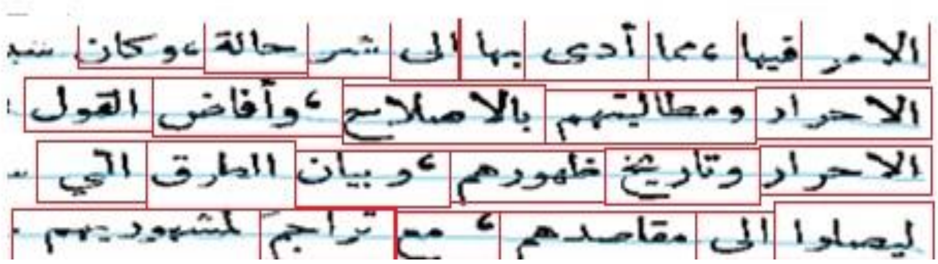


Figure 5-24 an intersection of two piece

5.6.2 De-warping by means of baselines and top lines

In the next endeavour, we decided to get top lines in addition to baselines to be able to determine more specific pieces. Consequently, it is important to detect top lines autonomous of baselines and convert both baselines and top lines to the analogous flat lines. In this approach, top lines and baselines would be parallel in the final reconstructed image, dewarped image, and this will resolve the difficulty of letters contraction automatically (Figure 5.25). In other words, in typical cases top lines and baseline should be parallel, and it is only due to the letter distortion that baseline and the top line may not be parallel in some pieces.

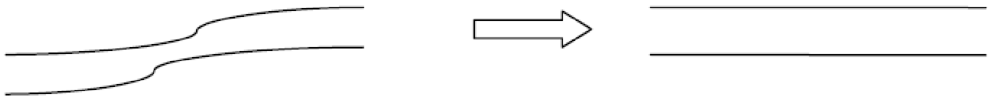


Figure 5-25 baseline and top line prior to and after dewarping

To detect top lines, the analogous technique of the suggested baseline detection method by Piecewise Linear Curve Smoothing (Rahnemoonfar, 2010). However, in this situation in state centric bottom points of components, central top points of them are selected and chosen of upward viewing angularity, the downward viewing angles are calculated; consequently, the in demand angles will be computed according to the equation 5.5.

$$\varphi = \begin{cases} \theta & if\, f'' \geq 0 \\ 2\pi - \theta & if\ f'' < 0 \end{cases} \qquad 5.5$$

Figure 5.26(a) presents the outcome top lines detected with the mentioned algorithm. As it can be observed from Figure 5.26(b), due to the distortion of some letters, top line and baseline are not parallel. By converting top lines and baselines to the identical flat lines, in the final de-warped image top line and baseline, would be parallel and distortion due to the retraction as well as arbitrary warping can be fixed.

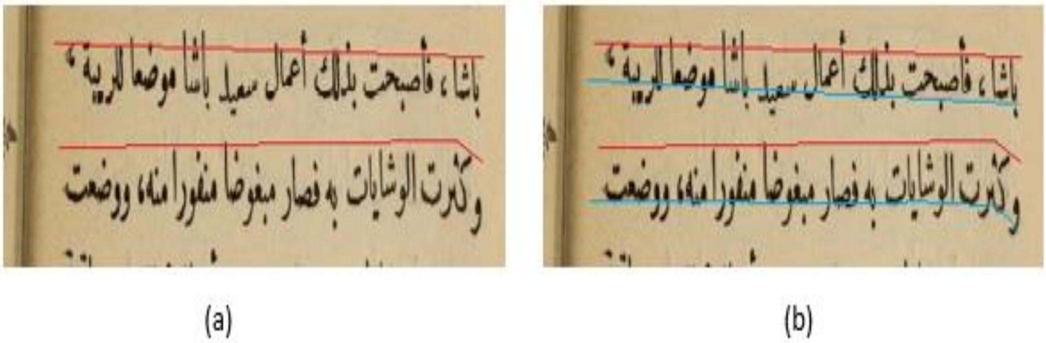


Figure 5-26(a) top line discovered with the suggested technique, (b) non-parallel baseline and top line due to the retraction

Presently there cannot be any perfectly independent and credible automated procedure for detection of the top lines due to the complex nature of the top line in Arabic scripts. Firstly, the number of ascenders is more than the descenders in Arabic; for example, in Arabic alphabet, there are only three characters with descender which do not sit on the baselines, while the number of letters with ascenders is more. Second, there are many additional diacritics in the top line such as accents, dots, etc; and third, the characters join by taking their relevant form, most of the Arabic letters are either dual-joining or right-joining, add a huge complication to the detection procedure of the exact place of the top line.

In general, considering baselines are not parallel too, there should be an algorithm to paradigm smooth differences between lines; for example, an algorithm should take care of reduction without going through the procedure of top line detection. This part will be improved further in the following section.

5.6.3 Dewarping by means of Grid-Based Modelling

Based on the accurate baseline detection introduced in the former section, an effective grid-based technique is proposed by Yang et al. (2017) to geometrically model and correct arbitrarily warped historical. The dewarping method with the global grid in this technique applies a conversion model to repair individual quadrilateral sub-grids (local meshes) to a rectangular form. While there are various obtainable transformation models for form restoration algorithms, for example, affine transformation, perspective transformation, bilinear transformation etc. the affine transformation is appropriate to be used in our situation since it can sustain the straight lines and the proportions of distances through the transformation procedure. The vertical position of the target rectangles top and bottom line is defined by the medium vertical status of identical rows in the mesh. The global grid building, and the simultaneous transformation of sub-grid can efficiently address historical scripts with the complicated layout. The horizontal situation of the goal rectangles left and right would be preserved at the horizontal status of the quadrilateral sub-grid. The elaborate steps in this technique are below:

- **Pre-process:** In preparation for the following procedures, connected components inside the grayscale (deformed) input image are classified, based on a criterion labelling approach applying pixel connectivity. Additionally, particular noise (small black parts) is refined out.
- **Global grid production:** First, the page is separated into areas (zones) applying a bottom-up technique based on local characteristics. The outcome is then labelled as text or non-text using a classification procedure that takes advantage of the same features. Lastly, vertical and horizontal dividing lines are specified and applied to improve the recognized leaf layout. This stage is particularly useful for documents with complex layouts, for example, newspapers and periodicals. Figure 5-27 displays result of the page analysis step for an instance document. The global grid is subsequently created based on the specified text regions.

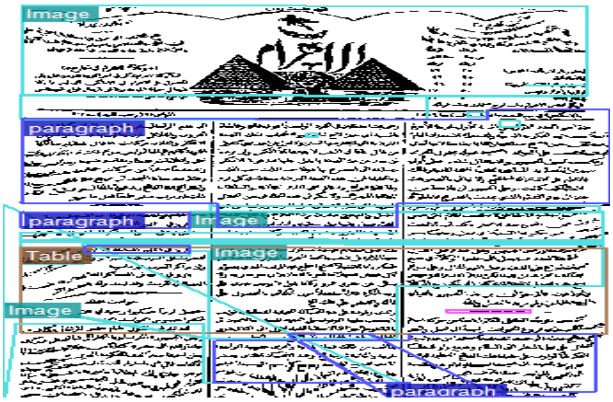


Figure 5-27: Separation into regions

- **Text Line Segmentation:** As a requirement for obtaining geometric deformities, text lines need to be detected. Hence, devoted text line segmentation algorithm by Daldali and Souhar (2018) is applied to the text ranges of the global grid.

- **Mesh production:** It firstly generates a primary rectangle- based mesh in terms of the medium width of connected components; and then discovers the nearest connected components of each point in the mesh and adapt its location accordingly.

- **Dewarping:** it applies a transformation paradigm to rectify individual quadrilateral subnets (local meshes) to a rectangular form. At its first step, the suggested dewarping technique bear in mind both global and local properties of the document image and we separately apply regression analysis and Root Mean Square Error (RMSE) measurement techniques to recognise outliers and correct mesh. The first stage is to rectify some local obscure point in the mesh by applying RMSE measure. Presented a raw mesh $M = \{a_{ij} : i \in (1,2, \ldots\ldots R_g), j \in \{1,2, \ldots.C_g\}\}$, the location of each point a_{ij} are indicated as (x_{ij}, y_{ij}).

  Where; R_g is the radius of gyration

  C_g is the centre of gyration

  For every row line i, a linear polynomial equation can be denoted by equation (5.6).

$$y = a_i * x + b_i \tag{5.6}$$

The quality of appropriate equation 1 is computed by applying RMSE measure. If RMSE is big, the row line i has a poor adjustment to the linear polynomial equation, so this row line may have many outliers. For every point (x_{ij}, y_{ij}) of this row line i, the procedure by equalisation is shown in equation (5.7).

$$y_{ij} = \begin{cases} a_i * x + b_i, & if \quad RMSE \quad < \partial \\ y_{ij} & if \quad RMSE \quad \geq \partial \end{cases} \tag{5.7}$$

In equation 5.7, the factor ∂ is applied as a benchmark of RMSE to fix the local outliners in mesh. Fig.5.28. (a) and (b) presents a specimen document image being fixed out of the first stage.

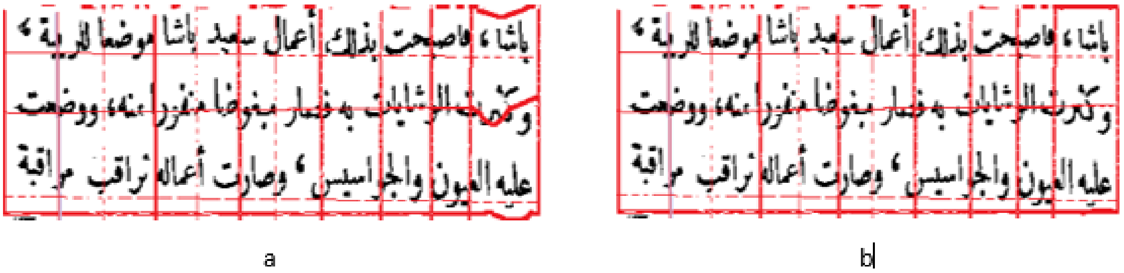


Figure 5-28: Mesh modification by RMSE measurement. a) Mesh with local obscure details. b). Outlier rectification by RMSE measure

The second stage is to rectify some global outliners in the mesh by applying regression analysis and linear polynomial fitting techniques. Firstly, it applies the average tendency of each row line for distinguishing the outliers from inaccurate text line segmentation. For every row line i, the total of slope M_i can be measured by equation 5.8.

$$M_i \ = \Sigma_2^{C_g}(|y_{i,j} - y_{i,j-1}|/|x_{i,j} - x_{i,j-1}|) \qquad (5.8)$$

The medium means of slope M_i of all row lines in manuscript zone can be computed as:

$$M_A = (\Sigma_2^{R_g} M_i)/(R_g - 1) \qquad (5.9)$$

Regarding the global impact of page curl, the manuscript with page curl is supposed have a nearby M_i for each row line. The domain of these row lines with nearby M_i is indicated as [a, b]; the row lines with distinct M_i are achieved by outliers from text line segmentation. Furthermore, the manuscript with arbitrary warp is assumed to have a M_A close to zero, which is usually minimal than the M_A of the manuscript with page curl. With regard to rectifying these outliers, for every row line i, third-degree polynomial equation 5.10 can be applied to adapt these row lines with near M_i.

$$y = A_i * x^3 + B_i * x^2 + C_i * x^1 + D_i \qquad (5.10)$$

Accordingly, for each point $(x_{i,j}\ ,\ y_{i,j})$ in this row line i, the rectification process is executed by equation (5.11).

$$y_{i,j} = \begin{cases} A_i * x^3 + B_i * x^2 + C_i * x^1 + D_i & if \;\; M_A \geq \beta \\ y_{i,j} & if \;\; M_A < \beta \end{cases} \qquad (5.11)$$

In equation 5.11, the parameter β is applied as a benchmark of M_i to rectify the global outliners in the grid. If there are some new potential obscure points generated by (5.11). A linear polynomial equation is assessed for each row line by applying the least-square algorithm. If an obscure point is recognised, its vertical location is substituted by the value generated by the linear polynomial equation.

Presented a linear polynomial equation by a row line of mesh:

$$y_{ij} = a * x_{ij+}b \qquad (5.12)$$

For every point (x_i , y_i) of this row line, to checked:

$$y_i = \begin{cases} a * x_i + b & if \; |y_i - a * x_i + b| \;\; \geq \delta H_c \\ y_i & if \;\; |y_i - a * x_i + b| \;\; < \delta H_c \end{cases} \qquad (5.13)$$

Where:

H_c : Medium height of elements in this zone.

δ : Parameter to rectify obscure point, from 0 to 1.

Figure5-29 (a) and (b) presents a specimen document image being rectified out of linear polynomial curve fitting.

The transaction with a massive volume of observed data with a low correlation, the curve fitting procedure can professionally filter the obscure points and modify them. However, in some instances, experiences demonstrate that the correlation coefficient of the marked points in each row line is sufficiently elevated. Here, the exploitation of the curve fitting method to facilitate all points of every row is not required. Figure5-30a) and b) shows dewarping result of a sample warped image by grid-based modelling.

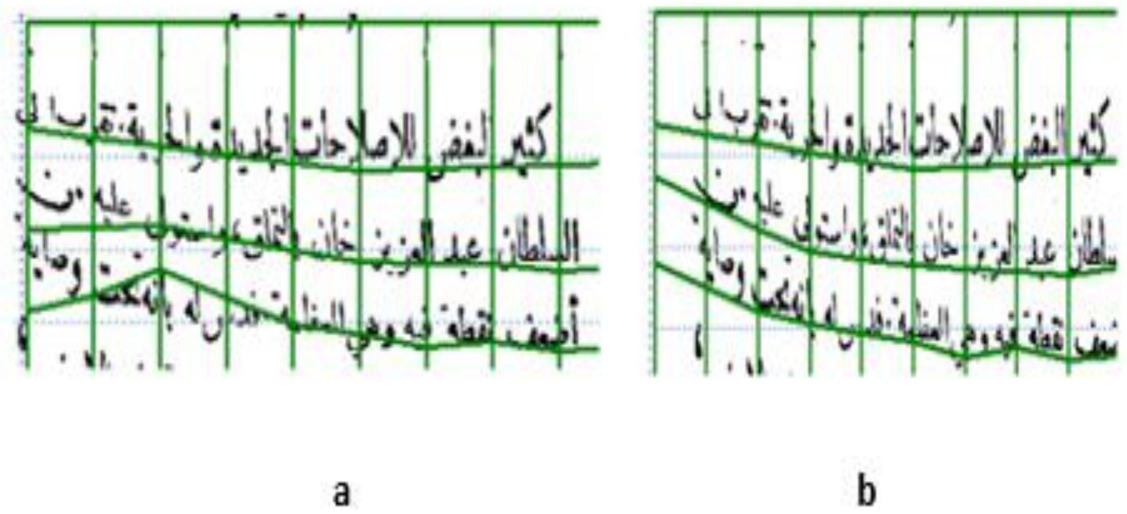


Figure 5-29Mesh modification by linear polynomial curve fitting. a) Mesh with global obscure points. b) Outlier rectification by linear polynomial fitting

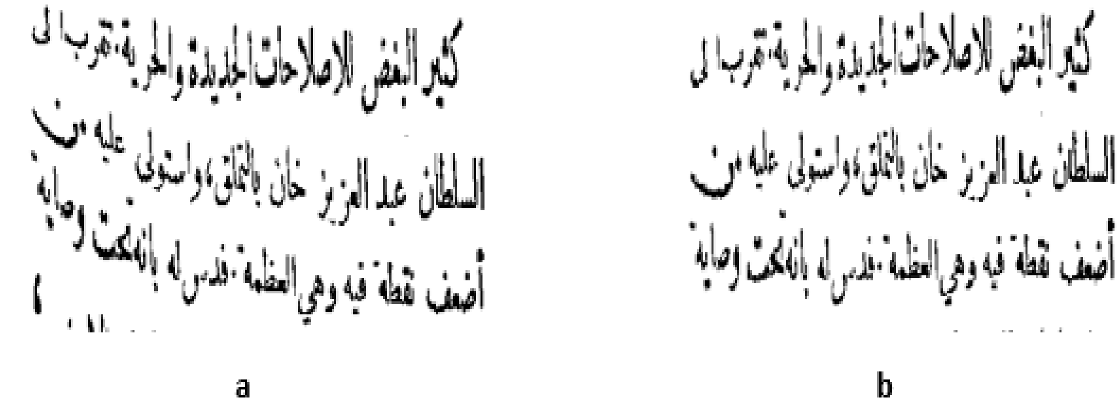


Figure 5-30: Dewarping by grid-based modelling. a) Original image. b) Dewarped image.

5.7 Conclusion

In this chapter, the System for dewarping Historical Arabic documents was proposed and explained in detail. The aim of the proposed system is to address the problem of geometric correction of Arabic historical documents and to overcome the weaknesses of previous methods. The proposed system consists of five stages: (1) Pre-processing steps (binarization, noise/dots removing), (2) Text line segmentation, (3) baseline estimation based on VD, (4) Dewarping Stage, and (5) evaluation Stage. The theory of baseline estimation based on exploited components of Voronoi Diagrams has been highlighted. The factors and equations used for this procedure have been explained, and the results of each part have been shown.

Chapter 6 - Evaluation methodology

6.1 Overview

In the former chapter, a dewarping system for historical Arabic documents was proposed. Despite the assortment of dewarping methods in the literature, no typical methodology presents for the of their performance; most of the valuations done so far focus on the optical of output and the input image (Brown and Tsoi, 2006). Moreover, in some situations, the recognition average of an OCR method is used for the evaluation of dewarping systems (Kim et al., 2015). In this manner, the rise of the recognition rate which OCR obtains between the original image and the dewarped image indicates the refinement of the dewarping technique with reference to the original image. The development of OCR as execution valuation metric highly depends on the implementation of the OCR itself. However, in the instance of historical manuscripts, OCR doesn't produce satisfactory results.

Four methodologies for the performance evaluation of the dewarping system are introduced in this chapter. The purpose of a restoration algorithm determines the quality of the performance of a restoration technique. Visual perception is still an important factor in evaluation, due to the visual quality of the output image is important for prints on demand objective. Visual inspection and final OCR accuracy depend on the line straightness and the flatness of the dewarped image. In this chapter, various supervised and unsupervised evaluation techniques are introduced for estimating the flatness of the final image. Lastly, a state-of the art system and commercial software are presented for the comparison target.

6.2 Visual Decision

The significance of geometric distortion correction for modern documents is placed on obtaining the printed script and it is recognisable by the traditional OCR. However, for historical documents, the purpose of digitisation is to create digital copies that represent the original printed content as faithfully as possible.

Specifically, while libraries allowed to make important works accessible online to more individuals, there is also a developing business sector for book reprints and in addition prints on request. Amateurs, different libraries and researchers, and in addition intrigued clients who look for them for their stylish esteem alone, will purchase reprints of scarce books and groups for their own usage. A great-quality copy can also mean a higher deals cost for libraries. Accordingly, it is essential that libraries should have high-quality copies with no geometric deformities. In this status, the visual ruling can be a standard for evaluating the image property.

Although visual experience could be a subjective method of evaluation, it still attains as a useful point for historical manuscripts in the prints on request.

Visual compare of input and output model is often considered a validation criterion in the scientific research. However, to support the observer in best evaluating the output models, in this research, a grid is set up on the image.

The straightness of text lines is the most significant agent which impacts both visual impression and also final OCR recognition rate.

In the subsequent sections, various supervised and unsupervised evaluation techniques are presented for determining the straightness of text lines in the final copy.

6.3 Supervised evaluation

Based on a presumption that the ‘right’ answer is known the supervised evaluation is operated. In the following, primarily it will be reviewed how the ground truth data are presented and then various supervised evaluation techniques are suggested accordingly.

6.3.1 Ground-truth data

The supervised performance evaluation is the mission of estimating how the real output of the method differed from the ideal output. The typical output describes the ground truth. The ground-truth is created manually by a human. For instance, in the state of evaluating a segmentation technique, the ground truth is presented by drawing of accurate image borders, or in the instance of evaluating OCR techniques, this is the ideal text which is typed by one or two labourers. For dewarping ground-truth in this research, one accessible tool for geometric rectification in PRImA research group sophisticated for the IMPACT project is applied (IMPACT, 2012). Implementing the whole step of dewarping manually, inclusive the line detection and image transform, is time manual portion of the system is text line drawing. For this intent, the user must first demarcate the four corners of a dewarping region in the image; then a grid is determined inside the region where the parallel lines of the grid are drawn by the user. Having specified the grid, each part is returned to its initial oblong shape.

The procedure of ground-truth is an effective and less time consuming procedure of producing a typical image, because it is manually created by Aletheia (Clausner et al., 2011a). Aletheia supports the user with a number of automated and semi-automated tools to produce a dewarped ground truth for an image.

6.3.2 Projection profile evaluation technique

Vertical projection profile of image is a reference for allocation of text lines in the image whilst text lines and space between them correspondingly, representative by peaks and troughs in the histogram. As it is demonstrate in chapter 3, vertical projection profile is the total of density value of a pixel in horizontal lines, vertical to the y axis and determined by: $I_{VP}\ [i]=\sum_{j=1}^{w} I(i,j)$ (6.1)

The measure of correlation coefficient between the histogram of the dewarped image and the ground truth results is one path to evaluate a dewarping technique. The concept is that if the image is warped, it is hard to detach text lines and the vertical projection profile will not be clear. Correlation coefficient estimates the force and the direction of a relation between two data strings. It is determined by the next term:

$$Corr(A,B) = \frac{Cov(A,B)}{\sigma(A).\sigma(B)} \qquad (6.2)$$

Where

$$Cov(A,B) = E[(A - \mu_A)(B - \mu_{AB})] \qquad (6.3)$$

In the above equation, A and B two data series reference how much two variables alteration together and Cov (A,B) is the covariance between them while the measure of variability of data is the standard deviation σ of a data series.

The correlation coefficient scales among +1 and -1. The greatest value (+1) reference a typical positive correspond between two data collections and it illuminates that two collections of data rising and drop together by comparable amounts. The nearer to one the correlation coefficient, the more similar are two data collections. The smallest value of -1 reveals that the two data collections progress in adverse directions, one growing when the other is reducing. The rate of zero mentions that there is a random association

between two groups or that is to say two data collections are uncorrelated. The position of text lines and the distance between them in the document image shows through peaks and troughs in the histogram, in the situation of the vertical projection profile. By producing the ground truth image which countered to be a typical image and computing the correlation between its histogram and the histogram of a dewarped image, one can measurement how correlated two images are or how the dewarped document image is near to the ground truth document image. The nearer correlation coefficient to one, the more matching the document image is to the ground truth. The correlation between the histogram of a plane image and the ground-truth image is prospective to be near one.
presents the ground-truth image, the dewarped image by mercantile software and their identical vertical projection profiles

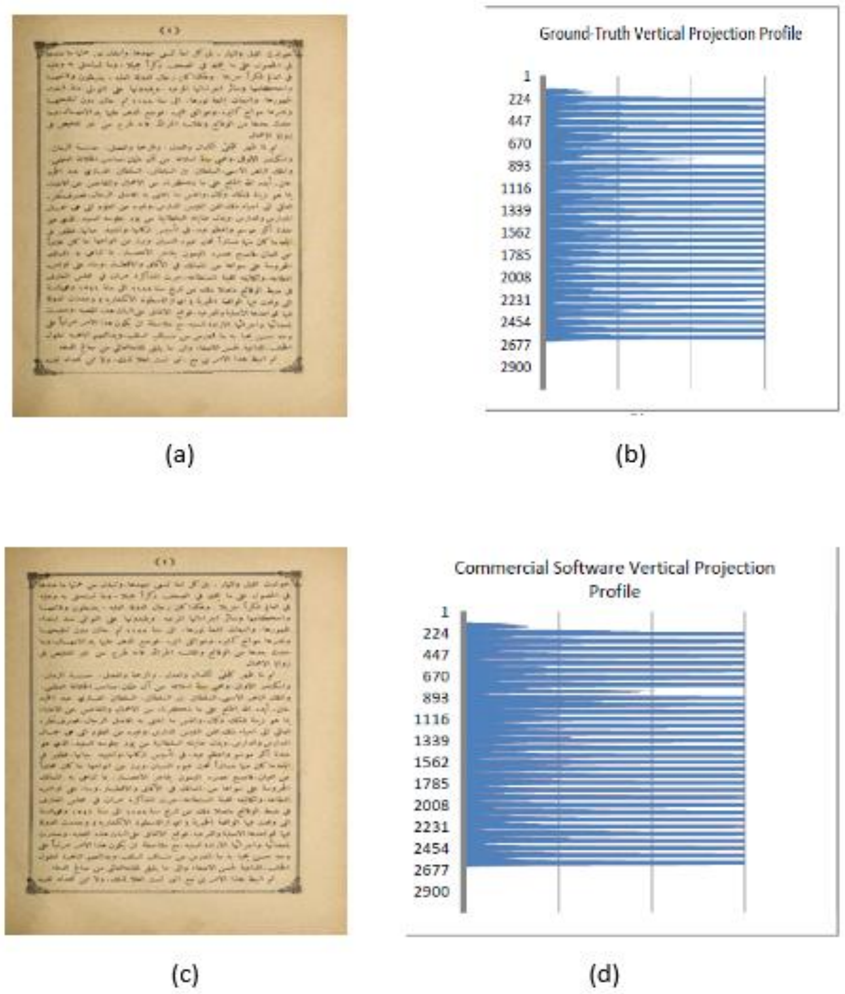


Figure 6-1:(a) Ground-truth image (b)vertical projection profile of ground-truth image (c) mercantile software dewarping method (d) vertical projection profile of mercantile software(Copyright: Qatar National Library)

6.3.3 Segmentation evaluation technique

Another technique for achievement evaluation is to contrast the segmentation results between the output of the dewarping method and the ground-truth image. The concept is that if the image is warped, areas are joined and the segmentation outcome is various from that of the ground-truth image. The more analogous the dewarped image is to ground truth, the more analogous are the segmentation consequence. In this technique, first the dewarped ground-truth image is produced; at the next stage on together dewarped ground-truth image and product image of a dewarping method, segmentation ground-trothing is completed and at the final stage the segmentation outcomes are evaluated.

The above method is illustrated in the following:

To produce some dewarped images with various procedures, additionally to dewarped image with Book-Restorer software, in the technique of producing ground-truth different intermediate image is also produced. This image is produced by a coarse grid (Figure 6.2a) rather of the fine grid which is applied for the final ideal ground-truth image (Figure6.2b).

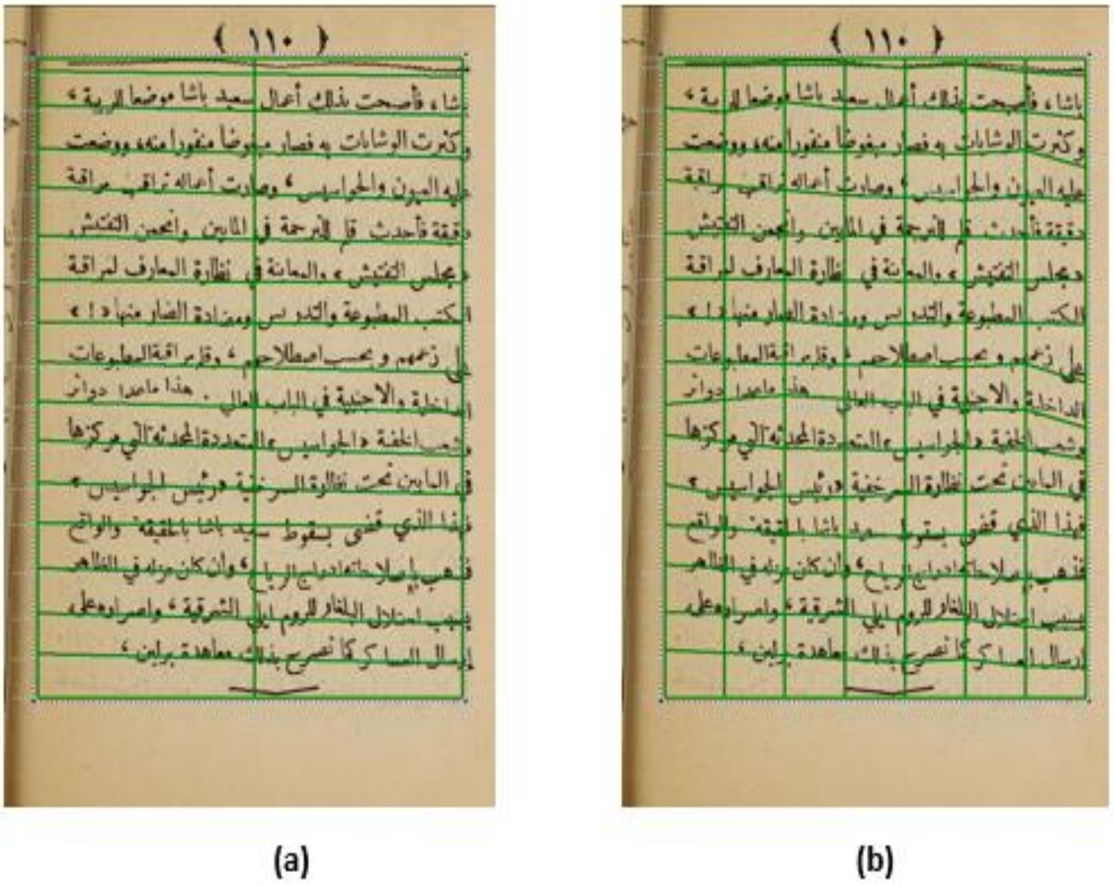
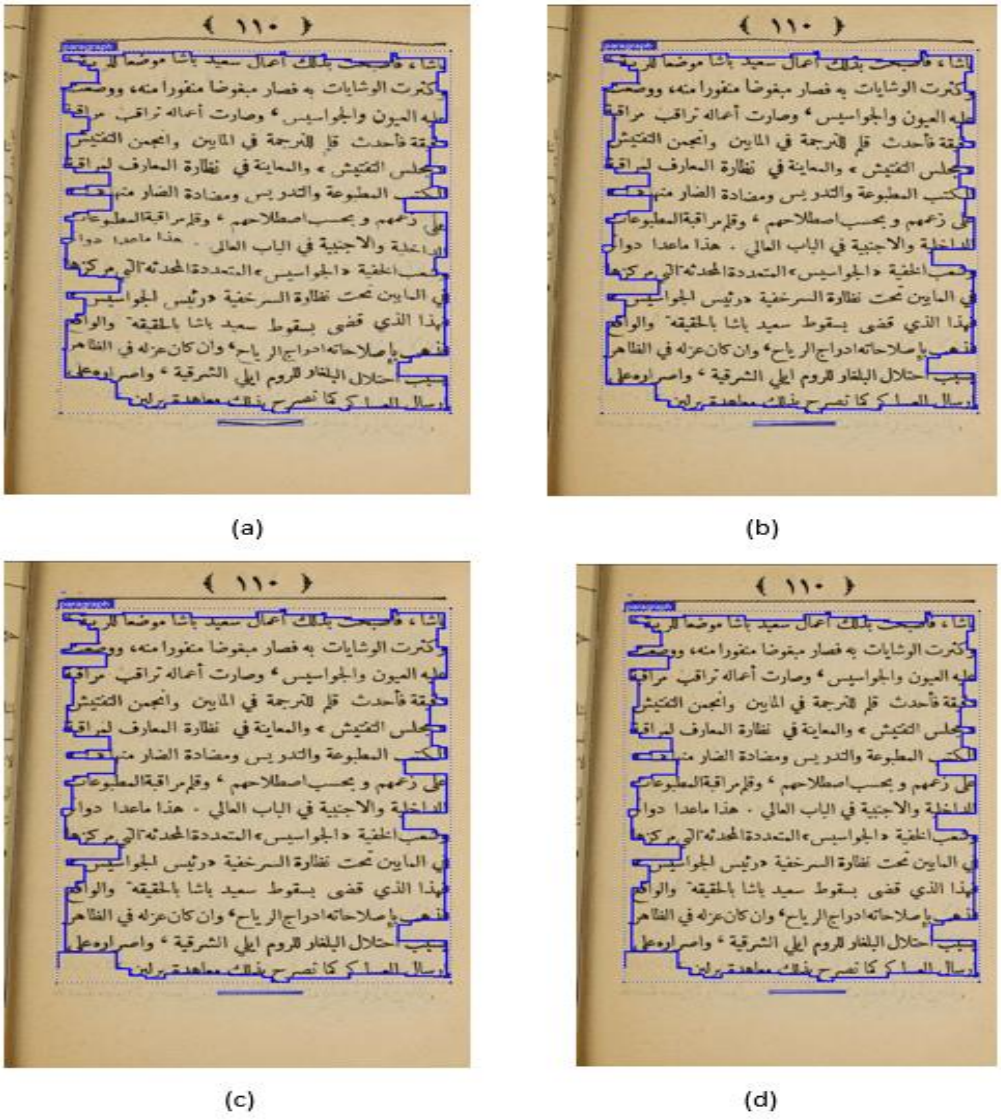


Figure 6-2:(a) coarse grid (b) fine grid (Copyright: Qatar National Library)

Heretofore, we have four document images; authentic image (Figure 6.3a), dewarped image by coarse grid (Figure 6.3b), Book-Restorer dewarped image (Figure 6.3c) and ground-truth image (Figure 6.3d). Another ground-truthing toolkit (Alethia) is applied to segment these four images. Line and region segmentation consequence on those four images, obtained from Alethia, are represented in Figure 6.3.


(a)

(b)

(c)

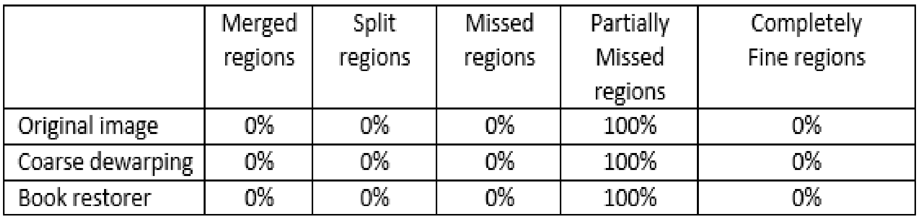
(d)

Figure 6-3 (a) authentic image (b) dewarped image by coarse dewarping (c) Book-Restorer results (d)dewarped image by ground truth dewarping

In the concluding stage in this experiment, a different tool such as Book-Restorer software in PRImA Research is applied   for evaluating the segmentation outcome. In this process, the dewarped image obtained from the dewarping ground truth tool is the standard and all the segmentation outcome from other images which are the authentic image, coarse grid, and Book-Restorer, will be contrasted with it.

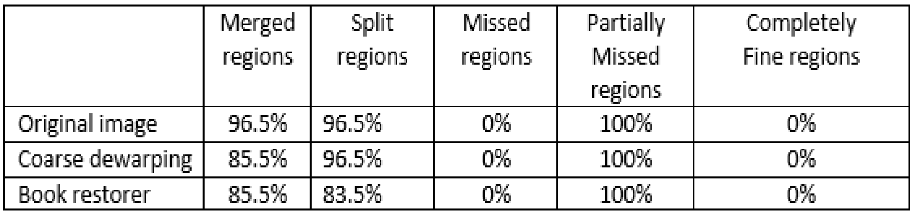
The segmentation evaluation toolkit examines the segmentation outcomes of every segmentation technique with the segmentation of a typical image and assesses the merged, split, missed, partially missed and perfectly fine regions and lines sequentially. Tables 6.1-6.3 presentation the segmentation evaluation outcomes from segmentation toolkit from the segmentation evaluation toolkit.

Table 6-1: Region segmentation evaluation outcomes



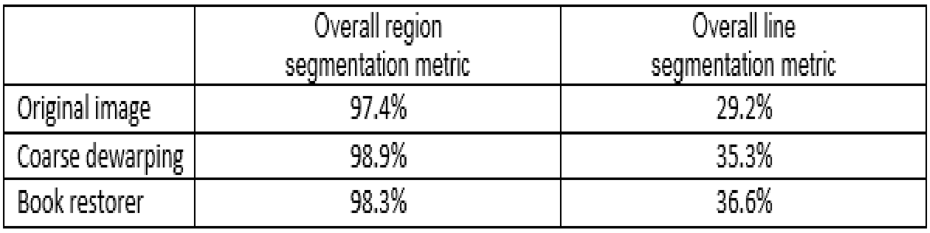
	Merged regions	Split regions	Missed regions	Partially Missed regions	Completely Fine regions
Original image	0%	0%	0%	100%	0%
Coarse dewarping	0%	0%	0%	100%	0%
Book restorer	0%	0%	0%	100%	0%

Table 6-2: Line segmentation evaluation outcomes



	Merged regions	Split regions	Missed regions	Partially Missed regions	Completely Fine regions
Original image	96.5%	96.5%	0%	100%	0%
Coarse dewarping	85.5%	96.5%	0%	100%	0%
Book restorer	85.5%	83.5%	0%	100%	0%

Table 6-3: Overall Line and Region segmentation evaluation outcomes



	Overall region segmentation metric	Overall line segmentation metric
Original image	97.4%	29.2%
Coarse dewarping	98.9%	35.3%
Book restorer	98.3%	36.6%

By considering at images, coarse dewarping appears flatter than the Book-Restorer outcomes; while, segmentation outcomes makes better on region evaluation but worse in line evaluation. Furthermore, though an authentic image is flatter than the Book-Restorer image, their overall region segmentation outcomes are the identical.

This process of evaluation also highly counts on segmentation outcomes; any mistake in segmentation technique or segmentation evaluation metrics lead to inaccurate outcomes in the dewarping evaluation.

In the following, different supervised evaluation methodology based on baselines positions in ground-truth and the resulting image is displayed.

6.3.4 Grid technique

Another supervised evaluation technique is to contrast the coordinate points of the effective productivity of a dewarping system with the ground truth one. To be qualified to contrast various methods generally and on the analogous coordinate positions, we have created a grid on the ideal ground truth image; the horizontal lines of the grid match to the text lines. The coordinate points of the ideal grid on the ideal flat image are registered. Following, the same grid is composed on the outcome of a dewarping technique. In the next stage, the user adapts the location of coordinate points until they find on the text line in the new image. Clearly the points which are previously on the baseline will maintain not changed. The variation between the coordinate points of the grid on the flat image and the modified grid on the product image of a dewarping method can be applied as a metrical for the evaluation. All the above procedures are performed by the ground-truth and described in section 5.6.3.

In brief, this technique involves four steps: 1- ground truthing, 2- producing a perfect grid on the ground truth image, 3- composing the perfect grid on the product image of a dewarped method 4- modifying the grid to the point that it matches to the text line on the result image.

For example, Figure 6.4a presents the ground-truth image and a typical grid located over it. Figure 6.4b displays the dewarped image by Book-Restorer software and the same grid composed on it. As it can be observed from Figure 6.4b, the grid does not correspond to the text lines and it should be modified to be altered on the text lines. Figure 6.4c shows modified grid on the dewarped image by Book-Restorer software. The variation between a typical defined grid and adjusted grid can be applied as a metrical for evaluation

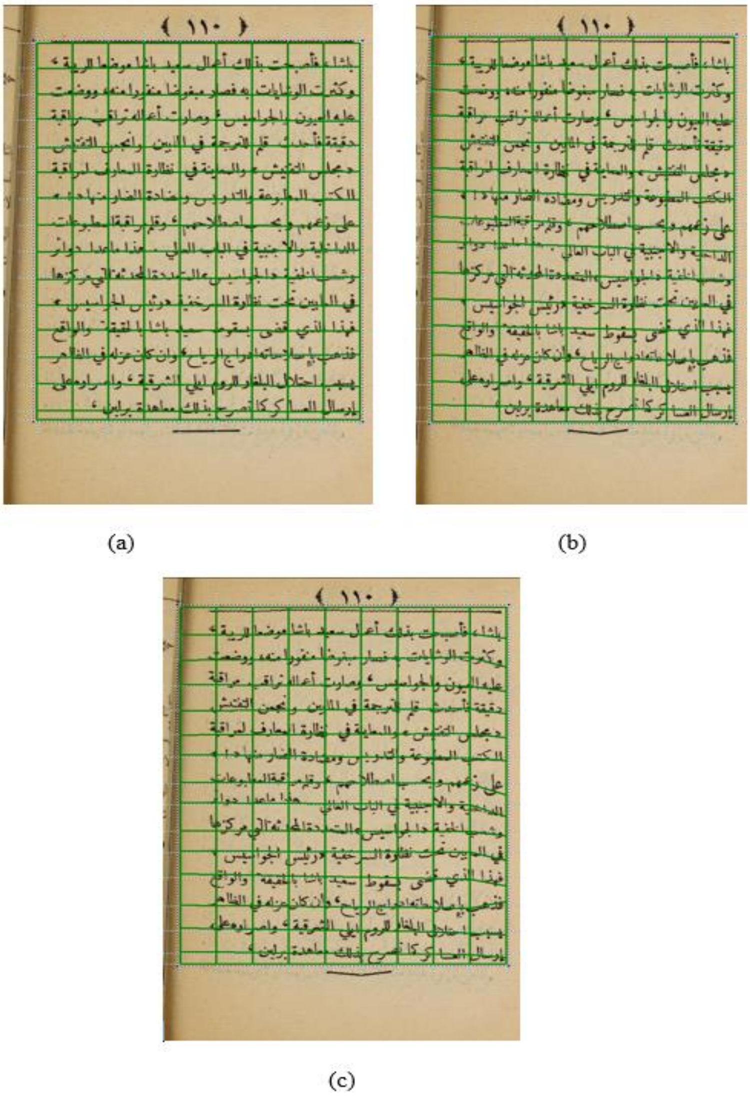


Figure 6-4:(a) typical grid on the ground truth dewarped image (b) typical grid on the dewarped image by Book Restorer (c) modified grid on the dewarped image by Book Restorer

The evaluation methodology applied in this process is based on supervised evaluation with (manually produced) ground-truth data. After the baselines are marked manually on both the authentic warped document image and outcome document image. Then, according to equation (6.4), computed the average baseline straightness of the authentic and the enhanced image. The outcomes of two extra geometric correction techniques are compared one is a state-of-the-art page-curl correction technique produced for IMPACT by NCSR (Stamatopoulos et al., 2011b) and the leading commercial product Book Restorer(Restorer, 2018).

Consequently, to estimate dewarping by measure the "straightness" of baselines, the average proportion of the sum of sub-cross-region by quadrilateral-area in each baseline is measured. As presented in Fig. 6-5, the sub-cross-area indicates the region of the sub-region formed by the baseline and mean Y line.

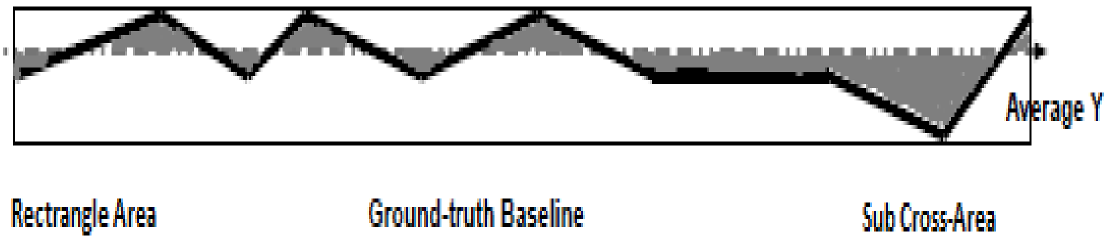


Figure 6-5: Rectangle region and sub-cross-region of a baseline

Commonly, for correctly straight text lines, the proportion of sub cross region to bordering on box area is lower (higher accuracy), whilst, for massively warped text lines, this proportion is supposed to be significantly greater (lower accuracy).

Consequently, the accuracy of handling arbitrary warping in a document image with N baselines can be solved by the following equation (6.4):

$$\text{Accuracy} = 1 - \frac{\sum_1^N \frac{\sum_1^M sub_{ij}}{\text{Rec}_j}}{N} \qquad (6.4)$$

Where:

sub_{ij} : Area of one sub cross-area in a marked baseline.

Rec_j : Area of bordering on the box of a marked baseline.

N : The numeral of baselines marked in a manuscript image.

M : The numeral of sub-cross ranges in one marked baseline.

6.3.5 Evaluation technique based on manually identified of points on text lines

Stamatopoulos et al. (2009) proposed the evaluation methodology, at the first step, user have to identify several points on restricted number of text lines of the authentic warped image. The points must be chosen in the midst of the main frame of the words and also on the long text lines which represent the image. Then at the next step, by applying the SIFT transform (Lowe, 2004) the manually identified points of the authentic warped image are correspond with the analogous points of the dewarped image. The integration of every cubic polynomial curve is used as the reference of the performance of the dewarping method. The utility of this technique is that it does not count on OCR for evaluation and it could be quicker than techniques which fully rely on ground-truth.

However, this method has some restrictions. Choosing some representative lines on the page is substantially relies on the user and it could be perfectly subjective. Besides in arbitrary warping every line has its own deformation and choosing a restricted number of lines (i.e. 2, 5) in the page with about 30 arbitrary text lines is not an inclusive evaluation. Additionally, if there is distorts short lines it cannot be estimated by this evaluation methodology because the short lines are ignored in this technique.

6.4 Unsupervised evaluation

Gaining accurate ground truth requires intensive work, is time exhaustion, and there is no warranty that the user can always determine the baseline accurately and consistently in all images.

Challenges in getting the ground truth in supervised techniques, such as an imprecise description of the baseline, and time-consuming make unsupervised techniques more appealing. In an unsupervised system, there is no request for ground-truth and each image can be estimated on its own. Here the concept is to have horizontal straight lines in the final dewarped image; consequently, any variation from the straight line is considered an error and reflects the performance of the dewarping technique.

To determine a metric for the unsupervised evaluation and to assess the deviation of each line from the straight line, the following procedure is performed. In the first step, the accurate baseline is detected. Here we have applied the suggested baseline detection technique illustrated in chapter 5. The definitive parametric equation of each line$f(t)$, after changing new coordinate points. The integration of the parametric equation of each line when t ranges between 0 and 1 (equation 6.4) references the area under the curve.

$$I = \int_0^1 |f(t)| \qquad (6.4)$$

The region under the curve which is a possible metric for evaluating a dewarping method references the perversion of each line from the horizontal straight line. The nearer the line to the horizontal straight line, the least is the rate of the integral. In this process, the integration or the area under the horizontal straight line is 0.The integration of a curve calculates the region between the curve and horizontal axis (x-axis). The origin of all coordinate points are calculated from the top-left corner of the in corner of the image; distinctly the region between each text line and the upper border of the image is not a perfect metric for dewarping evaluation because in this process even the integration of straight horizontal line is not zero but it is the area surrounded by the line and the upper boundary. Accordingly, for each text line a various coordinate system should be determined. The y axis of coordinate system continues the same for all text lines which is the left border of the image. However the origin and x-axis should be change for each text line.

The objective is to determine a level where the area under the curve which indicates the perversion of the curve from the straight line, becomes minimum.
The region under the curve with respect to the minimum rate of the curve, minimum level, can reference the perversion of the line from the straight line when the majority of points are existing on the straight line and there is an upward curl at either end of the curve (Figure 6.6). The same debate holds for the greatest point of the curve when there is a descending curl

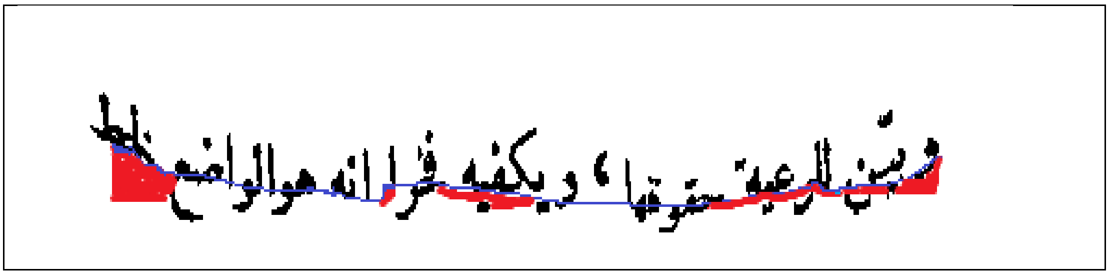


Figure 6-6 Integration in respect to the lowest level

However, in some instances, the region between the curve and minimum level can create an extra shift and therefore extra mistake. For instance, in Figure 6.7a, there is a minimum point on the curve which is the ravine of the curve; therefore, the integral of the curve in consideration to the minimum level provides error at both sides of the ravine. The average altitude of the curve can also produce extra shift and extra an error with the same cause.

The level which is subsequently chosen as the x-axis for calculating the integration in regard to that level, is the level where most points of the curve are existing at that level; the reason behind it is that in this case the integral of the most points of the curve would be zero and the mistake would be minimum. This level is the most horizontal portion of the curve. We call this level the modal level.

Figure 6.7b presents the state where the integration of the curve is computed in respect to the modal level. It can be observed from the Figure and as certained by numerical values that in this instance the integration of the curve will decline, and it does not contain any constant shift in the integration.

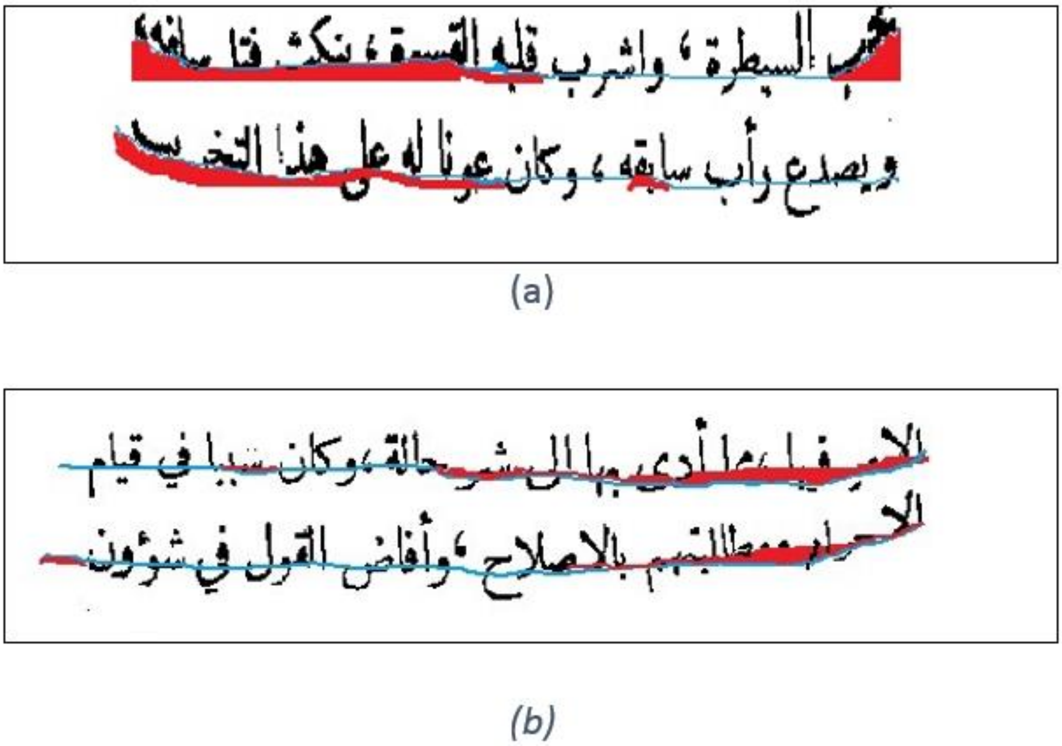


Figure 6-7 Integration of the baseline curve in respect to (a) minimum level and (b) modal level

Although, for integration calculation, the modal level is the better one which produces the minimum mistake, the relationship between maximum level, minimum level and modal level can be used for explaining the form of a curve. For example, when the integration in respect to minimum level is equivalent to that of the modal level, the curve has a concave form; furthermore, when the integral in regard to maximum level is equivalent to that of the modal level, the curve has a cambered form. The baseline has a wavy style, when the modal level integration is among minimum level and maximum level integration. The same analysis is appropriate to the integration value of the function and its ultimate value. These procedures can assist in defining the kind of deformation which occurs in each line and whether or not a dewarping method has extracted the whole distortion, part of it or has produced a new type of deformation.

Having determined the suitable level for computing the integration of each line, now a convenient metric for dewarping evaluation can be specified.

Two metrics can be acquainted for this objective; the first metrics references the perversion of each line from the horizontal straight line. Consequently, each line can have a metric referencing its straightness. Metric E_j reference the perversion of each line from the accurate line and is determined separately for each line in the image as next:

$$E_j = \frac{\int_0^1 |f_j^{Mod}(t)|}{L_j} \qquad (6.5)$$

Where $\left|f_j^{mode}(t)\right|$ the unlimited value of the j^{th} line function is computed in respect to the modal level and L_j is the length of the j^{th} line.Separator the integration by its length normalised the metric. This is essentially important for the images which do not have text lines with the same tallness. The less is the perversion of the line from the straight line, the reduce the value of the E_j.

The second metric is assigned awarding to the next equation; it shows the precision of each line in a dewarped image.

$$A_j = 100(1 - \frac{E_j}{E_J^0}) \qquad (6.6)$$

Where E_J^0 , is the identical metric in the authentic image and E_j ,is the perversion of line j from the upright line. The better is the precision of the line with a dewarping technique, the higher the value of the A_j. A perfect horizontal upright line has the exactness of 100% while a line which has the same deformation as the authentic image has the precision of 0%

The overall precision for the image is obtained by the averaging of all A_j measures in a leaf:

$$A = \frac{\sum_{j=1}^{N} A_j}{N} \tag{6.7}$$

Where N is the quantity of lines in the image.

Figures 6.8 a and b present the authentic image and dewarped image with Book-Restorer software and Figures c and d demonstrate the integration of each line in respect to the modal level.

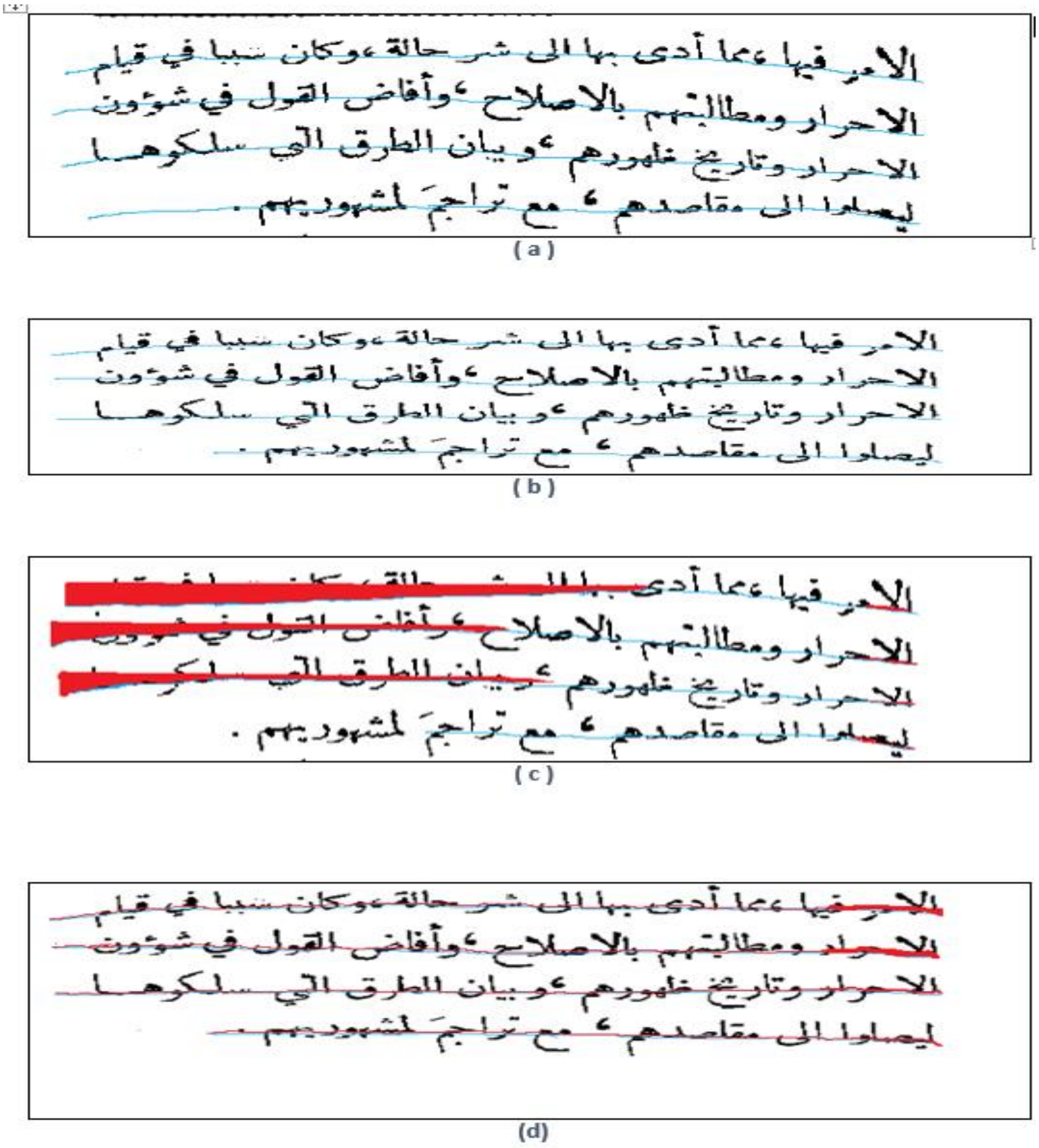
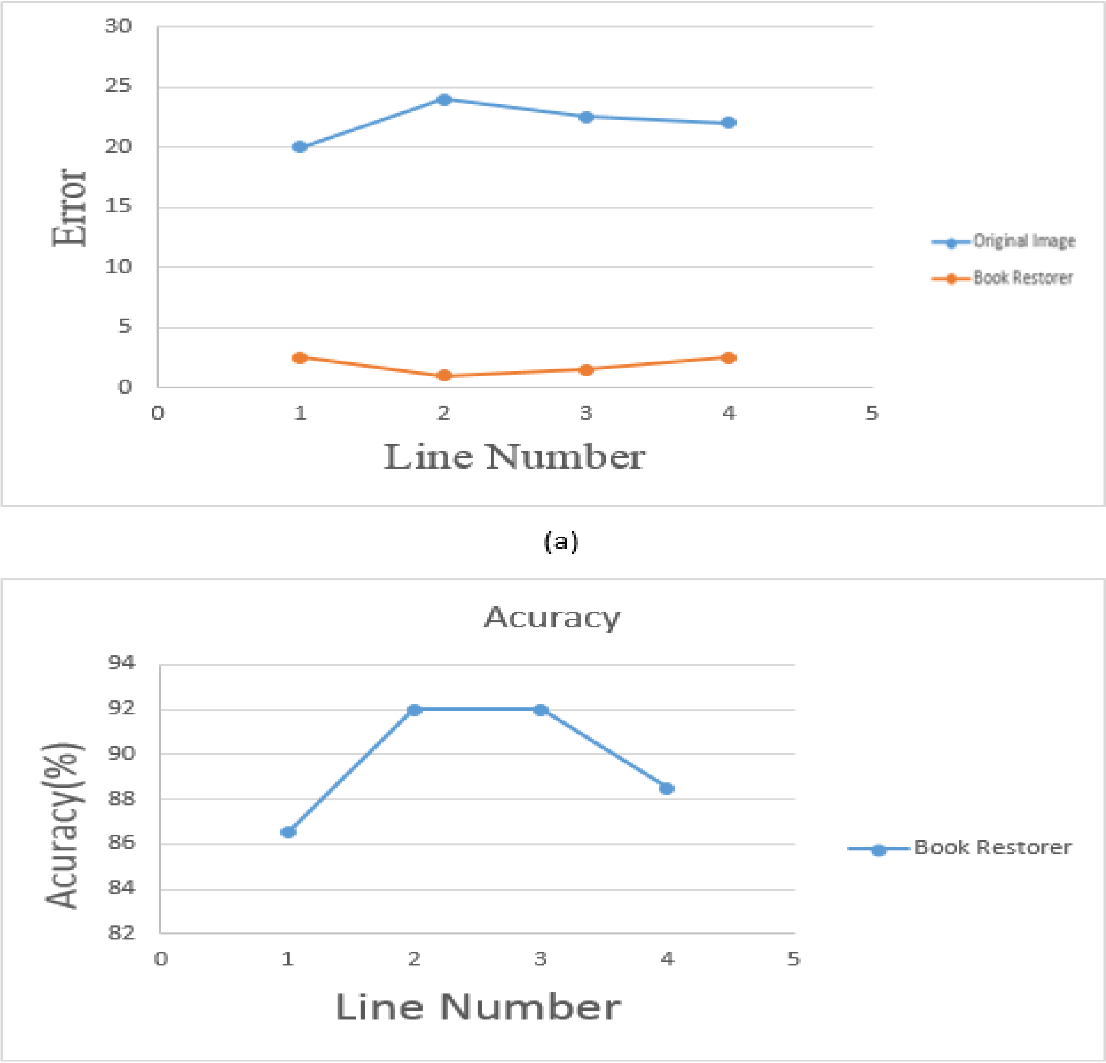


Figure 6-8 (a) Authentic image (b) Book-Restorer result; (c), (d): the integral in respect to the modal level

The results of computing the first and second metric which measurements the perversion of each line from the straight line for figure 6.8 is depicted in figure 6.9. It can be observed from Figure 6.8a that each line in the dewarped image has less perversion from the upright line than the identical line in the authentic image. Conforming to Figure 5.9a, the second line in the authentic image has the greatest mistake whilst the first and final line in the dewarped image has the highest error in respect to other lines. Figure 6.9b demonstrations the accuracy of every line represented by equation 6.6.



(a)

(b)

Figure 6-9: (a) perversion of every line from the straight line (b) accuracy

6.5 Benchmarks

One system of evaluating the execution of the suggested dewarping system is to assess it with the state-of-the-art techniques. The majority of suitable techniques in the literature are prepared for page curl for modernistic documents. Between them we obtained the following systems more general which are more suitable for historical scripts as well.

6.5.1 State-of-the art dewarping technique

The majority of the techniques in the literature are performed for new documents with page curl as the only deformation. Consequently, they are not suitable for more challenging statuses, for instance, arbitrary warping in historical manuscripts. To make a fair compare, only techniques which are prepared for such an objective are chosen.

It is debated in chapter 3 that 3D dewarping techniques which apply specific setups, for example structured light or 3D laser scanner, appear to be able to address arbitrary image deformation inclusive arbitrary warping and fold; although, many images are previously collected and are preserved in various libraries around the world and it is definitely very expensive and even sometimes difficult to reproduce images utilising any given specific scanner or camera. In our instance, images are previously taken with several copiers and scanners and there is no additional 3D information given by libraries. There is not any information about the copier or scanner only knowledge is one single image. Since 3D dewarping methods need the extra 3D model in addition to 2D image, or some additional knowledge about the camera's structure or condition, they are not appropriate in our situation. The only technique in the literature which applies only the information in a single document image and is prepared for random warping in historical manuscripts is the technique presented by Stamatopoulos et al. (2011a) which is the further extension of the method by Gatos et al. (2007). The specifics about this method can be found in chapter 3.

The state-of the art dewarping technique introduced in (Stamatopoulos et al., 2011a) is indicated to the NCSR technique in the following chapter. NCSR stands for "National Center for Scientific Research (Demokritos)", where the technique is established.

6.5.2 Commercial software

To obtain an equitable comparison with commercial software once more, products which are created for random warping and historical manuscripts must be used. The software that is applied repeatedly by libraries for geometric rectification of historical manuscripts is Book Restorer (Restorer, 2018). It is listed in the software documentation that, "to obtain an optimal outcome whatsoever the situation of the document is, i2S developed modern geometrical correction task obtainable in Book Restorer software. This superior mission was registered at the international level." Geometrical rectification task in Book Restorer "rectifies distortions caused by scanning on a vertical scanner (book bend) or the existence of crumpled leaves". The software has various selections for geometric rectification. The user can choose one or extra of these choices based on how strong is the geometric deformation.

6.6 Summary

In this chapter several supervised and unsupervised methods for the evaluation of dewarping performance are described. Unsupervised techniques can assess the performance dependent of the final text lines must be straight. Supervised evaluation techniques require a typical image to assess the product image based on the ground-truth one. All the supervised evaluation techniques inclusive projection profile, segmentation and grid techniques require a ground truth image. Since the process of ground-truth is an effective and less time-consuming procedure of producing a typical image because it is manually created by Aletheia. However, in the projection profile and segmentation methods, the exact straightness of baselines cannot be evaluated but rather, their convergent positions are assessed; any fault in segmentation evaluation procedures also impacts the dewarping evaluation technique. The grid technique was used in this research because it is able to evaluate the correct straightness of baselines with complex layouts.

Chapter 7 - Experimental Results and Discussion

7.1 Overview

In this chapter, the outcome of the dewarping system proposed in chapter 5 are offered. The evaluation of the system and the compare between the suggested technique, the master state-of-the art technique and the advance commercial software system is based on the supervised evaluation approach introduced in chapter 6. This chapter starts with the description of the dataset utilised. The sample of the images used in the dataset, investigate the text line segmentation techniques results, test the developed baseline detection method results and Implementing of selected procedure for dewarping the performance of the suggested dewarping system on each example. Lastly, the chapter finishes with the analytical results for the whole dataset.

7.2 Generating the dataset

In this research by Dulla (2018) , a dataset of warped historical Arabic documents is created to incorporate undesirable mutilation that might be available in the last pictures, for example, curl and arbitrary warping. Dataset constructing is based on two core actions were chosen and delivery of images and metadata (individually by each library partner), supported by ground truth creation for distinct subsets.

7.2.1 Data Gathering

The image Gathering process for producing a dataset of such size had to be very directly and strictly defined and followed. To make sure that that the dataset comprises a sufficient number of documents with both simple and complex layouts in each of the content categories, the first step of document choice is an off-line expert-driven activity. Once the documents are selected, digitisation is started. This step has been completed by following strict conventions. First, the collected documents which had been scanned by libraries at 350 dpi and in 24-bit colour have been processed later. The second step image select was accompanied by metadata collection (see below). Owing to privacy issues, it is not easy to gain access to a large number of printed Arabic historical documents. Also, this is due to the lack of many kinds of distortion of historical Arabic documents. Subsequently, reasonable care is taken to create each document synthetic images. They are generated from a few base images by warping text lines close to the ones above, and text lines as well as overlap area are finely ground truths. In this respect 62 images in total were designated where 57.69% have arbitrary warping, 23.08% page curl, 9.6% fold. To measure if the stages of proposed system have any harmful effect on the images with straight lines were selected. Figure 7.1 demonstrations the distribution of the images in the dataset. It should be observed that most images have a combination of two or more kinds of distortions, i.e. page curl and arbitrary warping.

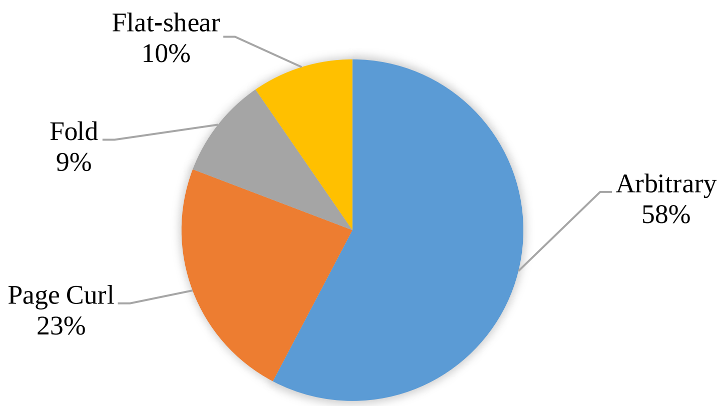


Figure 7-1 Distribution of the images in the dataset

.

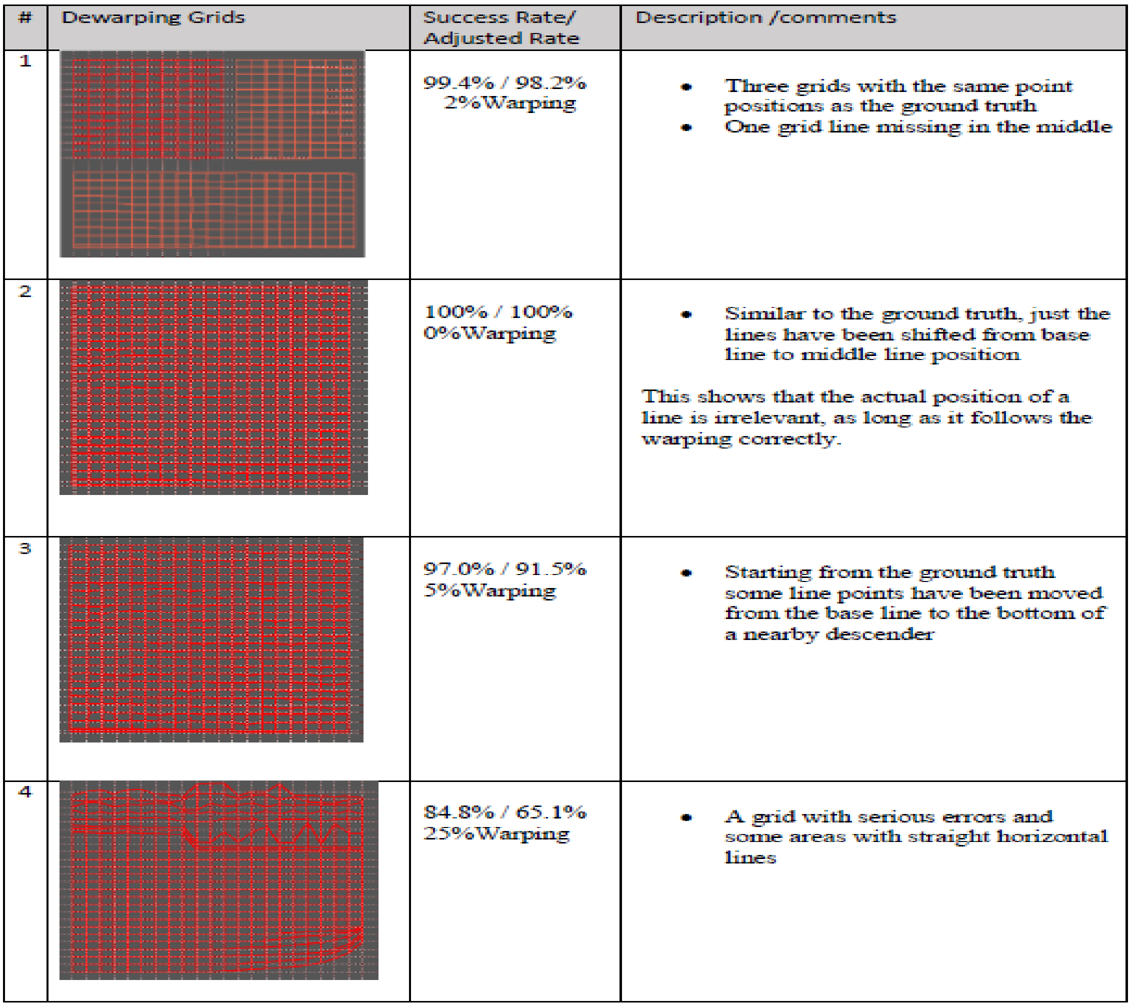
Once every accessible data was of satisfying quality, the pre-processing of the images could be carried out in order to be ingested into the dataset. The source images were presented in differing formats, such as uncompressed TIFF, JPEG, It was, therefore, important to standardise to an open and easy-to-use format so that images in the dataset could be easily used by a wide range of several tools.

7.2.2 Ground Truth Creation

One of the essential features of this dataset is the vital amount of definite and high quality ground truth accessible and the extension for its use. Ground truth, in this context, is a particular and formalized replica of what is really present on the physical page or, to place it in different words, what the right analysis/recognition approach is expected to come as result. Accordingly, it has to be designed (or at least tested) by a human. Aletheia is an advanced system for accurate and yet cost-effective ground truthing of large amounts of documents It supports top-down and bottom-up ground truthing. The format of the ground truth changes depending on the responsibility it is connected to (such as region outlines for segmentation or Unicode text for OCR). The ground truth is stored in the XML format which is part of the PAGE (Page Analysis and Ground truth Elements) representation framework. Ground truth is a vital advantage not only for improving and training new methods (such as adaptive text line segmentation or lexicon-based post correction) but also for performance evaluation as well as modification of end-to-end digitization workflows. The researcher tested the evaluation method on different manually created grids for a warped historical document to formulate and prepare for the dissimilar percentage (%) of warping on the image which are; 0% warping, 5% warping, 25% warping, 50% warping and 75% warping The ground truth grids and automatically generated dewarping grids are matched line by line where, each horizontal line of one set of grids is consequently matched against a nearby line of the other set of grids as shown in Table7-1.

Applying the ground truth baseline as a reference, two error values have been calculated that based on average distance to result baseline and average slope (angle) difference. The two values are calculated for each horizontal pixel position of a text line and combined using the arithmetic mean. The accomplishment rate for one single baseline is the harmonic mean of the non-linear distance success and the angle difference success.

Table 7-1: Generated dewarping grids are matched line by line



#	Dewarping Grids	Success Rate/ Adjusted Rate	Description /comments
1		99.4% / 98.2% 2%Warping	• Three grids with the same point positions as the ground truth • One grid line missing in the middle
2		100% / 100% 0%Warping	• Similar to the ground truth, just the lines have been shifted from base line to middle line position This shows that the actual position of a line is irrelevant, as long as it follows the warping correctly.
3		97.0% / 91.5% 5%Warping	• Starting from the ground truth some line points have been moved from the base line to the bottom of a nearby descender
4		84.8% / 65.1% 25%Warping	• A grid with serious errors and some areas with straight horizontal lines

7.3 Dataset description

The dataset contains a widespread assortment of documents reflecting both the holding of major Arabic and some European libraries. The dataset contains over 50 images originating from four different national and important libraries throughout the world such as Qatar Digital Library, the British library, the Library of Congress and Juma Al-Majid Centre for Culture and Heritage. The table 7-2 below presents some statistics about libraries books providers forward including numbers of contributed images.

Table 7-2: A list of including numbers of contributed images

Library	Country	Number
Qatar Digital Library	Qatar	50
British library	UK	10
Library of Congress	USA	10
Juma Al-Majid Center for Culture and Heritage	United Arab Emirates	10

7.3.1 Images

The presented database has covered a number of advantages over existing databases. It has variable page layouts and degradations that challenge text line segmentation and baseline detection techniques. Disregarding the page layout typically leads to an under-segmentation of text lines. The erraticism of page layouts in historical documents is much bigger than in the current book. In addition to some forthright features such as headings, subtitles, paragraph divisions, page numbers and page separators, there are some specific features such as ornamental letters, drop letters and catchwords. In terms of age, the majority of the manuscripts in the dataset – almost 90% – were produced in the 19th or early 20th century (particularly in the case of newspapers). Moreover, 23% were produced in the 17th or 18th century, with the remainder of the images ranging as far back as the 15th century. Although all probable steps were used to avoid some issues, in a dataset it is inevitable. There is a very small percentage of images where the publication year is not known, or was not made available to us (marked as "?" in the table below). A complete itemization of the age of the original documents in combination with the document types is presented in Table7-3.

Table 7-3: Document type and production century distribution

Century	Book Page	Newspaper page	Legal Document Page	Journal Page	Other Document Page	Unclassified Page	TOTAL
17	10	0	0	0	2	0	12
18	20	5	0	0	0	0	25
19	30	5	0	0	0	3	38
20	60	10	10	5	5	0	90
?	25	5	0	0	4	1	35

7.3.2 Metadata

According to Dulla (2018), metadata which is the XML-based representation that represents information at the document, page, and zone levels. It has many uses including merging/splitting functions and reading order.

In order to allow users to professionally search the repository, a huge set of metadata is preserved as part of the dataset. All accessible metadata is indexed and can be used as search parameters to access particular images or sets of images within the total dataset. Moreover, Dulla (2018) indicated that, the metadata has been designed under the following steps:

- Bibliographic knowledge — title, author, publicatiodate and location, document type, page number,
- Physical properties — language, script, typeface and number of columns.
- Copyright data — copyright holder, contactdetails, publishing permissions, content provider’s reference
- Managerial information original filename, access log,
- Digitisation data — resolution, bit depth, image dimensions, scanner used, file type, compression algorithm and quality, source of digitisation (paper,microfilm, etc.)

7.4 Text line segmentation technique results

In this research, various results are obtained when applying several techniques for text line segmentation, as mentioned in section 3.3 on the same samples with the various proportion of warping of historical Arabic documents. To determine the appropriate text line segmentation method, the results were compared under following success rate with the different ratio of warping.

A dataset of historical Arabic images are consisting of over 50 images with degraded document images by different percentage and various layouts.

Different results are obtained when applying the same technique on same samples with increasing percentage warp of historical Arabic documents as we have seen in figure 7-2. Figure 7-3 shows the results of the Hybrid Approach based on grouping of connected components, while figure 7-4 displays the results of Projection-based methods and Figure 7-5 illustrates the results of watershed transform on different warp samples for the same document image. Images (a), (d) are original images and images (b), (e) are after applying the proposed technique, watershed transform. Figure 7-6 presents the results of text line extraction method that uses seam carving procedure.

Based on our observations and as shown in Figure 7-3, Figure 7-4 and Figure 7-5, the Hybrid Approach is not flexible with diacritics and dots. Projection-based method works perfectly for printed text, where the text lines are comparatively and have appropriate space between lines and are straight. The watershed method is efficiently accurate of the situations, depending on CC analysis makes it highly unproductive in the state of warped or low-contrast manuscripts. However, the results of seam carving procedure have some advantages such as the method that can be used to various styles of writing whether it be printed or handwritten, and in diverse languages, which makes it a language-independent technique.

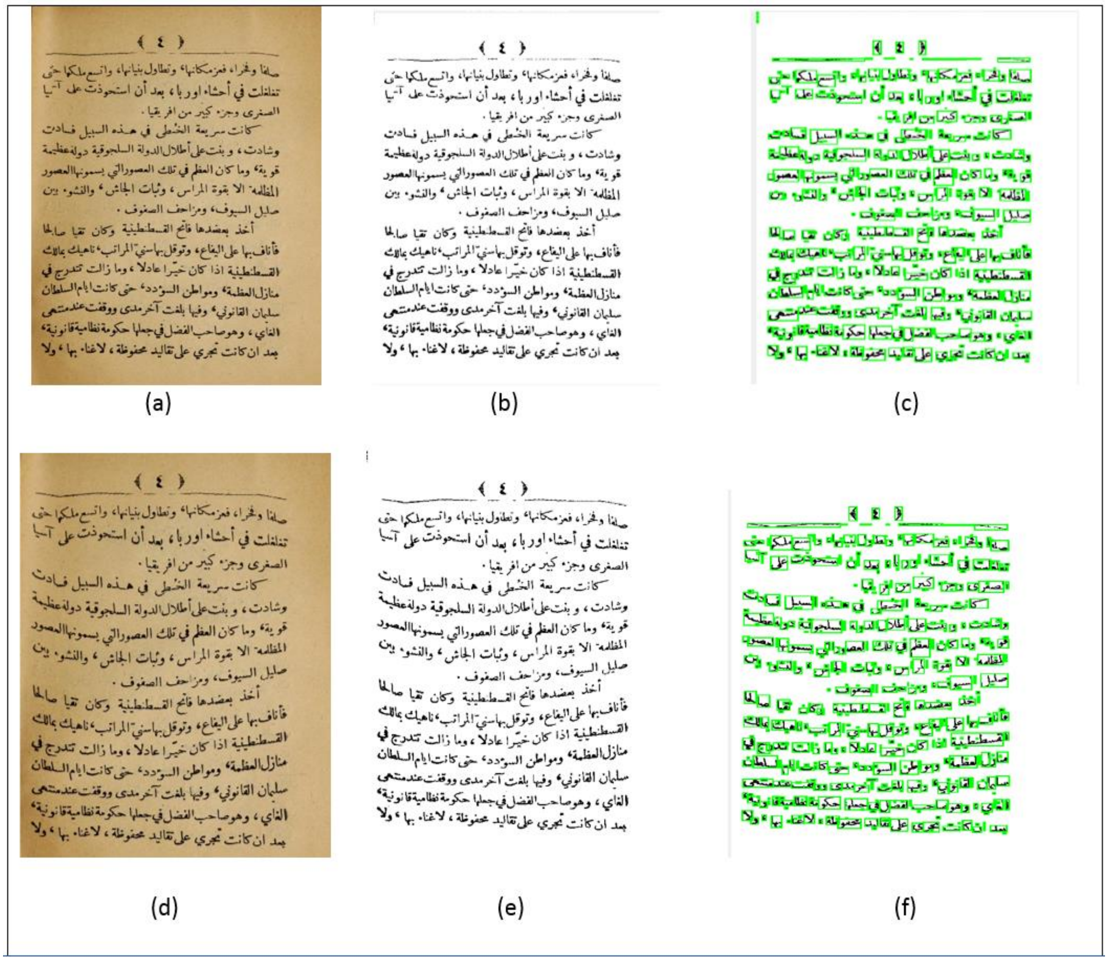
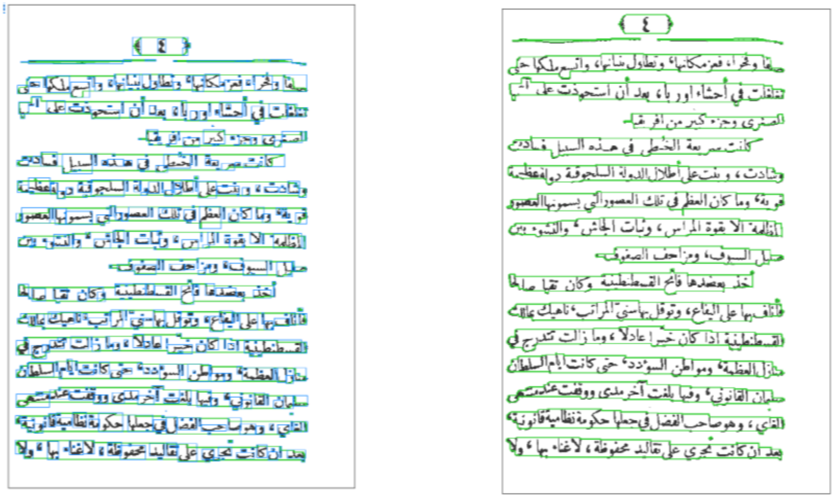


Figure 7-2:Different results are obtained images (a), (d) original image, images (b), (e) after applying pre-processing and after text line segmentation images (d), (f)



(a) Connect components (b) segmented text lines

Figure 7-3 Result of Hybrid Approach

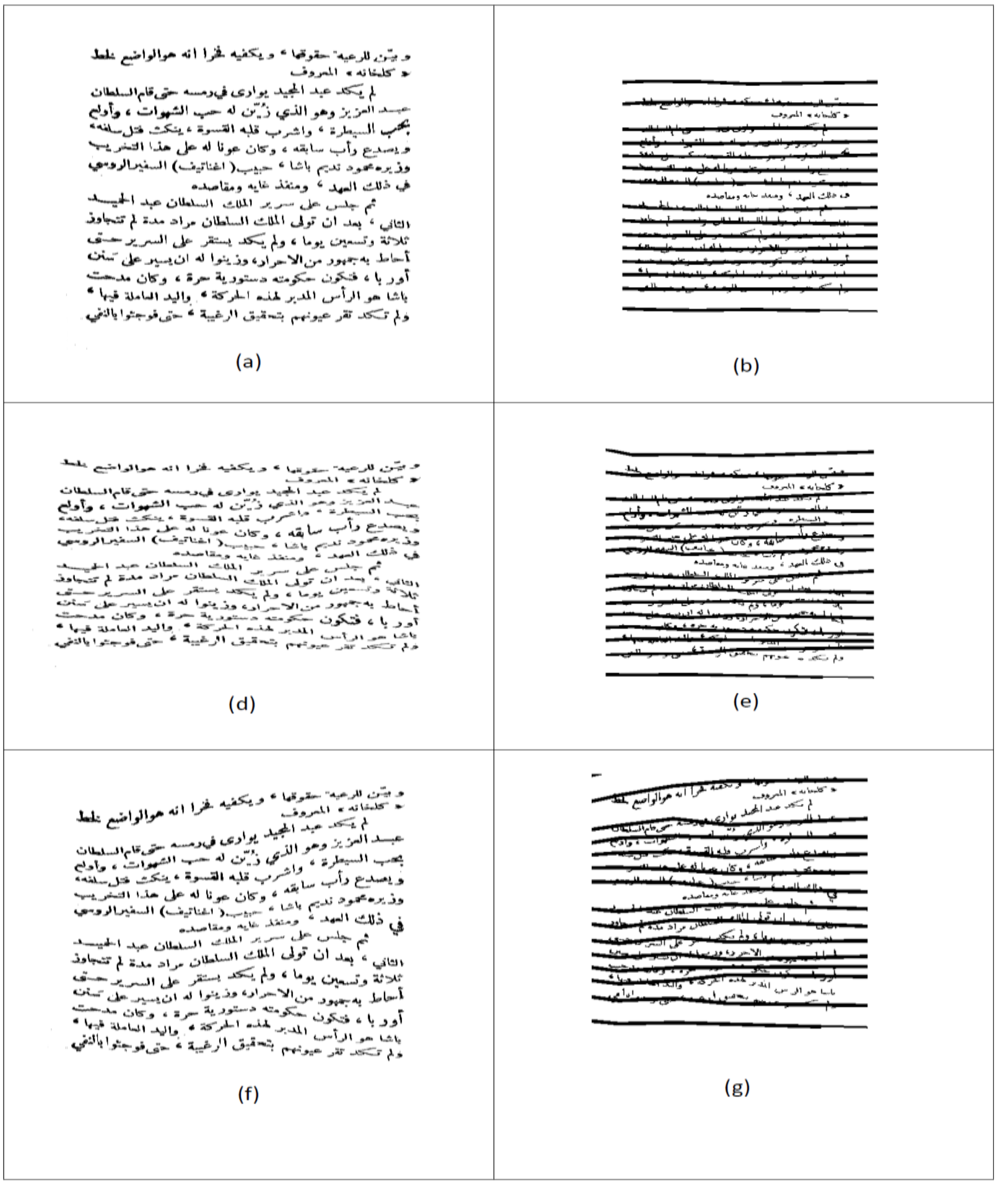


Figure 7-4 Result of Projection-based methods on different warped samples (a), (d) and (f) segmented text lines (b), (e) and (g)

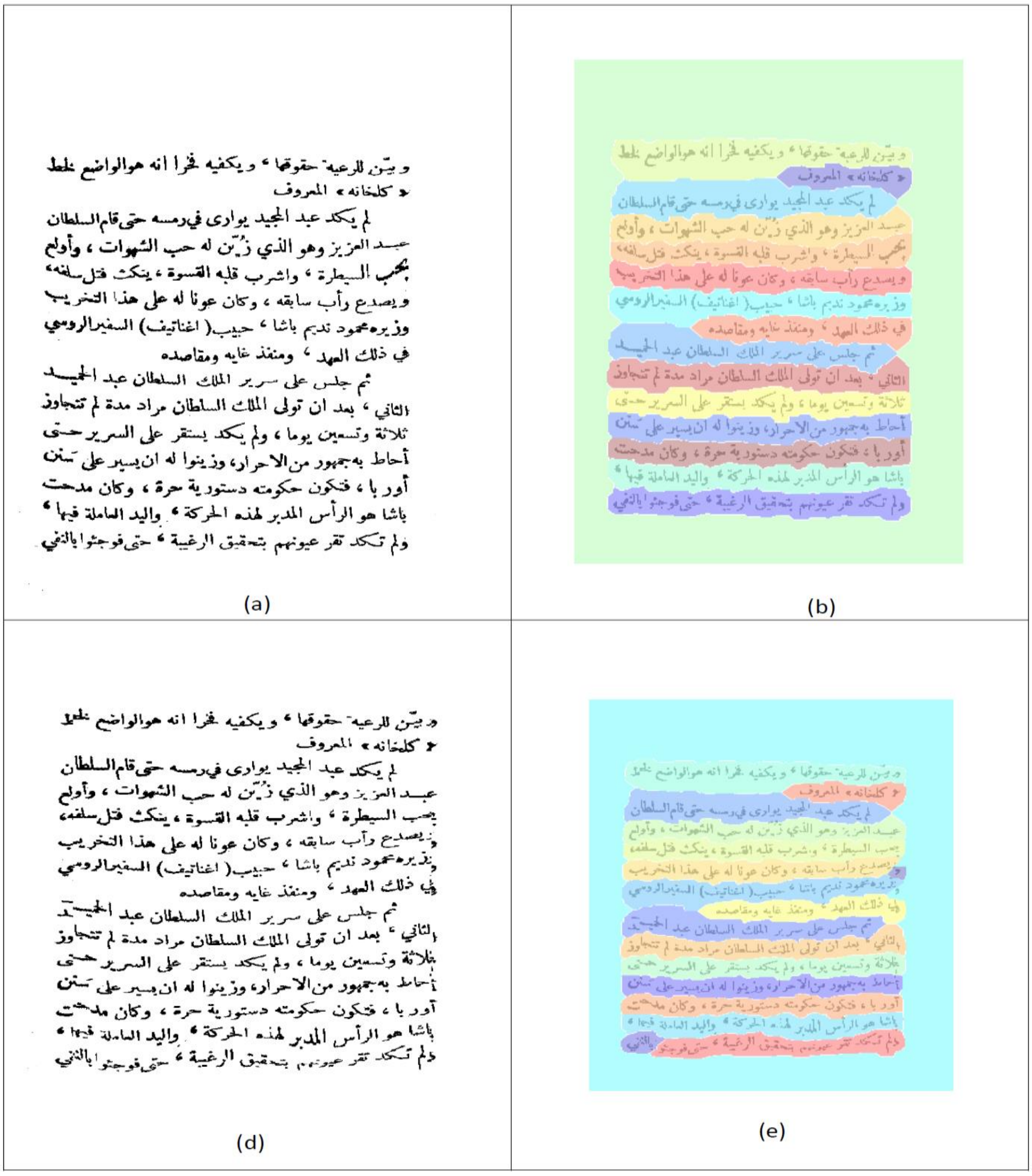


Figure 7-5 Result of watershed transform on different warped samples (a) and (d), (b) and (e) segmented text lines

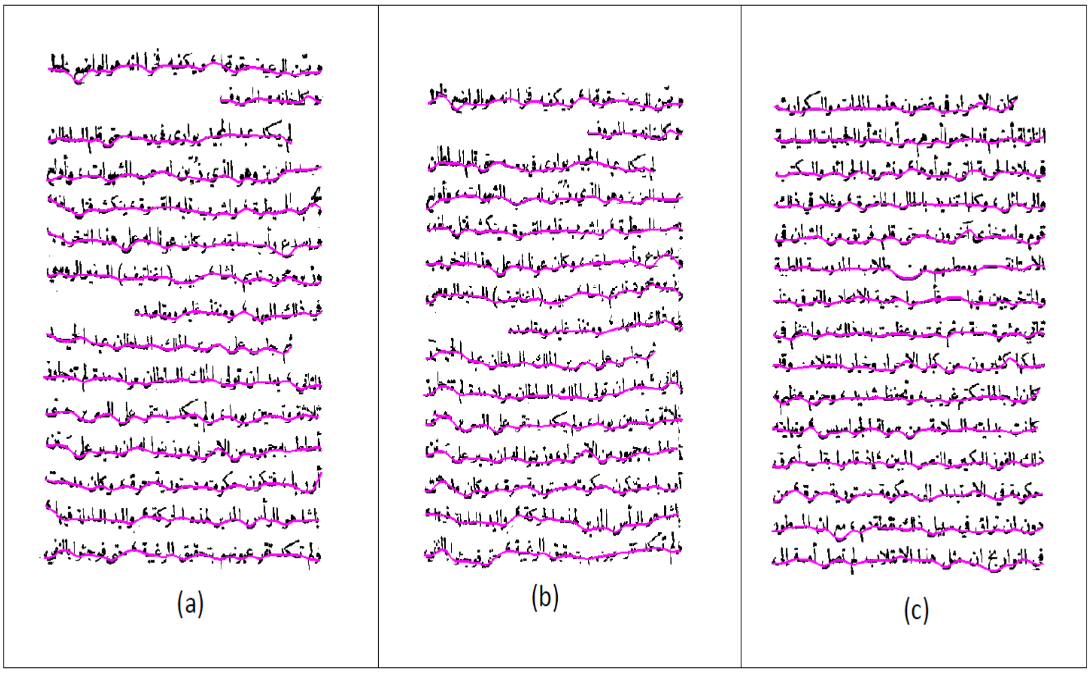


Figure 7-6: Result of smearing method on different warped samples (a), (b) and (c)

The tables and Figures illustrate the percentage of text line segmentation success rate on images that have been warped from different proportion. The percentage of warping is measured by Dewarping Evaluation tool. The data in table (7-4) and Figures 7:7, 7:8, 7:9, 7:10 and 7:11 show success rate on image with 0% warping, 5% warping, 25% warping, 50% warping and 75% warping, respectively.

The result (Table 7.4) shows that the seam carving method, clearly delivers better results than the other strategies on the literature review. It can be observed while the percentage of warping increased. At the time, there was a decrease in the performance of all segmentation techniques. It can be also noticed the majority of these methods fail when applied on a more difficult layout page.

Table 7-4: showing the success rate with different five proportions of: 0%, 5%, 25%, 50% and 75% warping

Method	The success rate with				
	0% Warping	5% Warping	25% Warping	50% Warping	75% Warping
Smearing method	98.5%	96.5%	90.5%	75.5%	60.5%
Watershed transform	94.9%	92.9%	86.9%	60.9%	50.9%
Hybrid approach	93.2%	90.2%	81.2%	51.2%	40.2%
Projection profile based	90.9%	88.9%	75.9%	45.9%	35.9%

According to (Table 7.4) the percentages of success rate in the four different methods ranged from 35.9% (with 75% warping by projection profile based) to 98.5% success rate at smearing method). It is clear to see that the highest percentages are recorded at the success rate with 0.0% warping in all methods that ranged between 90.9% (projection profile based) to 98.5% (smearing method).

Moreover, the percentages data showed that smearing method has the highest success rate percentages at all percentages of warping that ranged from 60.5% (75% warping) to 98.5% (0.0% warping). Furthermore, the projection profile based had the lowest success rate percentages at all rates that ranged from 35.9% (75% warping) to 90.9% (0.0% warping). The remaining methods; watershed transform and hybrid approach have recorded moderate percentages compared with the other methods (projection profile based and smearing method). Watershed transform method had slightly higher percentages than hybrid approach at all rate warping. The differences between the two methods ranged from 1.7% success rate (with 0.0% warping) to 10.7% success rate with

(75% warping). Also is it clear to see that, in general, the differences between all methods increasing with increasing rates of warping.

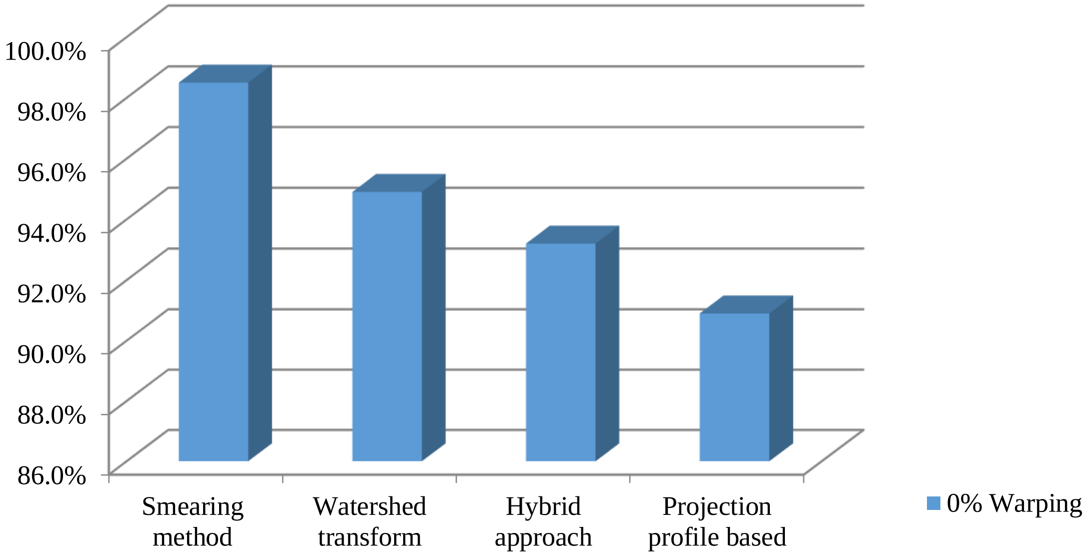


Figure 7-7: Graph showing the success rate with 0% warping

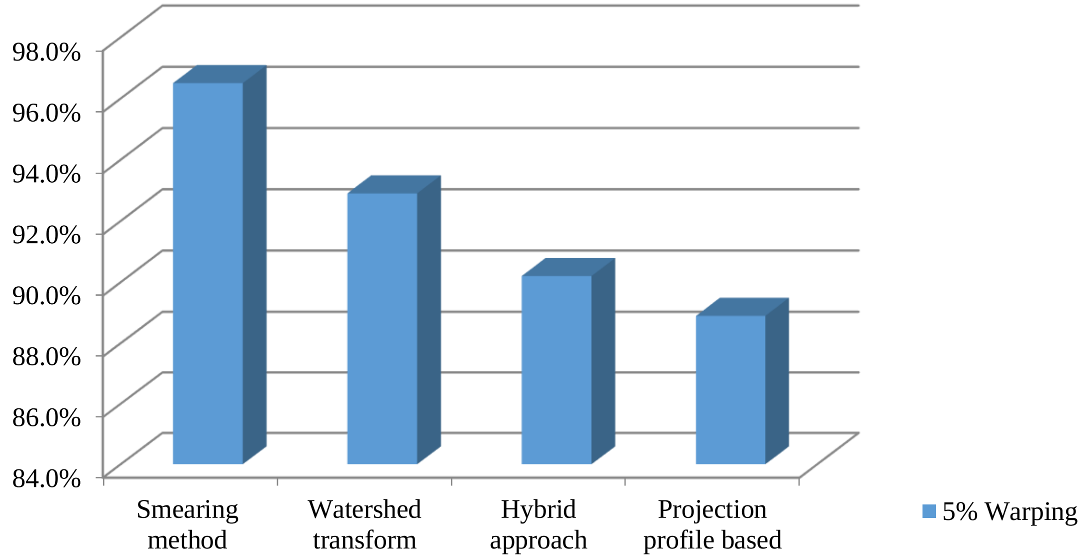


Figure 7-8 Graph showing the success rate with 5% warping

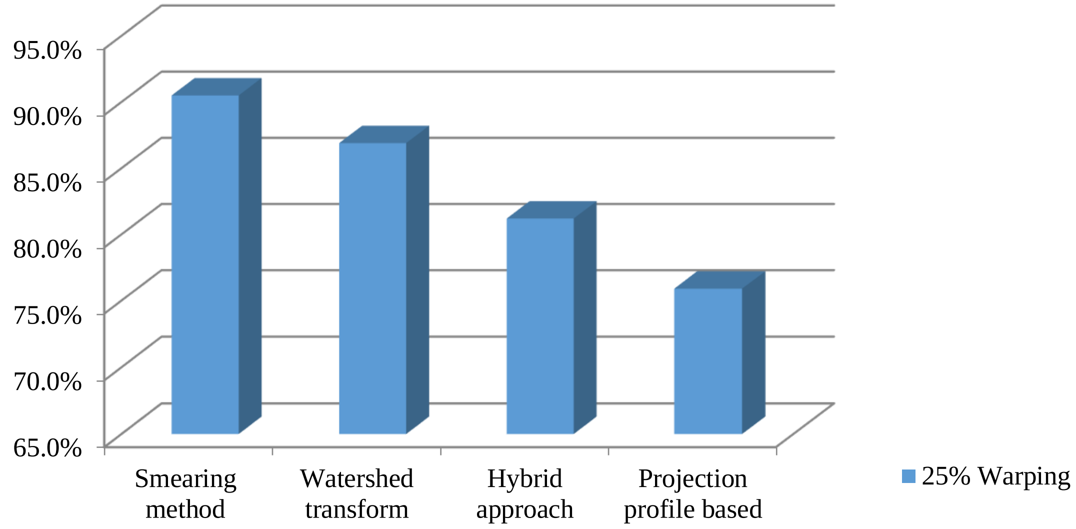


Figure 7-9 Graph showing the success rate with 25% warping

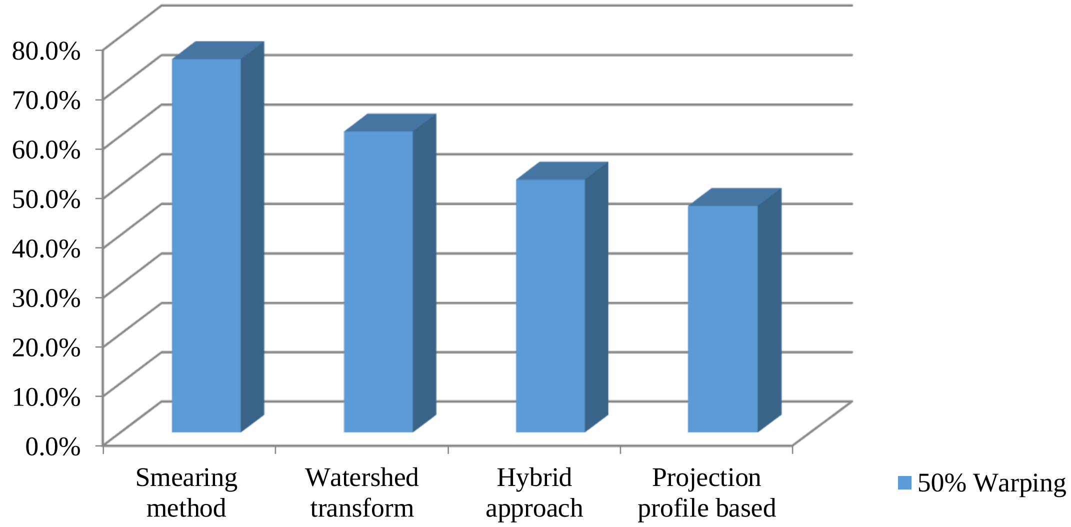


Figure 7-10 Graph showing the success rate with 50% warping

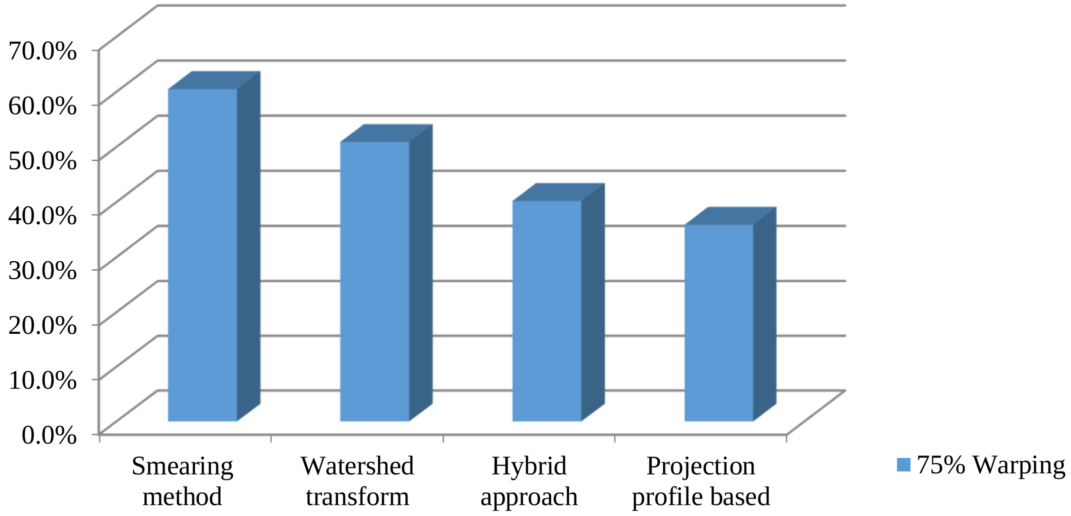


Figure 7-11: Graph showing the success rate with 75% warping

7.5 Test the developed baseline detection method results

With the above-mentioned method suggested in section 5.4.2, text lines are accurately segmented but the precise baseline is not detected and yet, one of the most important steps is baseline detection and straightness. Many experiments are conducted on the dataset created in section 7.2 to validate the abilities of the proposed technique on various warped documents images. The majority of the techniques failed to distinguish the right baseline when short characters and extensive diacritics exist see Table 4-3. In the following section, the reliability and flexibility of a new method for base line detection will be evaluated. The proposed base line detection method not only takes into consideration unforeseen geometric distortions, but also it distinguishes specific main components of the text line. Al-Shatnawi (2016) suggested baseline estimation technique based on the exploited components of Voronoi Diagrams (VD).This method based on potential points located in the same connected component and the angle between two adjacent edges the method works properly with the skewed and noisy images and with or without diacritics. However, it is not, able to estimate the baseline in document image with arbitrary warping.

The suggested efficient novel baseline estimation method in this research-based on the modified skeleton of VD. It is a multistage method with preliminary phases, VD constriction, and baseline estimation process phases as shown in Figure 7-9. The baseline ground truth is applied to evaluate our baseline detection technique. The basic ground truth image of the IFN / ENIT database and our dataset which created in section 7.2 are used to assess our baseline estimate. For the target of evaluation, we compare our technique based on two kinds of experiments visual and analytical experimentation.

Fig7-12. displays qualitative outcomes of the Pechwitz (Pechwitz and Margner, 2002), Boukerma (Boukerma and Farah, 2010) and suggested technique on images with diverse situation. Fig7-13 shows qualitative results to the same methods applying on our dataset created in section 7.2. The results attest to the accuracy of the suggested technique to obtain baseline in the state of text lines with various arbitrary warping, and in the case with of long words or short with different slant angles. However, Pechwitz method is poor in the case the case with of long words or short with various slant angles baseline relevant objects. Later on, Boukerma method solves Pechwitz difficulties but fail in the state of text lines with various arbitrary warping

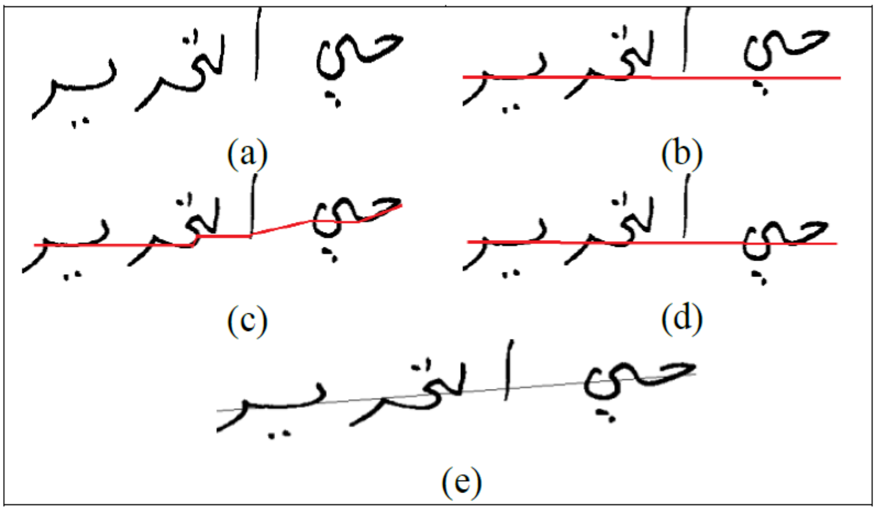


Figure 7-12:Arabic baseline evaluation techniques result applying: (a) the input image (b) Pechwitz, (c) Boukerma, and (d) the proposed method (e) ground truth images ae07_010.

:

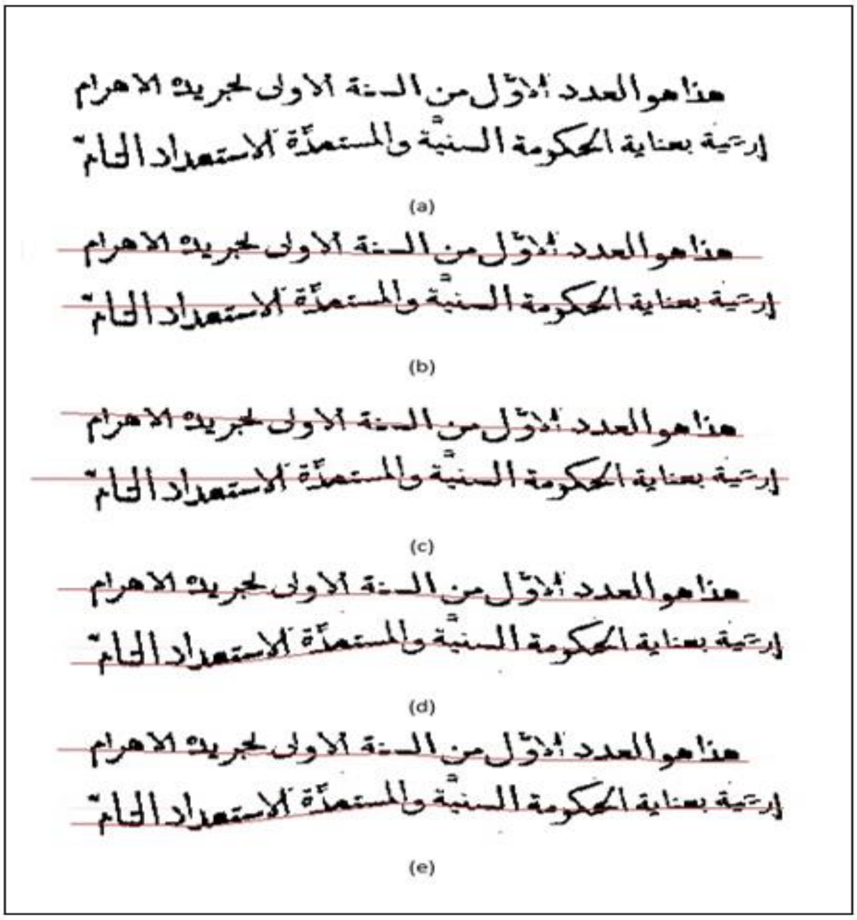


Figure 7-13:Arabic baseline evaluation techniques result applying on our dataset : (a) the input image (b) Pechwitz, (c) Boukerma, and (d) the proposed method (e) ground truth images

The analytical consequences of the suggested method were implemented to investigate the performance of our suggested technique. The distance among the points in evaluated baseline and the baseline ground truth image from IFN/ENIT and our dataset are computed by using the rightmost and the leftmost points for the baseline. For the quality measure, the error in baseline pixels are classified into six groups (0, 5, 10, 15, 20, 25). Table 7-5 explains the result of our suggested technique in comparison with the current works of Pechwitz, Boukerma techniques and ground truth image from IFN/ENIT.
Table 7-6 explains the result of our suggested technique in comparison with the present works of Pechwitz, Boukerma techniques and ground truth image from our dataset in section 7.2.

Table 7-5: Results of Pechwitz, Boukerma, ground truth IFN/ENIT and suggested baseline evaluation methods with several baseline error

Baseline error in pixel	Pechwitz method	Boukerma method	Proposed method
0	0.84	1.75	2.4
5	21.6	27.75	31.43
10	47.69	60.86	63.6
15	64.21	82.61	84.54
20	78.17	89.56	92.63

As shown in Table7-5, in the proposed technique 84.54% of all cases the error is less than 15 pixels, addition significant improvement from the previous techniques, and the correct extraction according to baseline in ground truth image was 2.4 %. However, more than 15 pixels the rate will be more than 92% as the suggested method determine "good" rows, extend the first good row out to column after determining skeletisation- axis from binary original image. This compare appears to be semi-optimal since the nature of the suggested baseline and the baseline in ground truth is lightly different. Pechwitz method based on the constructed linear regression of the features in polygonal approximated word skeleton. This method misses when the image contains very short words with separated characters only or when the image contains a warped word. In general, the error is less than 15 pixels in 64.21% of all situations and only produced correct extraction was 0.84% according to baseline in ground truth image. Boukerma method based on the determined group of points on subword to evaluation technique. However, this approach defeats when the subword includes large diacritics and small letters. Commonly, this method can obtain 1.75% according to baseline in ground truth image and 82.61 % of all states the error is smaller than 15 pixels. Besides, the defeat of this method is also given by the incorrect choice of local minima point situated at the bottom curve of small descenders.

Figure 7-14 presents the outcome of compare between Pechwitz, Boukerma and proposed baseline estimation methods with different baseline error in the case we using ground truth IFN/ENIT.

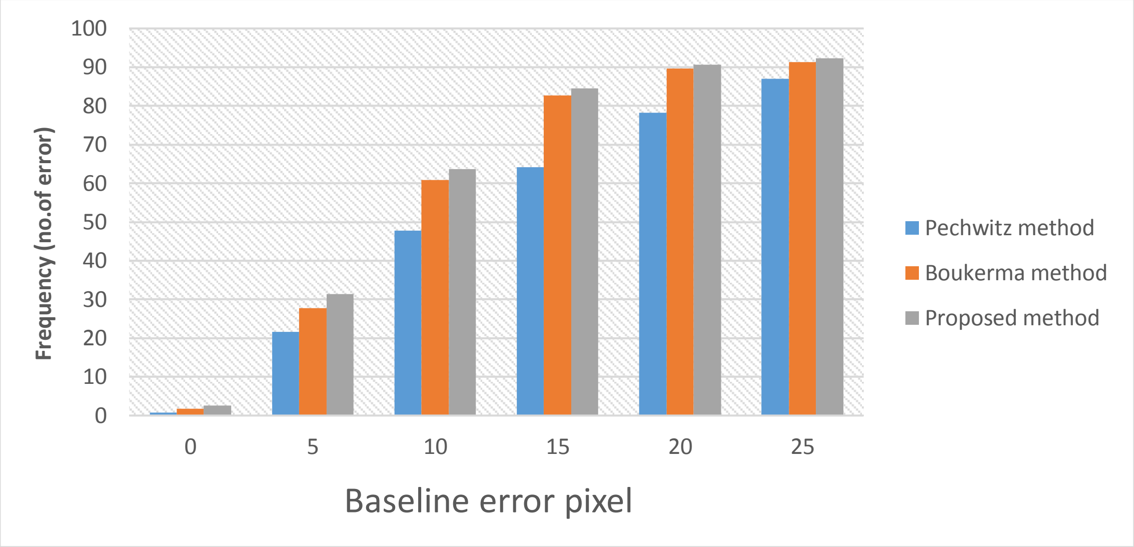


Figure 7-14: Compare between Pechwitz, Boukerma and proposed baseline estimation methods with different baseline error in the case by using ground truth IFN/ENIT.

Table 7-6: Results of Pechwitz, Boukerma, ground truth our dataset and suggested baseline evaluation methods with several baseline error

Baseline error in pixel	Pechwitz method	Boukerma method	Proposed method
0	0.91	1.85	2.6
5	21.8	27.85	31.53
10	47.89	60.96	63.8
15	64.41	82.81	84.74
20	78.37	89.76	90.93
25	87.9	91.9	92.87

As presented in Table7-6, in case when using ground truth our dataset in section 7.2. In the suggested technique 84.74% of all states the error is less than 15 pixels, addition significant improvement from the former techniques, and the precise extraction according to baseline in ground truth image was 2.6 %. However, more than 15 pixels the proportion will be more than 92% as the suggested method after determining skeletisation- axis from binary original image, calculated vertical Profile Projection over skeleton Image. Finally detect the blobs for every row, for every blobs detect baseline and link all the baseline of every blobs in each row. This compare appears to be semi-optimal because the nature of the recommended baseline and the baseline in ground truth is minimally different.

Pechwitz method failures when the image contains very short words with separated characters only or when the image contains a warped text line. In overall, the error is less than 15 pixels in 64.41% of all situations and only produced correct extraction was 0.91% according to baseline in ground truth image. Boukerma method defeats when the subword contains large diacritics and small letters. Generally, this technique can achieve 1.85% according to baseline in ground truth image and 82.81 % of all states the error is smaller than 15 pixels. In addition, the defeat of this technique is also given by the wrong choice of local minima point located at the bottom curve of small descenders or warped word. Figure 7-15presents the outcome of compare between Pechwitz, Boukerma and proposed baseline estimation methods with different baseline error in the case we using ground truth our dataset. In comparison with previous methods, the result of the suggested method gives the best performance in terms of visual experiment and analytical experiments.

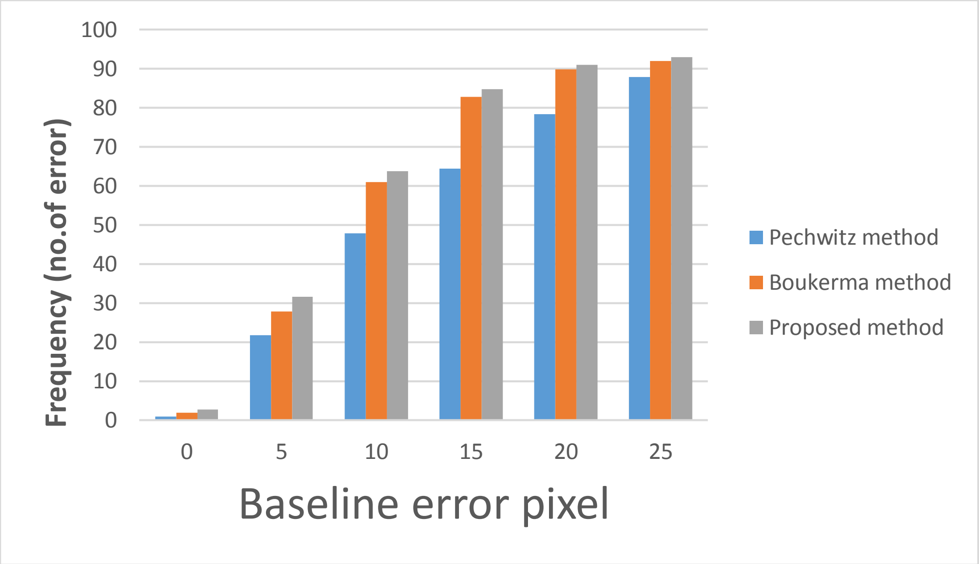


Figure 7-15: Compare between Pechwitz, Boukerma and proposed baseline estimation methods with different baseline error in the state by using ground truth of our dataset.

7.6 Implementation of selected procedure for dewarping

The dewarping method with global grid in the mentioned method in section 5.6.3 applies a transformation model to correct individual quadrilateral sub-grids (local grids) to a rectangular form. In this step results are contrasted with the state-of-the-art technique (NCSR), and commercial software (Book-Restorer). Several experiments have been taken out to assess the effectiveness of the suggested dewarping technique. The evaluation methodology applied in this research is based on supervised evaluation with (manually generated) ground-truth data. The experiments intend to contrast the performance of proposed approach with the current state of the art of geometrical correction methods. The experiment is performed with a different and representative sample of 25 arbitrarily warped historical Arabic manuscripts image with the complex layout from our dataset created section 7.2. Baselines on both the authentic warped document image and outcome document image are identified manually. The accuracy of each manuscript is estimated by the medium baseline straightness of the original and the corrected image is computed according to equation (6.4). Figure 7.16 presents the original image, the generated image after vertical and after horizontal dewarping. It can be seen from Figure 7.16a the original image has a smooth curl in the horizontal direction and warp in the vertical direction. These two deformities will be taken off by vertical dewarping (Figure 7.16b) and horizontal dewarping sequentially. The concluding image after vertical and horizontal dewarping is represented in Figure 7.16c.

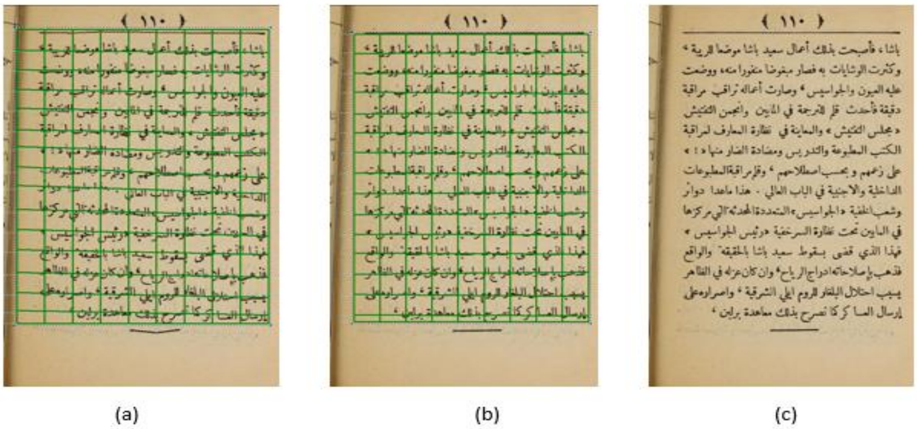


Figure 7-16:(a) original image (b) after vertical dewarping (c) final dewarped image (a grid is composed on all images to support watching the straight and curvy lines)

In the next stage, the results of the recommended method are compared with the state-of-the-art method (NCSR), PRImA method and commercial software (Book-Restorer). It can be seen from Figure 7.17 b that Book- Restorer results still have some distortion and page curl is not fixed perfectly. In order to input image in the NCSR technique should be binarised, the output image is binarised certainly. It can be seen from Figure 7.17c that page curl is not completely extracted from the NCSR technique. Figure7.17d depict the dewarped image by the PRImA solution which slightly different than, the technique recommended in figure7.17e.

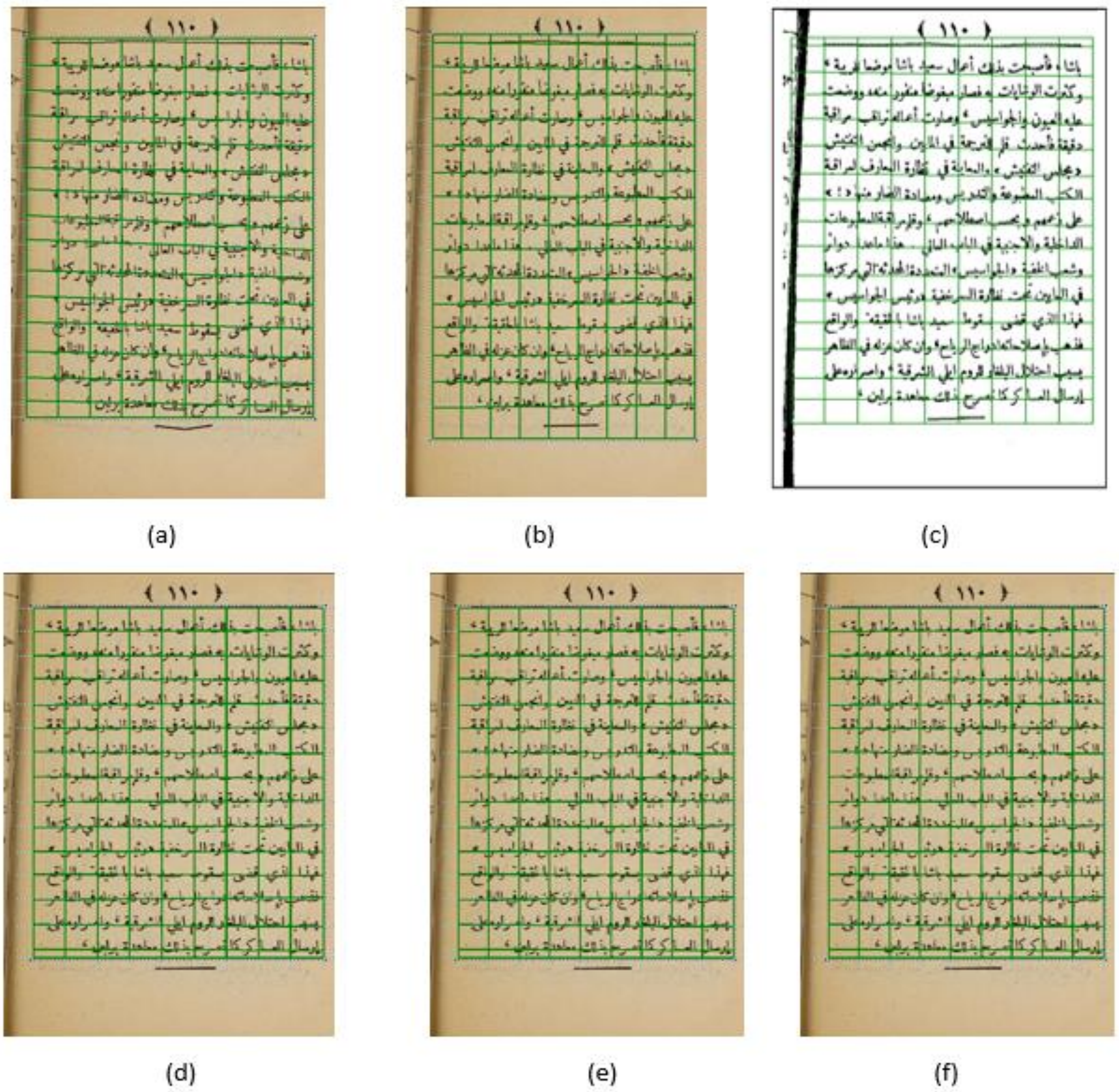
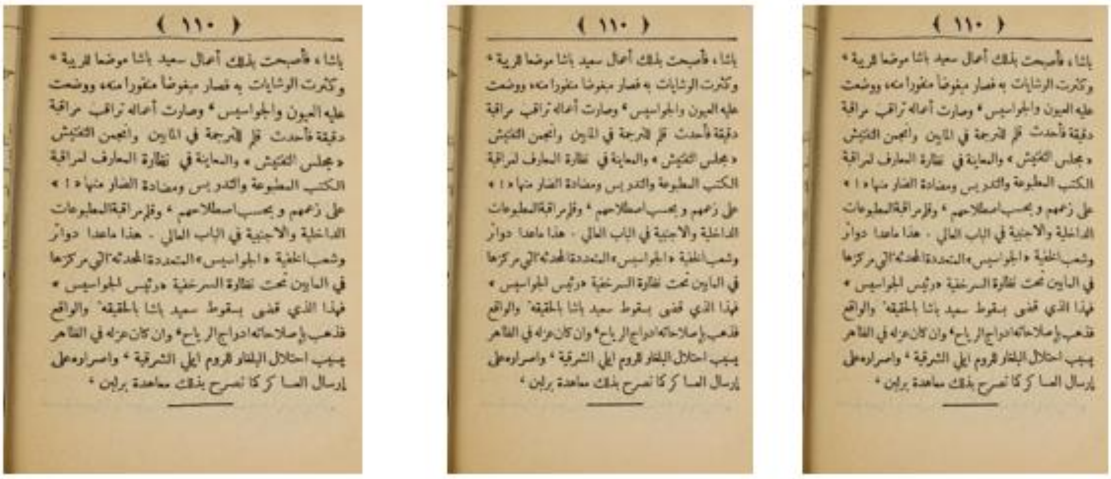


Figure 7-17:(a) Original image (b) Book-Restorer (c) NCSR (d) PRImA method (e) ground-truth (f) the proposed method.

It can be seen from Figure 7-18 and 7-19 depict the technique recommended and shows the result of the NCSR technique for comparison.

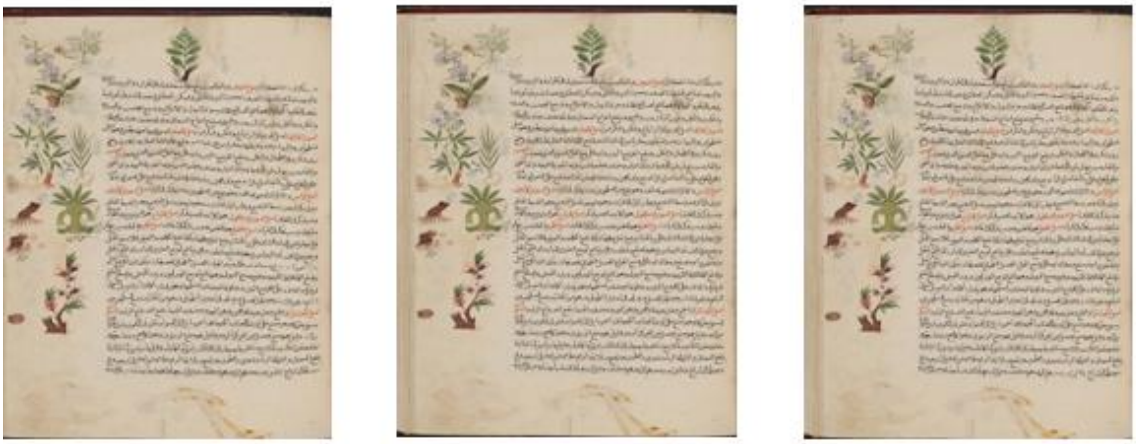


a) Original Image

Dewarped Image by b) Proposed Method

c) NCSR method

Figure 7-18: Dewarping of a sample image with simple layout



a) Original Image

Dewarped Image by b) Proposed Method

c) NCSR method

Figure 7-19: Dewarping of a sample image with complex layout.

7.6.1 Experiments to improve overall accuracy

To evaluate the results quantitatively, the supervised evaluation approach explained in section 6.3.4 was performed. The empirical outcomes for the test group are shown in Fig.7-20. It can be seen from Table 7-7 the results display that the recommended method and Prima method improve these 25 historical document images with arbitrary warping by growing the accuracy from an average of 70% to 90%. The NCSR technique can roughly get better the accuracy from an average of 70% to 85%.

Table 7-7: The empirical outcomes for the test set of image

Image ID	Original image	Book Restorer	NCSR method	Prima method	Proposed method
1	0.74	0.78	0.78	0.86	0.87
2	0.76	0.75	0.75	0.86	0.88
3	0.79	0.84	0.85	0.87	0.88
4	0.76	0.78	0.81	0.85	0.88
5	0.77	0.85	0.85	0.85	0.86
6	0.72	0.77	0.77	0.89	0.9
7	0.74	0.83	0.75	0.85	0.89
8	0.72	0.72	0.82	0.84	0.83
9	0.75	0.81	0.83	0.83	0.84
10	0.72	0.85	0.81	0.90	0.90
11	0.73	0.81	0.79	0.88	0.89
12	0.77	0.83	0.80	0.86	0.90
13	0.73	0.82	0.80	0.87	0.87
14	0.70	0.79	0.80	0.86	0.89
15	0.77	0.77	0.78	0.89	0.90
16	0.83	0.76	0.80	0.89	0.90
17	0.72	0.75	0.80	0.89	0.89
18	0.75	0.81	0.79	0.77	0.83
19	0.76	0.82	0.80	0.85	0.86
20	0.72	0.84	0.85	0.84	0.85
21	0.73	0.83	0.84	0.86	0.84
22	0.74	0.82	0.83	0.87	0.86
23	0.73	0.81	0.84	0.86	0.88
24	0.72	0.80	0.83	0.85	0.90
25	0.71	0.81	0.82	0.84	0.90

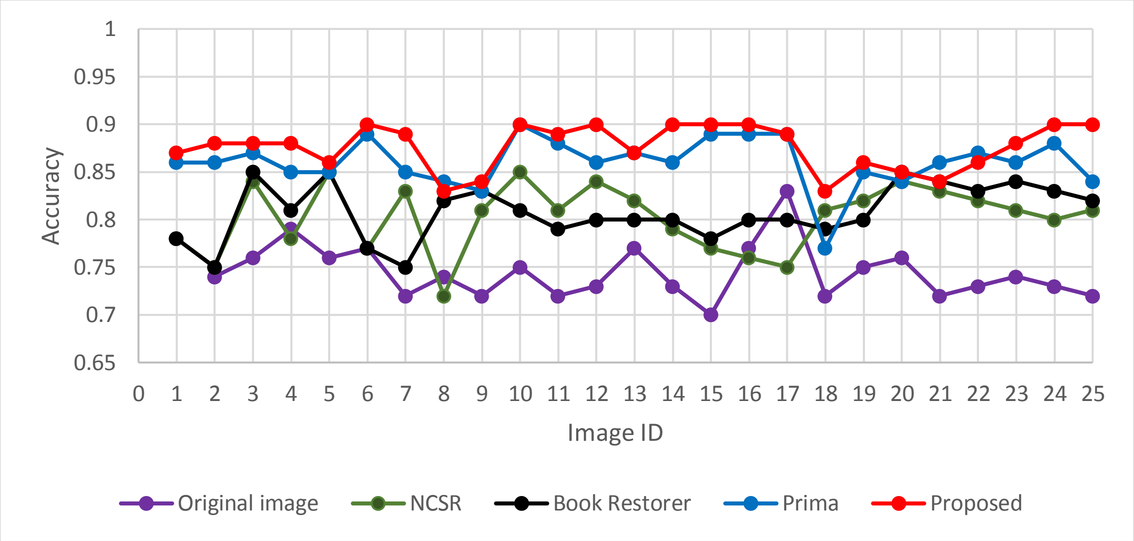


Figure 7-20: Evaluation results of NCSR, Book Restorer™ and proposed method on arbitrarily warped documents

The accuracy of each technique for each image is described in Table 7.7. In particular on some images (ID: 2, 4, 7, 12, 14, 15), recommended method has obvious enhancement than Prima method. Only on image 21, 22 recommended method slightly lower accuracy than Prima solution. Generally, the recommended technique performs better than both the NCSR technique and Book Restorer™ on rectifying document images with arbitrary warping.

7.6.2 Experiments to comparison Pixel error

The second experiment applying the distributions of points of the marked baselines in to produce a qualitative pixel error comparison between various geometrical correction techniques. This test looks forward to evaluating the recommended method for global rectification performance. The experiment is executed with the same sample as the previous experiment. Baselines on both the original warped document image and outcome document image are identified manually. The error of every document image is computed sequentially by standard mean of pixel error (SME), maximum pixel error (MPE), and standard derivation of pixel errors (STD) as described in Tables7. 8, 9, 10. Given that there are **N** baselines being identified in a document image, and each baseline is identified with **M** points; then there are completely **N *M** points being allotted in a document image.

For each baseline i, it can compute the average straight line Y, then the pixel error of every point can be determined in equation 7.1.

$$\lvert PixError_i^j = \lvert Y_i^j - Aver\_Y_i\rvert \qquad (7.1)$$

$PixError_i^j$: Pixel error of a distinct point j in the baseline i of the net.

Y_i^j: Effective value of Y axis of Point(i, j).

$Aver\_Y_i$: Average value of Y axis of all points in the baseline i.

7.6.3 The results

Tables 7-8, 7-9 and 7-10 show evaluation results on Mean of Pixel's error, Max of Pixels' error and standard derivation of pixel errors from four methods of the accuracy of images, which are Book Restorer, NCSR method, Prima method and Proposed method for 25 images.

7.6.3.1 The results computed on standard mean of pixel error (SME).

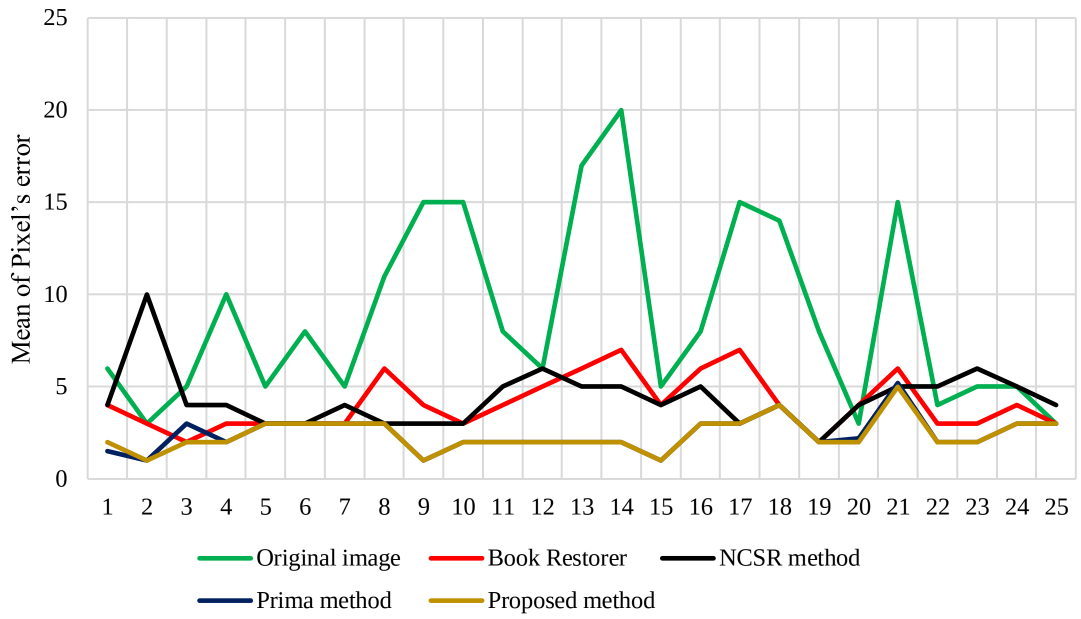
The shows data (Table 7-8) represent the error values in the 25 original images and the error values recorded on the same 25images after processing with these images using four methods which are Book restorer, NCSR, Prima and Proposed methods. The original image recorded the highest values, with an average of 8.8 that ranged from 3.0 for images number 2, 20 and 25 to 20.0 for image number 14.

NCSR method has the highest value with an average with 4.4 and it ranges between 2.0 (image 19) and 10.0 for the image number 2 (Table 7-8 and Figure 7.23). Therefore, it can be say that, NCSR method is considered as the least efficient of the four methods with the highest values of error in the marked based line. The method of Book restorer has relatively lower values (4.1) than the method (NCSR method) that ranges from 2.0 for the images number 3 and 19 to 7.0 for images number 14 and 17, which means, the efficiency of Book restorer method is not good enough. It is clear to see that, both methods of prima and proposed have the lowest values with an average of 2.44 and 2.4respectively. Where the prima method ranges from 1.0% (images 2, 9 and 15) to 5.2% (image 21) meanwhile, proposed method ranges between 1.0% (at the same images of the previous method (prima). Therefore, it can be say that, proposed method is the best method with the lowest value of error in the marked based line.

The results (Table 7-8) shows that, standard deviation values of the original image method is relatively high with 5.0 with comparing to the mean average (8.8). This values indicates the fact that, the accuracy values of the 25 images vary. On the other hand, the values of standard deviation of the remaining four methods Book restorer, NCSR, Prima and Proposed are relatively low with 1.5, 1.5, 0.9 and 0.9, respectively. It can be seen that, images number 21 and 14 that held the highest percentage with average of the four methods with only 7.24 and 7.20, respectively. On the other hand, image number 15 with 3.0.

Table 7-8:the accuracy of each method for each image computed by by standard mean of pixel error (SME)

Image ID	Original image	Book Restorer	NCSR method	Prima method	Proposed method
1	6	4	4	1.5	2
2	3	3	10	1	1
3	5	2	4	3	2
4	10	3	4	2	2
5	5	3	3	3	3
6	8	3	3	3	3
7	5	3	4	3	3
8	11	6	3	3	3
9	15	4	3	1	1
10	15	3	3	2	2
11	8	4	5	2	2
12	6	5	6	2	2
13	17	6	5	2	2
14	20	7	5	2	2
15	5	4	4	1	1
16	8	6	5	3	3
17	15	7	3	3	3
18	14	4	4	4	4
19	8	2	2	2	2
20	3	4	4	2.2	2
21	15	6	5	5.2	5
22	4	3	5	2	2
23	5	3	6	2	2
24	5	4	5	3	3
25	3	3	4	3	3



7-21:Evaluation outcomes on Mean of Pixel's Error

7.6.3.2 The results computed on maximum pixel error (MPE)

This table 7.9 shows the values of error that have obtained by four methods that have been computed by maximum pixel error (MPE). It is clear to see that, the accuracy image computed maximum pixel error (MPE) recorded the highest percentages in general comparing with both; computed by standard mean of pixel error (SME) and computed by standard derivation of pixel errors (STD).

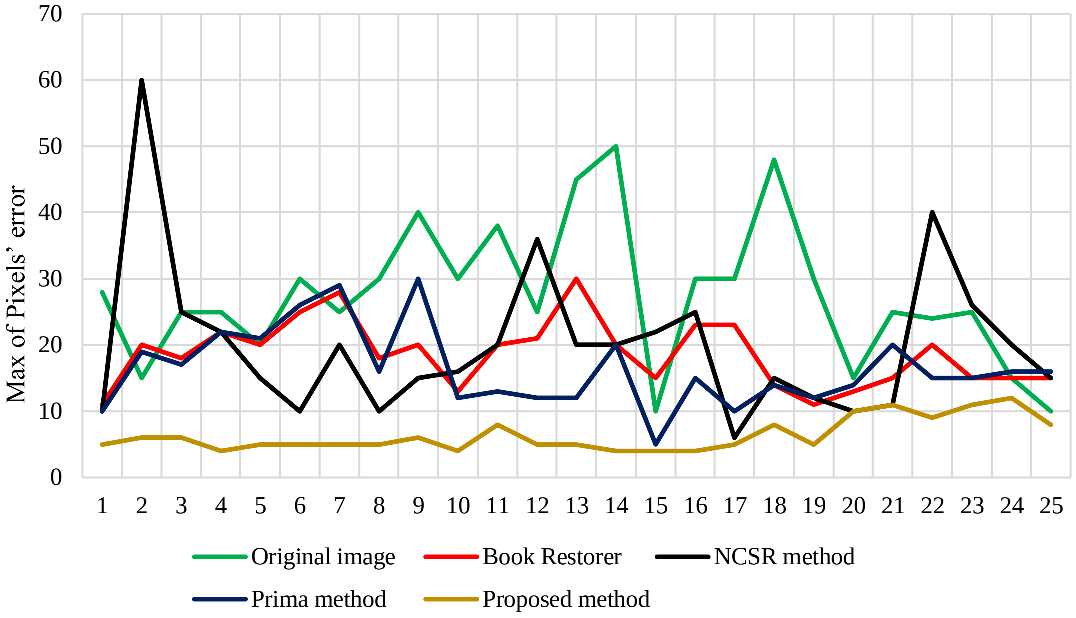
The original image recorded the highest values, with an average of 27.5 that ranged from 10.0 for images number 15 and 25 to 50.0 for image number 14. NCSR method have the highest values with an average of 19.6 that ranges between 6.0 (image 17) and 60.0 which is the highest values of error in the marked base line that have recorded for the image number 2 (Table 7-9 and Figure 7.22). The methods of Book restorer and Prima have high with relatively lower than NCSR method with an average of 18.8 and 16.6, respectively, that ranges from 11.0 (image 1 and 30) to 30 (image 13) and from 5.0 for the images number 15 to 30.0 for image number 9, respectively. Therefore, it can be say that, both methods of NCSR, Book restorer and Prima are considered as the least efficient of the four methods with the highest values of error in the marked based line. The proposed method has the lowest percentage with slightly higher than 7.0 that ranges from 4.0 that recorded for the images number 25 to 25.0 for images number 4, 10, 14, 15 and 16.

Therefore, it can be say that, proposed method is the best method with the lowest value of error in the marked based line

Table 7-9: the accuracy of each method for each image computed by maximum pixel error (MPE)

Image ID	Original image	Book Restorer	NCSR method	Prima method	Proposed method
1	28	11	10	10	5
2	15	20	60	19	6
3	25	18	25	17	6
4	25	22	22	22	4
5	20	20	15	21	5
6	30	25	10	26	5
7	25	28	20	29	5
8	30	18	10	16	5
9	40	20	15	30	6
10	30	13	16	12	4
11	38	20	20	13	8
12	25	21	36	12	5
13	45	30	20	12	5
14	50	20	20	20	4
15	10	15	22	5	4
16	30	23	25	15	4
17	30	23	6	10	5
18	48	14	15	14	8
19	30	11	12	12	5
20	15	13	10	14	10
21	25	15	11	20	11
22	24	20	40	15	9
23	25	15	26	15	11
24	15	15	20	16	12
25	10	15	15	16	8

On the other hand, image number 25 with an average of 10.0%. The results (Table 7-8) shows that, standard deviation values of both methods: original, prima and NCSR are high with 5.3%, 6.3% and 10.7%, respectively comparing to their averages of 18.8%, 16.6% and 25.7%, respectively. This value indicates the fact that, these data are more systematic. On the other hand, the values of standard deviation of the remaining three methods of NCSR, and Proposed are relatively low with 11.7%, 4.6 respectively comparing to their averages of 19.6% and 7.2% respectively. This value indicates the fact that, these data are more dispersed.



7-22:Evaluation outcomes on Max of Pixels' Error.

7.6.3.3 The results computed on standard derivation of pixel errors (STD)

The Table 7-10 shows the results in percentages have obtained by four methods that have been computed by standard derivation of pixel errors (STD). It is clear to see that, the accuracy image computed by standard derivation of pixel errors (STD) has recorded the lowest percentages in general comparing with the other ways; computed by standard mean of pixel error (SME) and by maximum pixel error (MPE).

As the previous, the data shows, original image recorded the highest percentages, with an average of 5.9 that ranged from 10.0 for images number 15 and 25 to 50.0 for image number 14. Both NCSR and Book restorer methods have the highest values of error in base line (Table 7-9) with an average of 3.7 and 3.1, respectively, that ranges from 1.8 (image 19) to 10.0 for the image number 2 and from 1.8 for the image number 10 to 6.0 for image number 19, respectively, (Table 7-10 and Figure 7.23). Therefore, it can be say that, both methods of NCSR and Book restorer are considered as the least efficient of the four methods with the highest values of error in the marked based line.

Table 7-10: The evaluation results computed on standard derivation of pixel errors (STD)

Image ID	Original image	Book Restorer	NCSR method	Prima method	Proposed method
1	5.8	2.1	3.5	2.0	1.9
2	3.0	4.0	10.0	1.8	1.8
3	5.0	2.8	4.0	2.0	1.9
4	6.0	2.6	4.0	2.0	2.2
5	4.0	2.8	2.4	2.2	1.9
6	5.0	3.8	3.0	2.4	2.3
7	4.0	3.6	5.0	1.8	2.0
8	7.5	3.4	2.0	1.8	2.0
9	8.0	2.1	2.2	2.2	2.2
10	8.0	1.8	2.6	2.1	2.4
11	5.8	2.2	3.0	2.2	2.1
12	4.2	3.6	4.2	2.1	1.8
13	9.0	4.0	3.0	2.2	1.8
14	11.0	5.0	3.8	2.2	1.7
15	2.4	3.5	3.0	1.0	1.5
16	6.0	5.0	5.0	2.4	1.8
17	8.0	6.0	2.0	1.8	1.8
18	9.0	2.8	2.4	2.5	2.0
19	6.0	2.4	1.8	2.0	2.0
20	2.2	2.0	4.0	2.2	2.0
21	11.0	2.7	2.5	2.4	2.4
22	5.0	2.4	5.0	1.8	2.0
23	5.2	2.2	5.0	2.0	1.5
24	5.2	2.6	5.0	2.0	1.5
25	2.2	2.4	4.0	2.0	2.0

On the other hand, both prima and proposed method have the lowest values of error in base line of images with 2.0 and 1.9 that ranges from 1.0 to 2.5 and from1.5 to 2.4, respectively. Therefore, it shows that, the proposed method is the best method with the lowest value of error in the marked based line.

Moreover, image number 15 with an average of 2.28. The results (Table 7-8) shows that, standard deviation values of the methods of Proposed, Prima, Book, NCSR and Original and are 0.25, 0.30, 1.07, 1.69 and 2.49, respectively. These values indicate the fact that, methods of Proposed, Prima are more systematic where, NCSR and Original more dispersed.

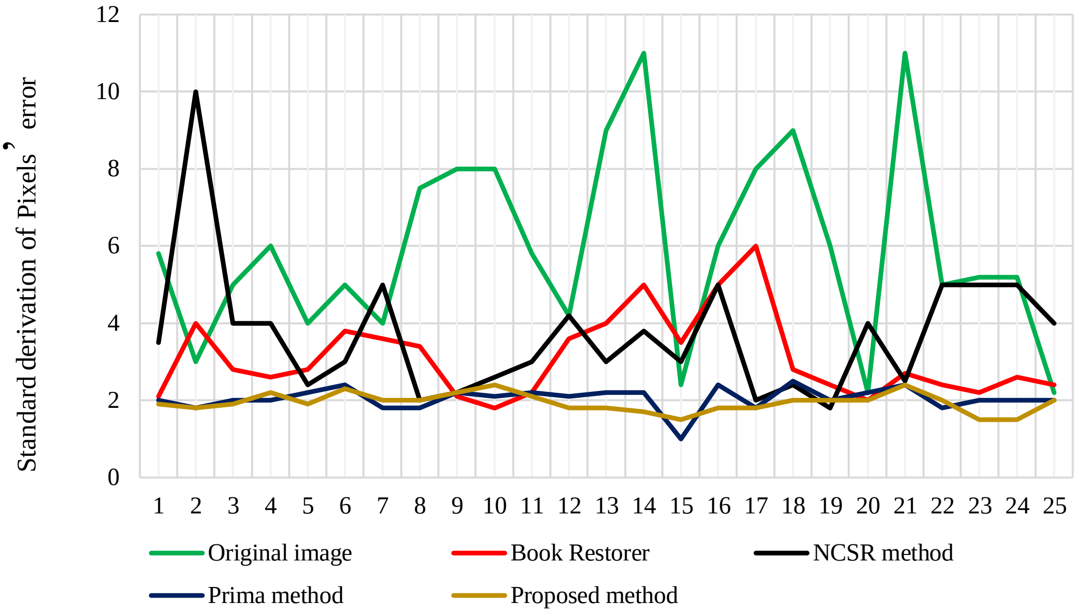


Figure 7-23: Evaluation outcomes on evaluation outcomes on STD.

Then, SME, MPE and STD of pixels error can be sequentially estimated and presented in Fig.7-21,22,23. The results show that the proposed method in brown line can considerably get better these 25 historical manuscripts images with arbitrary warping by reducing the pixel errors, for example SME from an average of 10 pixels to 2 pixels, MPE from an average of 20-30 pixels to 10-20 pixels, STD from an average of 4-6 pixels to 2 pixels. The NCSR technique in black line and the Book Restorer software in red line can also reduce the pixel errors, despite their execution, are not as good as the recommended one. In general, the proposed method performs better than both the NCSR method and Book Restorer on rectifying manuscript images with arbitrary warping. In effect, STD and SME reflect a global correction ability of dewarping techniques, which is analogous to the overall accuracy.MPE reflects a local correction ability of dewarping techniques, which describes if the methods produce some individual obscure points with worse impact on the authentic document image. Concerning these two aspects, the recommended techniques have both better performances than current state-of-the-art techniques.

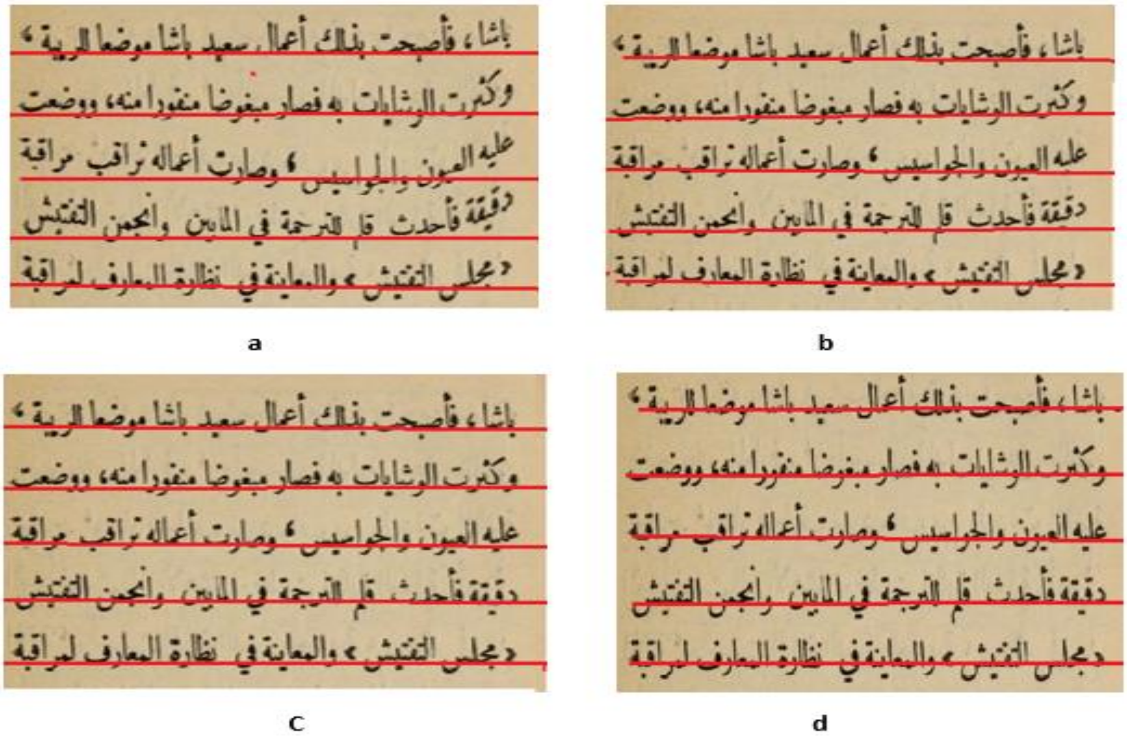


Figure 7-24: Performance of various correction methods processing image10, a) arbitrary warping region. b) Rectification by Book Restorer) Rectification by NCSR. d) Rectification by the recommended method.

Figure7-25 presents some section in the sample image 10 to show the performance of various dewarping techniques. It can be observed that the arbitrary warping has been considerably decreased. It can be noted that the arbitrary warping has been considerably improved. But, considering that the global region impacted by arbitrary warping is rather small in the image, the overall accuracies of these three dewarping techniques are slightly different.

7.6.4 Experiments to correction performance local dewarping

The third experiment is prepared to represent the local rectification performance of the recommended dewarping technique. With reference to the result of the first experiment, for every document image, with reference to the result of the first experiment, for every document image, there are N baselines being identified in a document image, and each baseline is computed the straightness accuracy; then the original document image and dewarped document image would have the correspond straightness accurateness of each baseline. So the refinement of straightness accuracy for each baseline is identified as:

$$Gain\_S\_Line_i = Dewarp\_S\_Line_i - Ori\ \_\ S\_Line_i \qquad (7.2)$$

Where:

$Gain\_S\_Line_i$: Development of straightness accuracy for the baseline i of the grid.

$Dewarp\_S\_Line_i$: Straightness accurateness for the baseline i of the grid in the dewarped image.

$Ori\ \_\ S\_Line_i$: Straightness accurateness for the baseline i of the grid in origin image.

Meantime, a standard is defined to test if the baseline is enhanced. If the development of straightness correctness is over 0.01, it indicates that the baseline is enhanced. If the betterment of straightness accuracy is least than -0.01, it means that the baseline is getting worse. If the development of straightness accuracy is among -0.01 and 0.01, it indicates that the baseline is kept the same.

The local rectification performance of the addressed dewarping technique can be shown by the percentage of them over the whole number of baselines.

Percentage_Line_Improved = Number_Line_Improved / N;

Percentage_Line_Worse = Number_Line_Worse / N;

Percentage_Line_Same= Number_Line_Same / N;

7.6.4.1 The experimental outcomes of line enhanced

Table 7-11 shows four methods of line enhanced, which are book restorer, NCSR method, prima method and proposed method for 25 images. It is clear to see that proposed method recorded the highest percentage that ranged from 75.0% (image number 12) to 97.3%.00 for both images 7 and 8. his is followed by prima method that has an average of 86.0% that ranged between 65.0% (image 5) and 97.0% for the same previous images both images 7 and 8 (Figure 7.25). The method of book restorer and NCSR method have relatively lower percentages 73.0% and 71.0% respectively. Where, the book restorer method ranged between 10.0% and 96.0% for images number 13 and 17 respectively.

Table 7-11: the percentages of line enhanced

Image ID	Book Restorer	NCSR method	Prima method	Proposed method
1	84.0%	80.0%	92.0%	95.0%
2	34.0%	30.0%	80.0%	85.0%
3	60.0%	75.0%	89.0%	90.0%
4	80.0%	60.0%	82.0%	86.0%
5	78.0%	62.0%	65.0%	85.0%
6	84.0%	80.0%	95.0%	95.0%
7	90.0%	60.0%	97.0%	97.0%
8	50.0%	94.0%	97.0%	97.0%
9	89.0%	77.0%	93.0%	95.0%
10	85.0%	97.0%	90.0%	85.0%
11	80.0%	90.0%	94.0%	94.0%
12	80.0%	38.0%	70.0%	75.0%
13	96.0%	87.0%	95.0%	95.0%
14	95.0%	97.0%	95.0%	95.0%
15	60.0%	70.0%	90.0%	93.0%
16	70.0%	30.0%	90.0%	93.0%
17	10.0%	97.0%	95.0%	96.0%
18	93.0%	60.0%	68.0%	80.0%
19	80.0%	70.0%	96.0%	95.0%
20	92.0%	94.0%	93.0%	93.0%
21	80.0%	97.0%	93.0%	95.0%
22	60.0%	50.0%	70.0%	85.0%
23	65.0%	60.0%	73.0%	85.0%
24	68.0%	60.0%	74.0%	82.0%
25	69.0%	58.0%	70.0%	85.0%

With respect of the NCSR method it recorded relatively lower percentages that ranged from 30.0% for both images number 2 and 16 to 97.0% for the images number 10, 17 and 21. It indicates that Image 14 and 13 held the highest percentage with average of the four methods with 96.0% and 93.0%, respectively. On the other hand, image number 2 held the least average which was 57.0%.

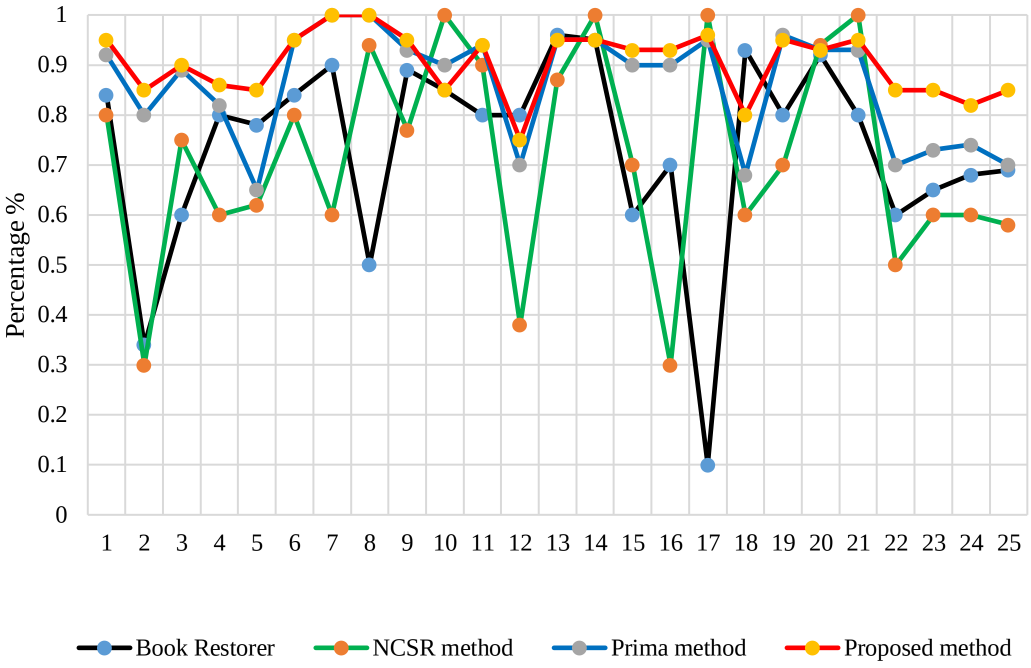


Figure 7-25:proportion of improved lines.

7.6.4.2 The experimental outcomes of line unchanged

Table 7-12 shows four methods of line unhanged, which are book restorer, NCSR method, prima method and proposed method for 25 images. It is clear to see that, the proportions of line unchanged, generally recorded lower percentages with comparing with the previous of “the percentages of line enhanced” where data (table 7:12) recorded was completely opposite to the previous method (proposed method) where the proposed method has recorded the lowest percentage with an average of 89.0% that ranged from 2.0% (image number 1) to 20.0% for images number 18. The NCSR method recorded the highest percentage with an average of 23.0% that ranged from 0.06 (images number 8 and 10) to 45.0% for both images 24 and 25.

Table 7-12: the proportions of line unchanged

Image ID	Book Restorer	NCSR method	Prima method	Proposed method
1	9.0%	9.0%	2.0%	2.0%
2	34.0%	30.0%	8.0%	7.0%
3	30.0%	8.0%	5.0%	5.0%
4	12.0%	30.0%	10.0%	6.0%
5	22.0%	34.0%	36.0%	18.0%
6	16.0%	20.0%	8.0%	5.0%
7	14.0%	30.0%	5.0%	5.0%
8	20.0%	6.0%	5.0%	5.0%
9	8.0%	20.0%	10.0%	10.0%
10	13.0%	6.0%	11.0%	5.0%
11	14.0%	18.0%	10.0%	9.0%
12	14.0%	38.0%	30.0%	12.0%
13	8.0%	7.0%	8.0%	5.0%
14	5.0%	8.0%	10.0%	10.0%
15	40.0%	30.0%	10.0%	5.0%
16	18.0%	50.0%	12.0%	10.0%
17	90.0%	8.0%	15.0%	5.0%
18	14.0%	38.0%	36.0%	20.0%
19	20.0%	28.0%	16.0%	5.0%
20	15.0%	18.0%	13.0%	10.0%
21	18.0%	10.0%	12.0%	12.0%
22	26.0%	10.0%	25.0%	15.0%
23	40.0%	45.0%	30.0%	10.0%
24	28.0%	45.0%	24.0%	12.0%
25	29.0%	46.0%	26.0%	15.0%

This is followed by book restorer method that has an average of 22.0% that ranged between 5.0% (image 14) and 90.0% for the image number 17 (Figure 7.26). The method of prima method had relatively lower percentages with an average of 15.0% that ranged from 2.0% for image number 1 to 0.36 for both images number 5 and 18. It can be seen that images number 22 and 16 that held the highest percentage with average of the four methods with only 31.0% and 29.0%, respectively. On the other hand, images number 12, 13, 7 and 9 held the least average which were 7.0%, 8.0% and 9.0%. The results (Table 7-12) shows that, standard deviation values for the four methods; are 17.1%, 14.6%, 10.0% and 4.6% for the methods of Book, NCSR, Prima and Original, respectively. This value indicates the fact that, all methods are more dispersed, methods of Book, NCSR, Prima, with more symmetric for Original image.

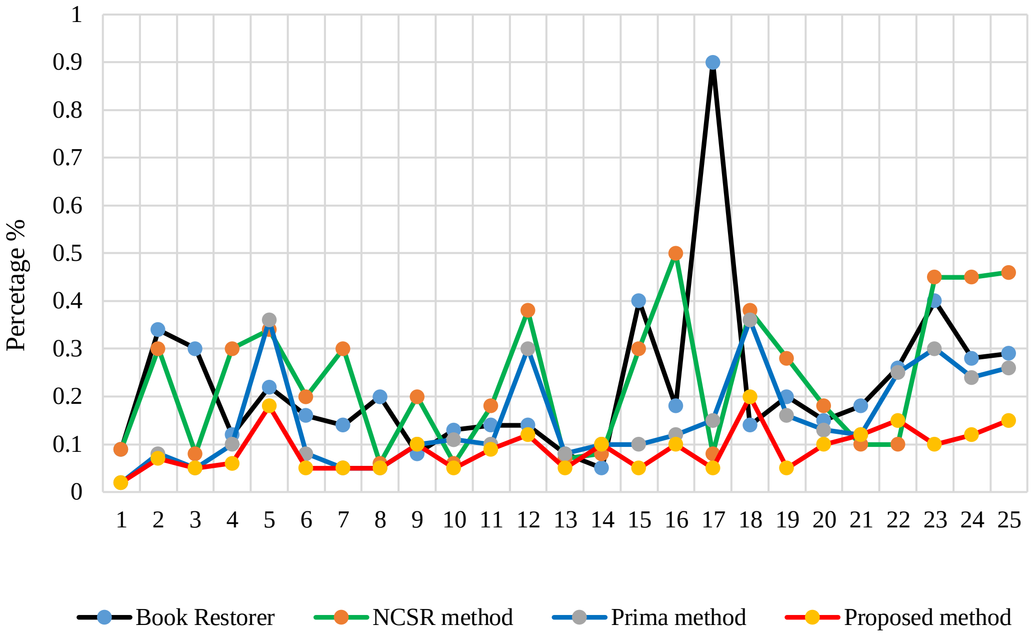


Figure 7-26: Proportion of unchanged lines.

7.6.4.3 The experimental outcomes of damaged lines.

It is clear to see that (Table 7-13), in general the percentages of "damaged lines recorded the lowest percentages". The data shows that, proposed method recorded the lowest percentage with an average of 9.0% that ranged from 5.0% (at images number 1, 2, 6, 7, 8, 11, 13, 16 and 17) to 25.0% for image number 25. This is followed by prima method that has an average of 10.0% that ranged between 5.0% (images number 6, 7, 8, 11, 13, 17, 19 and 20) and 28.0% for the same previous image number 25 (Figure 7.23). The method of book restorer and NCSR method have relatively higher percentages than both methods of "prima and proposed" with averages of 15.0% for each method. Where, the book restorer method ranged between 4.0% and 30.0% for images number 7 and 22 respectively. With respect of the NCSR ranged from 5.0% to 50.0% for the images number 14, and 22, respectively. It can be seen that images number 22 held the highest percentage with average of the four methods with 26.0%. On the other hand, image number 13 held the least average which was 6.0%.

Table 7-13: the percentages of damaged lines

Image ID	Book Restorer	NCSR method	Prima method	Proposed method
1	0.09	0.08	0.06	0.05
2	0.18	0.22	0.08	0.05
3	0.1	0.22	0.06	0.05
4	0.12	0.12	0.14	0.14
5	0.1	0.1	0.11	0.11
6	0.12	0.1	0.05	0.05
7	0.04	0.2	0.05	0.04
8	0.28	0.1	0.05	0.05
9	0.14	0.12	0.09	0.05
10	0.13	0.08	0.11	0.11
11	0.11	0.09	0.05	0.05
12	0.12	0.15	0.12	0.12
13	0.05	0.08	0.05	0.05
14	0.12	0.05	0.12	0.12
15	0.2	0.14	0.13	0.13
16	0.1	0.28	0.1	0.05
17	0.25	0.05	0.05	0.05
18	0.14	0.18	0.13	0.13
19	0.13	0.18	0.05	0.05
20	0.2	0.07	0.05	0.05
21	0.2	0.08	0.09	0.09
22	0.3	0.5	0.12	0.12
23	0.16	0.12	0.16	0.15
24	0.2	0.13	0.2	0.19
25	0.21	0.2	0.28	0.25

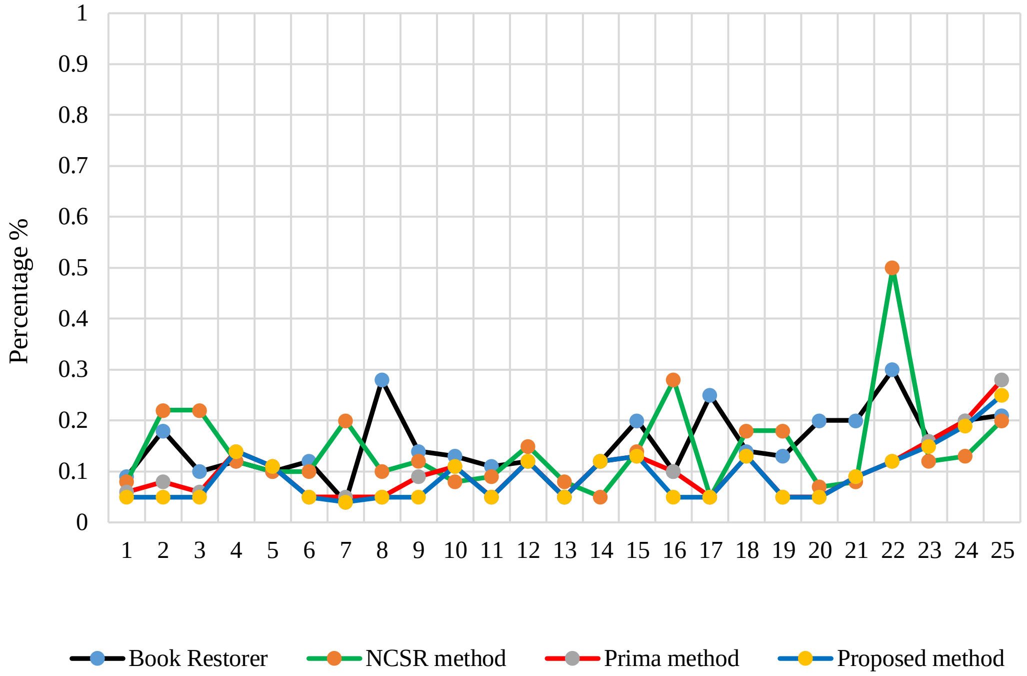


Figure 7-27:: Proportion of damaged lines

7.7 Summary

This chapter describes creating the dataset and summarises the techniques created in the research and the evaluation of their results. To conclude this chapter, the researcher carried three experiments in order to confirm the recommended technique performs better than both the NCSR technique and Book Restorer on rectifying document images with arbitrary warping. The produced outcomes from these experiments are presented, described and discussed in this chapter.

Chapter 8 - Conclusion and future work

8.1 Overview

The final chapter of this Thesis pertains the overall examination of designing a de-warping system for historical Arabic document images. This chapter summarizes the techniques performed in this research and their results followed by a review based on the aims and objectives presented in Chapter 1.

8.2 Summary

In summary, this Thesis includes three main parts: firstly, the theoretical background of the geometric correction issues, secondly, the description of the development of a new technique to solve this problem, and finally, an evaluation of the recommended method. More specifically, the various chapters of this Thesis were as follows.

In Chapter 1, after briefing introduction, the objectives were set up to achieve the aims of this research, the theoretical background of this research was presented in Chapters 2

Chapter 3 provides a comprehensive survey of the literature associated with geometric correction issues. Chapter4 illustrates the proposed methodology procedure and framework adopted in the research. Chapter 5 describes in detailed the proposed de-warping method which is able to deal with the combination of geometric distortions including page curl, arbitrary warping and folds in historical Arabic documents.

Chapter 6 shows the evaluation of the proposed methodology by performing the supervised evaluation and describe the significance of this evaluation approach in comparison with other evaluation techniques. Eventually, in chapter 7 the experimental outcomes and evaluation of the proposed system were introduced.

8.3 Aims and Objects revisited

The aims of this research were to identify a new efficient technique to rectify geometric deformations in historical Arabic documents including page curl, arbitrary warping and fold. The proposed system is capable to deal with various geometric distortions independently and in any conditions.

The system is capable of de-warping may types of deformation without performing any presumption about the type of distortions in the input image, and at the same time, it presents a very high average accuracy of 90% on a very diverse dataset of historical document images. The adaptability and efficiency of the system are obtained by different methods. Firstly, by using two cases of overlapping and weighted L2-norm, the system is capable of flexibly segmenting text line and can be used to various styles of writing whether it be printed or handwritten and able to address multilingual documents. Secondly, the suggested baseline detection process identifies descenders, ascenders and any unforeseen geometrical distortions, as well as detecting baselines precisely; in other words, unlike existing text line detection methods in the literature, the present method identifies and decreases error to an absolute minimum. Finally, with the assistance of baseline detection by Voronoi Diagram technique and both local and global de-warping, the dewarping system precisely flattens any deformed page, included documents with global shrinkage or local shrinkage in the letters level.

The suggested dewarping system is strong in existence of noise created due to bleed ink, show-through, spot and filth; indeed, any noise that residues after the binarisation step are extracted by a compound of noise removal, broken character restoration, and reference point detection steps. Broken characters affected by faded ink or special typefaces restored in the broken parts restoration step. The components bigger than the normal size, for instance, drop letters, ornamental letters or decoration will be extracted in the reference point detection of text line segmentation step. The components bigger than the normal size, for instance, drop letters, ornamental letters or decoration will be extracted in the reference point detection of text line segmentation step. In addition, since the proposed text line segmentation method uses almost all the components of the page, it works when various styles of writing and different font sizes coexist. Finally, the challenge of dense sections and irregular spaces between text lines and words are addressed by applying two key properties of the text line segmentation approach. Since the recommended method does not set a condition for any special language it can cope with multilingual documents. Based on accurate baseline detection, a performance evaluation methodology for dewarping tailored for historical documents is introduced in this research too; currently available evaluation processes frequently apply OCR, but in most situations, this is unsuitable for use with historical documents. The recommended evaluation method is, therefore, a precise and adjustable choice for both historical and modern documents

8.4 Contribution of the Study

Historical documents are considered one of the most important human wealth and a source of intellectual creation. Unfortunately, due to ageing impacts, multiple noises and arbitrary warping are found in the document image. Moreover, difficulties for several images of antique documents show defects of warps and curvatures of text lines.

This research contributes the development of a new dewarping system, which is the first for Historical Arabic Documents. This result is rectified based on different procedures that have been tested and evaluated using different classifiers. For this system, a dataset consisting of different warped historical Arabic documents has been created. One of the important advantages of this dataset is the important amount of detailed and high-quality ground truth available and the scope for its use. By warping text lines close to the ones above and text lines, as well as overlap area.

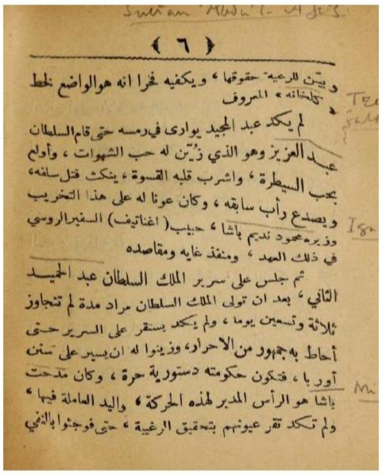
8.5 Research limitations and future work

Although the suggested method in this research is highly precise and adaptable, it still has some shortcomings and many aspects can be developed in the future.

The main restriction of the technique is its dependency on text lines and their baselines. When the manuscript image comprises a picture only or text inside a picture, the proposed de-warping system cannot be used effectively. To improve the adaptability of the system, one resolution is to make page layout analysis first of all and then use the recommended de-warping system in text segments. When the page consists peripheral texts, the proposed dewarping system cannot be applied directly to the whole page except if the text lines in the different rows are aligned and determined at the same level. Indeed, if the system is further improved to integrate marginal text segmentation, the suggested baseline detection can be used direct. Then by using the proposed technique based on VD and estimated the baseline in each part and in the whole page, the dewarping can be performed easily. Irregular lighting and shading in some sections of the page may reduce the accuracy of the system in identifying precise text lines and consequently de-warping the page. Shading rectification before any geometric correction will improve the accuracy of the system. Besides, if both the geometric and shading aspects are fixed, the final image is more suitable and benefits from a higher quality in print-on-demand.

Appendixes

Appendix A Different samples of Historical Arabic Documents



Sultan 'Abdu'l-'Aziz

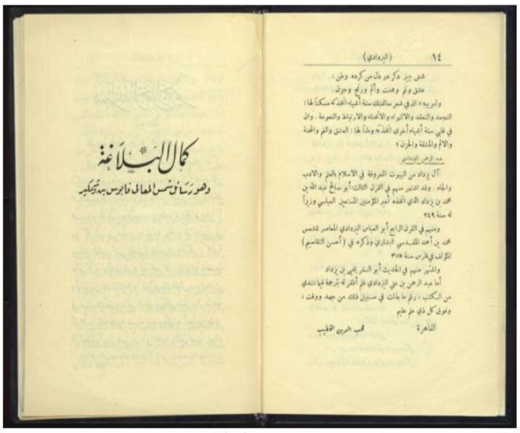
﴿ ٦ ﴾

ويبين للرعية حقوقها ، ويكفيه فخرا انه هوالواضع لخط
كلخانه » المعروف

لم يكد عبد المجيد يوارى في رمسه حتى قام السلطان
عبد العزيز وهو الذي زيّن له حب الشهوات ، وأولع
بحب السيطرة ، واشرب قلبه القسوة ، ينكث فتل سلفه
ويصدع رأب سابقه ، وكان عونا له على هذا التخريب
وزيره محمود نديم باشا ، حبيب (اغناتيف) السفيرالروسي
في ذلك العهد ، ومنفذ غايه ومقاصده

ثم جلس على سرير الملك السلطان عبد الحميد
الثاني ، بعد ان تولى الملك السلطان مراد مدة لم تتجاوز
ثلاثة وتسعين يوما ، ولم يكد يستقر على السرير حتى
أحاط به جمهور من الاحرار ، وزينوا له ان يسير على سنن
آوربا ، فتكون حكومته دستورية حرة ، وكان مدحت
باشا هو الرأس المدبر لهذه الحركة ، واليد العاملة فيها ،
ولم تكد تقر عيونهم بتحقيق الرغبة ، حتى فوجئوا بالنفي

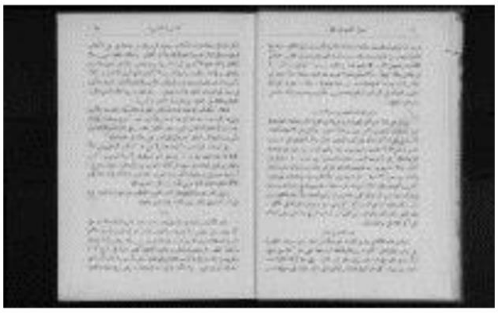
(a)



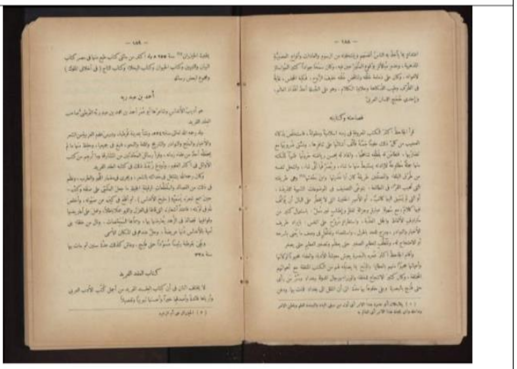
كمال البلاغة

وهو رسائل شمس المعالي قابوس بن وشمكير

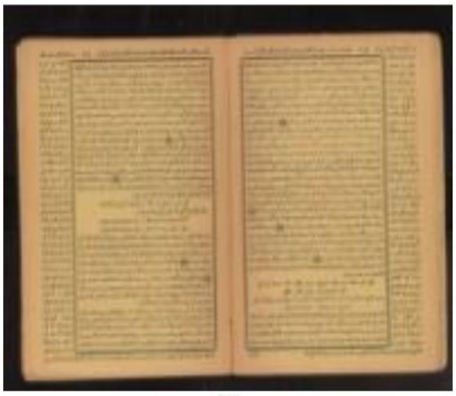
(b)



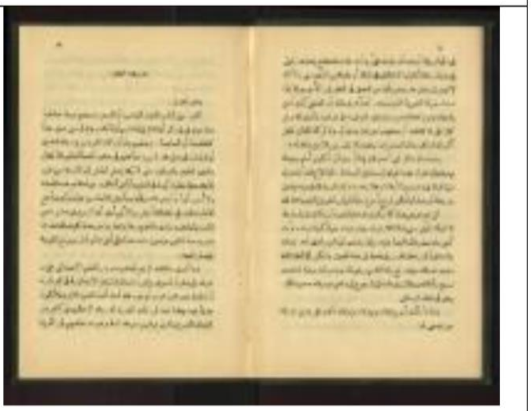
(d)



(e)

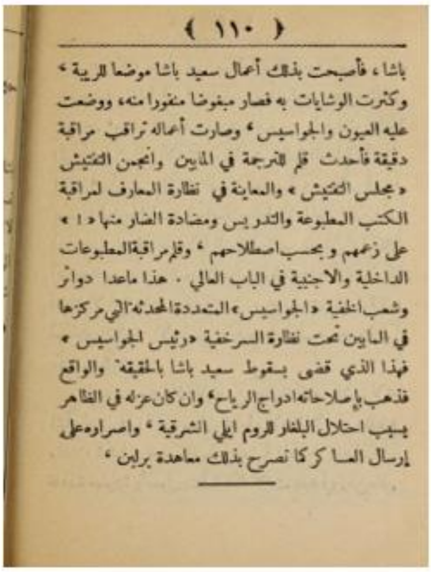


(f)



(g)

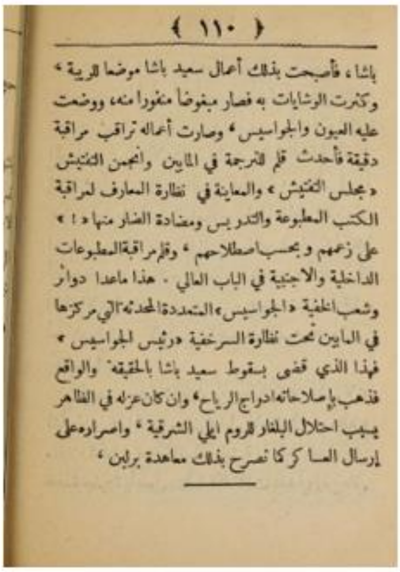
Appendix B synthetic warped images



{ ١١٠ }

باشا ، فأصبحت بذلك أعمال سعيد باشا موضعا للريبة ؛
وكثرت الوشايات به فصار مبغوضا منفورا منه، ووضعت
عليه العيون والجواسيس ؛ وصارت أعماله تراقب مراقبة
دقيقة فأحدث قلم الترجمة في المابين وانجمن التفتيش
« مجلس التفتيش » والمعاينة في نظارة المعارف لمراقبة
الكتب المطبوعة والتدريس ومضادة الضار منها « ! »
على زعمهم وبحسب اصطلاحهم ؛ وقلم مراقبة المطبوعات
الداخلية والاجنبية في الباب العالي . هذا ماعدا دوائر
وشعب الخفية « الجواسيس » المتعددة المحدثة التي مركزها
في المابين تحت نظارة السرخفية « رئيس الجواسيس »
فهذا الذي قضى بسقوط سعيد باشا بالحقيقة ؛ والواقع
فذهب بإصلاحاته ادراج الرياح ؛ وان كان عزله في الظاهر
بسبب احتلال البلغار الروم ايلي الشرقية ؛ واصراره على
إرسال العساكر كما تصرح بذلك معاهدة برلين ؛

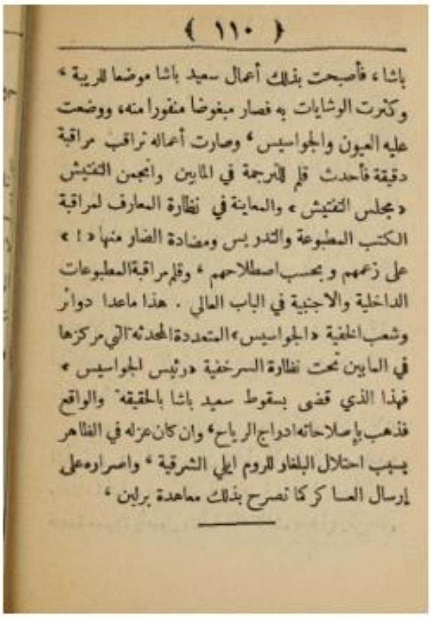
5% warped



{ ١١٠ }

باشا ، فأصبحت بذلك أعمال سعيد باشا موضعا للريبة ؛
وكثرت الوشايات به فصار مبغوضا منفورا منه، ووضعت
عليه العيون والجواسيس ؛ وصارت أعماله تراقب مراقبة
دقيقة فأحدث قلم الترجمة في المابين وانجمن التفتيش
« مجلس التفتيش » والمعاينة في نظارة المعارف لمراقبة
الكتب المطبوعة والتدريس ومضادة الضار منها « ! »
على زعمهم وبحسب اصطلاحهم ؛ وقلم مراقبة المطبوعات
الداخلية والاجنبية في الباب العالي . هذا ماعدا دوائر
وشعب الخفية « الجواسيس » المتعددة المحدثة التي مركزها
في المابين تحت نظارة السرخفية « رئيس الجواسيس »
فهذا الذي قضى بسقوط سعيد باشا بالحقيقة ؛ والواقع
فذهب بإصلاحاته ادراج الرياح ؛ وان كان عزله في الظاهر
بسبب احتلال البلغار الروم ايلي الشرقية ؛ واصراره على
إرسال العساكر كما تصرح بذلك معاهدة برلين ؛

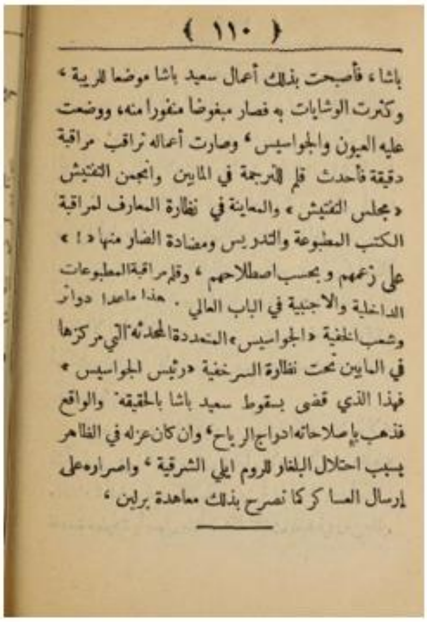
10% warped



{ ١١٠ }

باشا ، فأصبحت بذلك أعمال سعيد باشا موضعا للريبة ؛
وكثرت الوشايات به فصار مبغوضا منفورا منه، ووضعت
عليه العيون والجواسيس ؛ وصارت أعماله تراقب مراقبة
دقيقة فأحدث قلم الترجمة في المابين وانجمن التفتيش
« مجلس التفتيش » والمعاينة في نظارة المعارف لمراقبة
الكتب المطبوعة والتدريس ومضادة الضار منها « ! »
على زعمهم وبحسب اصطلاحهم ؛ وقلم مراقبة المطبوعات
الداخلية والاجنبية في الباب العالي . هذا ماعدا دوائر
وشعب الخفية « الجواسيس » المتعددة المحدثة التي مركزها
في المابين تحت نظارة السرخفية « رئيس الجواسيس »
فهذا الذي قضى بسقوط سعيد باشا بالحقيقة ؛ والواقع
فذهب بإصلاحاته ادراج الرياح ؛ وان كان عزله في الظاهر
بسبب احتلال البلغار الروم ايلي الشرقية ؛ واصراره على
إرسال العساكر كما تصرح بذلك معاهدة برلين ؛

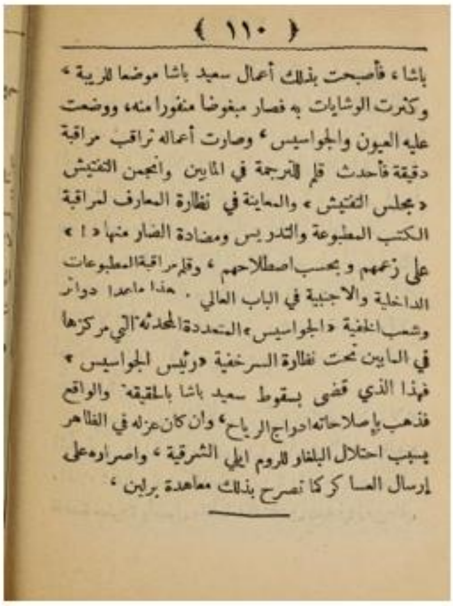
20% warped



{ ١١٠ }

باشا ، فأصبحت بذلك أعمال سعيد باشا موضعا للريبة ؛
وكثرت الوشايات به فصار مبغوضا منفورا منه، ووضعت
عليه العيون والجواسيس ؛ وصارت أعماله تراقب مراقبة
دقيقة فأحدث قلم الترجمة في المابين وانجمن التفتيش
« مجلس التفتيش » والمعاينة في نظارة المعارف لمراقبة
الكتب المطبوعة والتدريس ومضادة الضار منها « ! »
على زعمهم وبحسب اصطلاحهم ؛ وقلم مراقبة المطبوعات
الداخلية والاجنبية في الباب العالي . هذا ماعدا دوائر
وشعب الخفية « الجواسيس » المتعددة المحدثة التي مركزها
في المابين تحت نظارة السرخفية « رئيس الجواسيس »
فهذا الذي قضى بسقوط سعيد باشا بالحقيقة ؛ والواقع
فذهب بإصلاحاته ادراج الرياح ؛ وان كان عزله في الظاهر
بسبب احتلال البلغار الروم ايلي الشرقية ؛ واصراره على
إرسال العساكر كما تصرح بذلك معاهدة برلين ؛

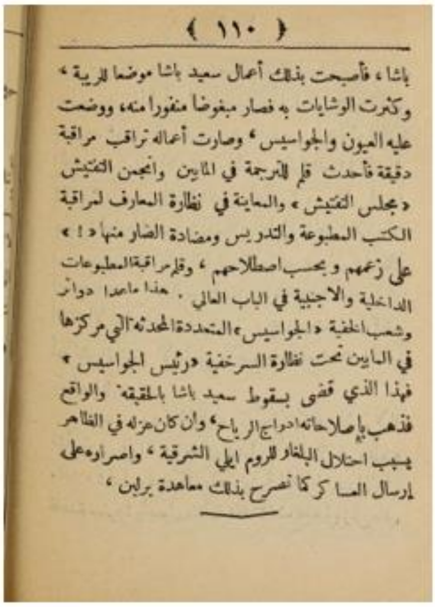
30% warped



{ ١١٠ }

باشا ، فأصبحت بذلك أعمال سعيد باشا موضعا للريبة ،
وكثرت الوشايات به فصار مبغوضا منفورا منه، ووضعت
عليه العيون والجواسيس ، وصارت أعماله تراقب مراقبة
دقيقة فأحدث قلم للترجمة في المابين وانجمن التفتيش
«مجلس التفتيش» والمعاينة في نظارة المعارف لمراقبة
الكتب المطبوعة والتدريس ومضادة الضار منها «!»
على زعمهم و بحسب اصطلاحهم ، وقلم مراقبة المطبوعات
الداخلية والاجنبية في الباب العالي . هذا ماعدا دوائر
وشعب الخفية «الجواسيس» المتعددة المحدثة التي مركزها
في المابين تحت نظارة السرخفية «رئيس الجواسيس»
فهذا الذي قضى بسقوط سعيد باشا بالحقيقة والواقع
فذهب بإصلاحاته ادراج الرياح ، وان كان عزله في الظاهر
بسبب احتلال البلغار للروم ايلي الشرقية ، واصراره على
إرسال العساكر كما تصرح بذلك معاهدة برلين ،

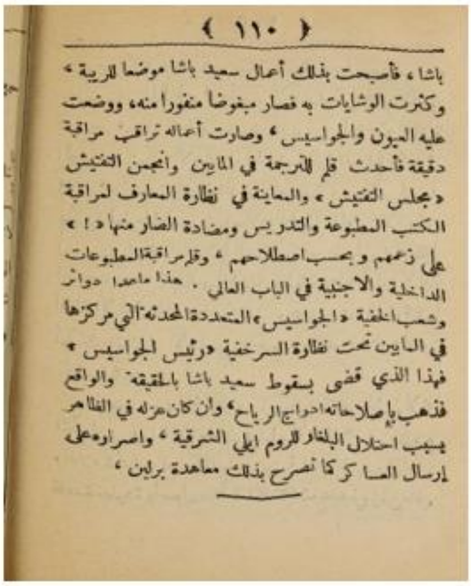
40% warped



{ ١١٠ }

باشا ، فأصبحت بذلك أعمال سعيد باشا موضعا للريبة ،
وكثرت الوشايات به فصار مبغوضا منفورا منه، ووضعت
عليه العيون والجواسيس ، وصارت أعماله تراقب مراقبة
دقيقة فأحدث قلم للترجمة في المابين وانجمن التفتيش
«مجلس التفتيش» والمعاينة في نظارة المعارف لمراقبة
الكتب المطبوعة والتدريس ومضادة الضار منها «!»
على زعمهم و بحسب اصطلاحهم ، وقلم مراقبة المطبوعات
الداخلية والاجنبية في الباب العالي . هذا ماعدا دوائر
وشعب الخفية «الجواسيس» المتعددة المحدثة التي مركزها
في المابين تحت نظارة السرخفية «رئيس الجواسيس»
فهذا الذي قضى بسقوط سعيد باشا بالحقيقة والواقع
فذهب بإصلاحاته ادراج الرياح ، وان كان عزله في الظاهر
بسبب احتلال البلغار للروم ايلي الشرقية ، واصراره على
إرسال العساكر كما تصرح بذلك معاهدة برلين ،

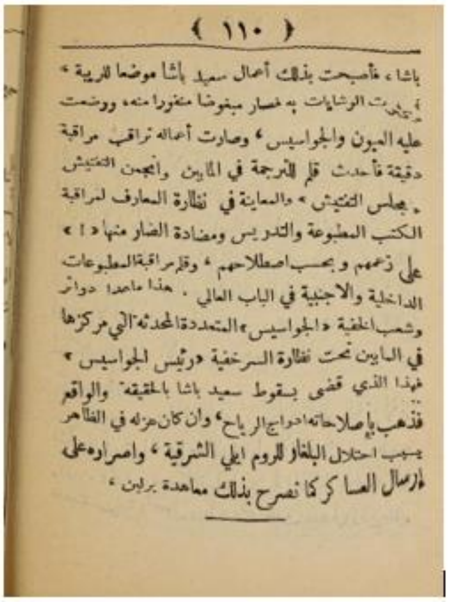
50% warped



{ ١١٠ }

باشا ، فأصبحت بذلك أعمال سعيد باشا موضعا للريبة ،
وكثرت الوشايات به فصار مبغوضا منفورا منه، ووضعت
عليه العيون والجواسيس ، وصارت أعماله تراقب مراقبة
دقيقة فأحدث قلم للترجمة في المابين وانجمن التفتيش
«مجلس التفتيش» والمعاينة في نظارة المعارف لمراقبة
الكتب المطبوعة والتدريس ومضادة الضار منها «!»
على زعمهم و بحسب اصطلاحهم ، وقلم مراقبة المطبوعات
الداخلية والاجنبية في الباب العالي . هذا ماعدا دوائر
وشعب الخفية «الجواسيس» المتعددة المحدثة التي مركزها
في المابين تحت نظارة السرخفية «رئيس الجواسيس»
فهذا الذي قضى بسقوط سعيد باشا بالحقيقة والواقع
فذهب بإصلاحاته ادراج الرياح ، وان كان عزله في الظاهر
بسبب احتلال البلغار للروم ايلي الشرقية ، واصراره على
إرسال العساكر كما تصرح بذلك معاهدة برلين ،

60% warped

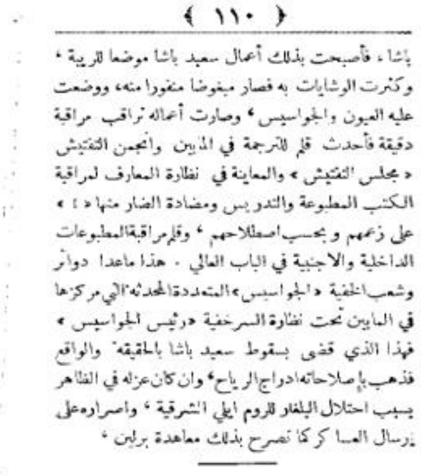


{ ١١٠ }

باشا ، فأصبحت بذلك أعمال سعيد باشا موضعا للريبة ،
وكثرت الوشايات به فصار مبغوضا منفورا منه، ووضعت
عليه العيون والجواسيس ، وصارت أعماله تراقب مراقبة
دقيقة فأحدث قلم للترجمة في المابين وانجمن التفتيش
«مجلس التفتيش» والمعاينة في نظارة المعارف لمراقبة
الكتب المطبوعة والتدريس ومضادة الضار منها «!»
على زعمهم و بحسب اصطلاحهم ، وقلم مراقبة المطبوعات
الداخلية والاجنبية في الباب العالي . هذا ماعدا دوائر
وشعب الخفية «الجواسيس» المتعددة المحدثة التي مركزها
في المابين تحت نظارة السرخفية «رئيس الجواسيس»
فهذا الذي قضى بسقوط سعيد باشا بالحقيقة والواقع
فذهب بإصلاحاته ادراج الرياح ، وان كان عزله في الظاهر
بسبب احتلال البلغار للروم ايلي الشرقية ، واصراره على
إرسال العساكر كما تصرح بذلك معاهدة برلين ،

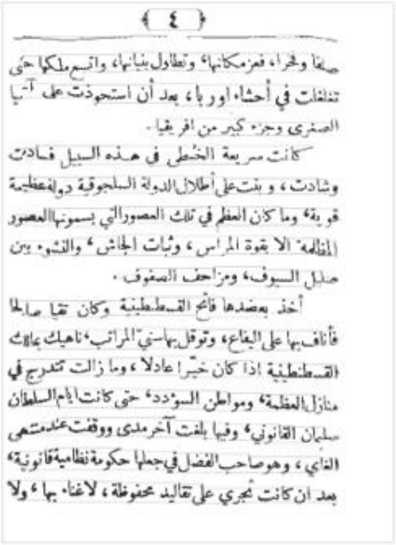
80% warped

Appendix B Result of images binarisation



{ ١١٠ }

باشا . فأصبحت بذلك أعمال سعيد باشا موضعا للريبة ،
وكثرت الوشايات به فصار مبغوضا منفورا منه، ووضعت
عليه العيون والجواسيس ، وصارت أعماله تراقب مراقبة
دقيقة فأحدث قلم للترجمة في المابين وانجمن التفتيش
« مجلس التفتيش » والمعاينة في نظارة المعارف لمراقبة
الكتب المطبوعة والتدريس ومضادة الضار منها « ! »
على زعمهم وبحسب اصطلاحهم ، وقلم مراقبة المطبوعات
الداخلية والاجنبية في الباب العالي . هذا ماعدا دوائر
وشعب الخفية « الجواسيس » المتعددة الحدثة التي مركزها
في المابين تحت نظارة السرخفية « رئيس الجواسيس »
فهذا الذي قضى بسقوط سعيد باشا بالحقيقة والواقع
فذهب باصلاحاته ادراج الرياح ، وان كان عزله في الظاهر
بسبب احتلال البلغار للروم ايلي الشرقية ، واصراره على
ارسال العساكر كما تصرح بذلك معاهدة برلين ،

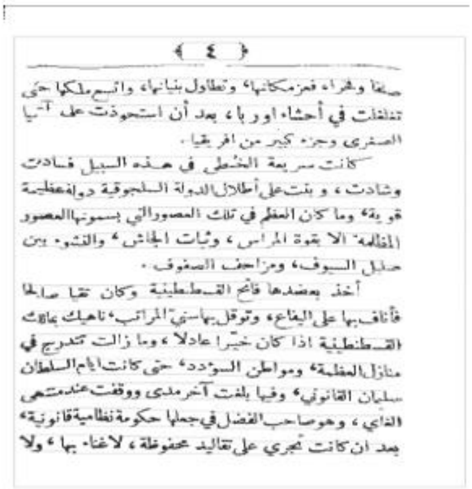


{ ٤ }

صيتا وفخرا ، فعز مكانها ، وتطاول بنيانها ، واتسع ملكها حتى
تغلغلت في أحشاء اوربا ، بعد أن استحوذت على آسيا
الصغرى وجزء كبير من افريقيا .

كانت سريعة الخطى في هذه السبيل فسادت
وشادت ، و بنت على أطلال الدولة السلجوقية دولة عظيمة
قوية ، وما كان العظم في تلك العصور التي يسمونها العصور
المظلمة الا بقوة المراس ، وثبات الجاش ، والنشوء بين
صليل السيوف ، ومزاحف الصفوف .

أخذ بعضدها فاتح القسطنطينية وكان تقيا صالحا
فأناف بها على اليفاع ، وتوقل بها سني المراتب ، ناهيك بملك
القسطنطينية اذا كان خيرا عادلا ، وما زالت تتدرج في
منازل العظمة ، ومواطن السودد ، حتى كانت أيام السلطان
سليمان القانوني ، وفيها بلغت آخر مدى ووقفت عند منتهى
الذاتي ، وهو صاحب الفضل في جعلها حكومة نظامية قانونية ،
بعد ان كانت تجري على تقاليد محفوظة ، لاغناء بها ، ولا

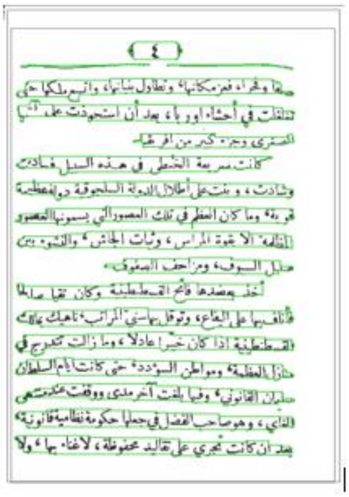


{ ٤ }

صيتا وفخرا ، فعز مكانها ، وتطاول بنيانها ، واتسع ملكها حتى
تغلغلت في أحشاء اوربا ، بعد أن استحوذت على آسيا
الصغرى وجزء كبير من افريقيا .

كانت سريعة الخطى في هذه السبيل فسادت
وشادت ، و بنت على أطلال الدولة السلجوقية دولة عظيمة
قوية ، وما كان العظم في تلك العصور التي يسمونها العصور
المظلمة الا بقوة المراس ، وثبات الجاش ، والنشوء بين
صليل السيوف ، ومزاحف الصفوف .

أخذ بعضدها فاتح القسطنطينية وكان تقيا صالحا
فأناف بها على اليفاع ، وتوقل بها سني المراتب ، ناهيك بملك
القسطنطينية اذا كان خيرا عادلا ، وما زالت تتدرج في
منازل العظمة ، ومواطن السودد ، حتى كانت أيام السلطان
سليمان القانوني ، وفيها بلغت آخر مدى ووقفت عند منتهى
الذاتي ، وهو صاحب الفضل في جعلها حكومة نظامية قانونية ،
بعد ان كانت تجري على تقاليد محفوظة ، لاغناء بها ، ولا

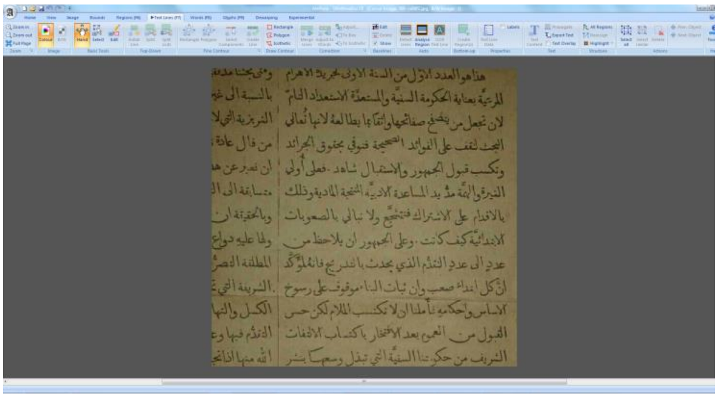


{ ٤ }

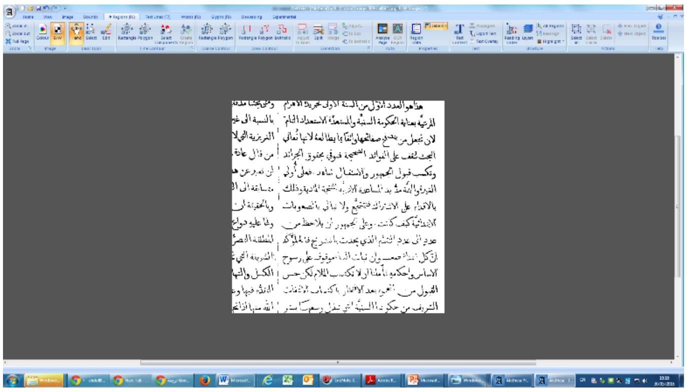
صيتا وفخرا ، فعز مكانها ، وتطاول بنيانها ، واتسع ملكها حتى
تغلغلت في أحشاء اوربا ، بعد أن استحوذت على آسيا
الصغرى وجزء كبير من افريقيا .

كانت سريعة الخطى في هذه السبيل فسادت
وشادت ، و بنت على أطلال الدولة السلجوقية دولة عظيمة
قوية ، وما كان العظم في تلك العصور التي يسمونها العصور
المظلمة الا بقوة المراس ، وثبات الجاش ، والنشوء بين
صليل السيوف ، ومزاحف الصفوف .

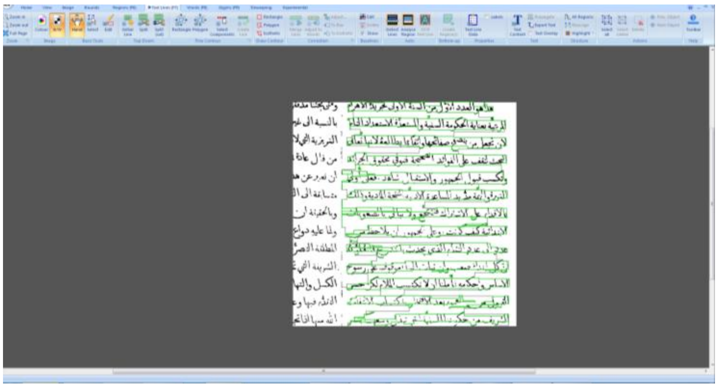
أخذ بعضدها فاتح القسطنطينية وكان تقيا صالحا
فأناف بها على اليفاع ، وتوقل بها سني المراتب ، ناهيك بملك
القسطنطينية اذا كان خيرا عادلا ، وما زالت تتدرج في
منازل العظمة ، ومواطن السودد ، حتى كانت أيام السلطان
سليمان القانوني ، وفيها بلغت آخر مدى ووقفت عند منتهى
الذاتي ، وهو صاحب الفضل في جعلها حكومة نظامية قانونية ،
بعد ان كانت تجري على تقاليد محفوظة ، لاغناء بها ، ولا



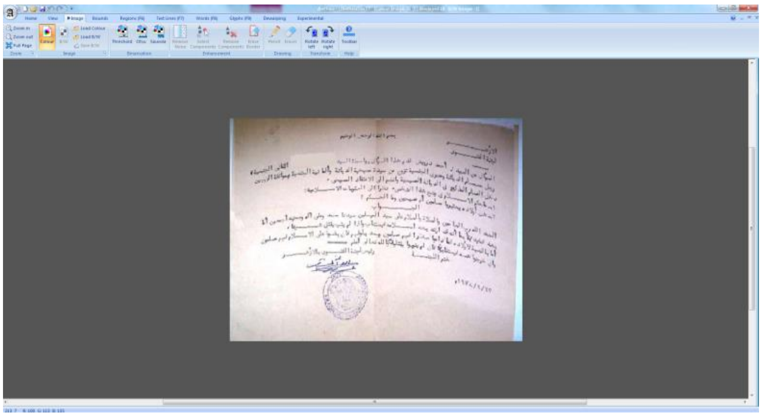
Aletheia application screen snapshots



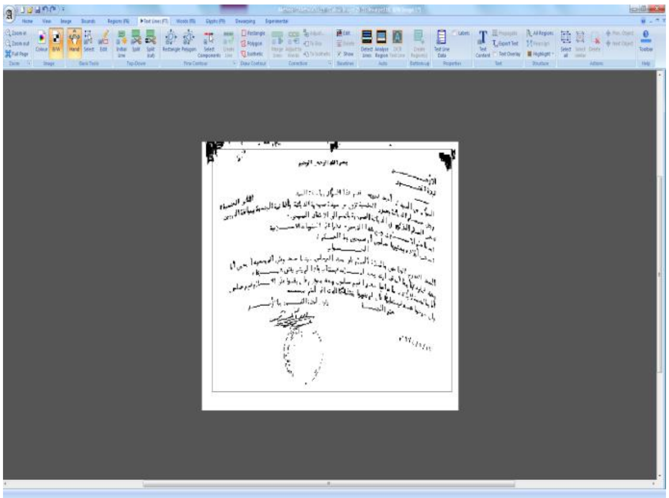
Binarise the colour image using an adaptive method Sauvola algorithm



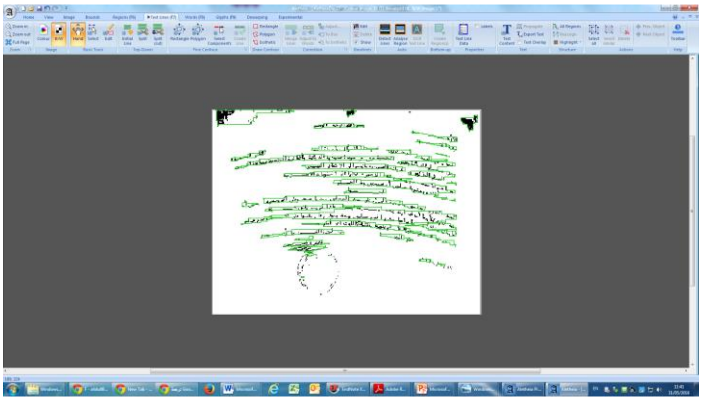
Result of detected text lines



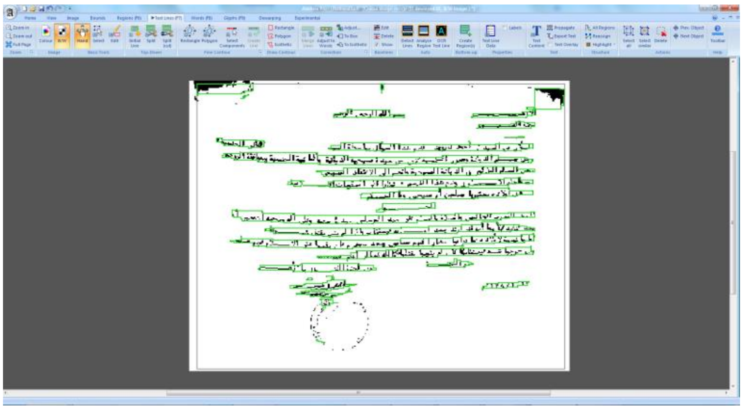
Loading warped image



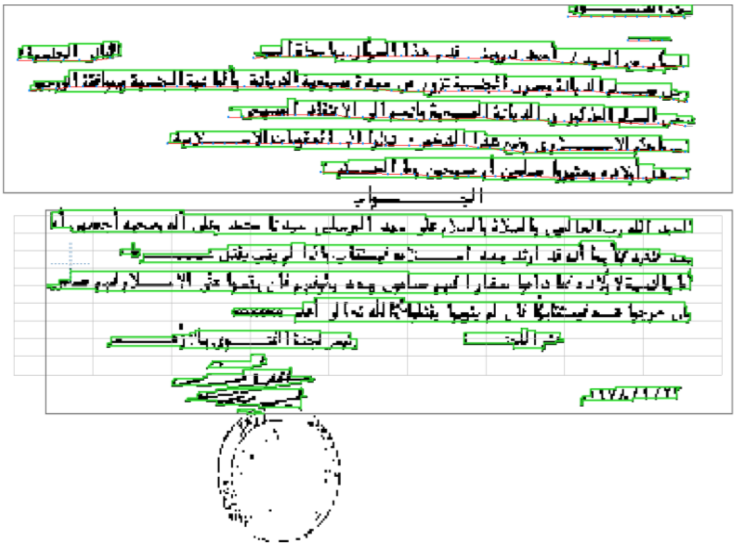
Binarise the colour image using an adaptive method Sauvola algorithm



Result of detected text lines

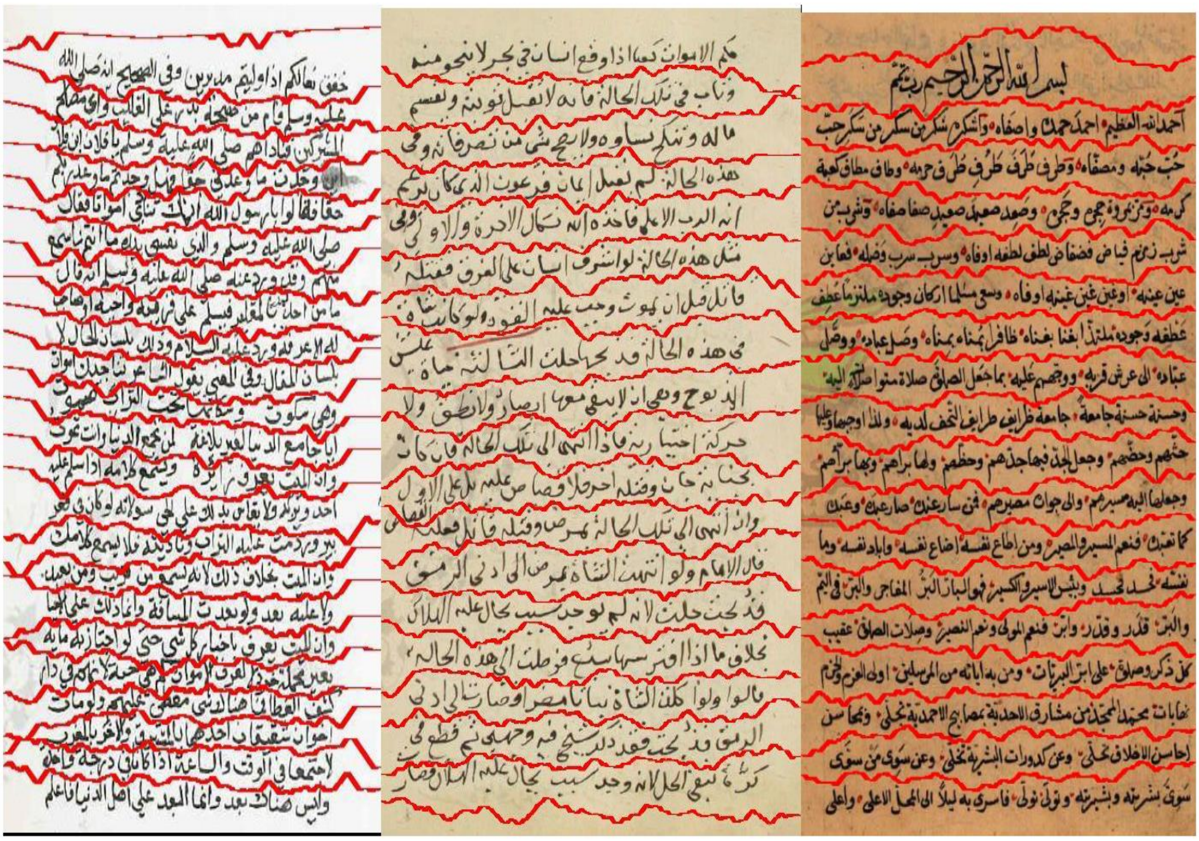


Result of detected text lines in the normal image

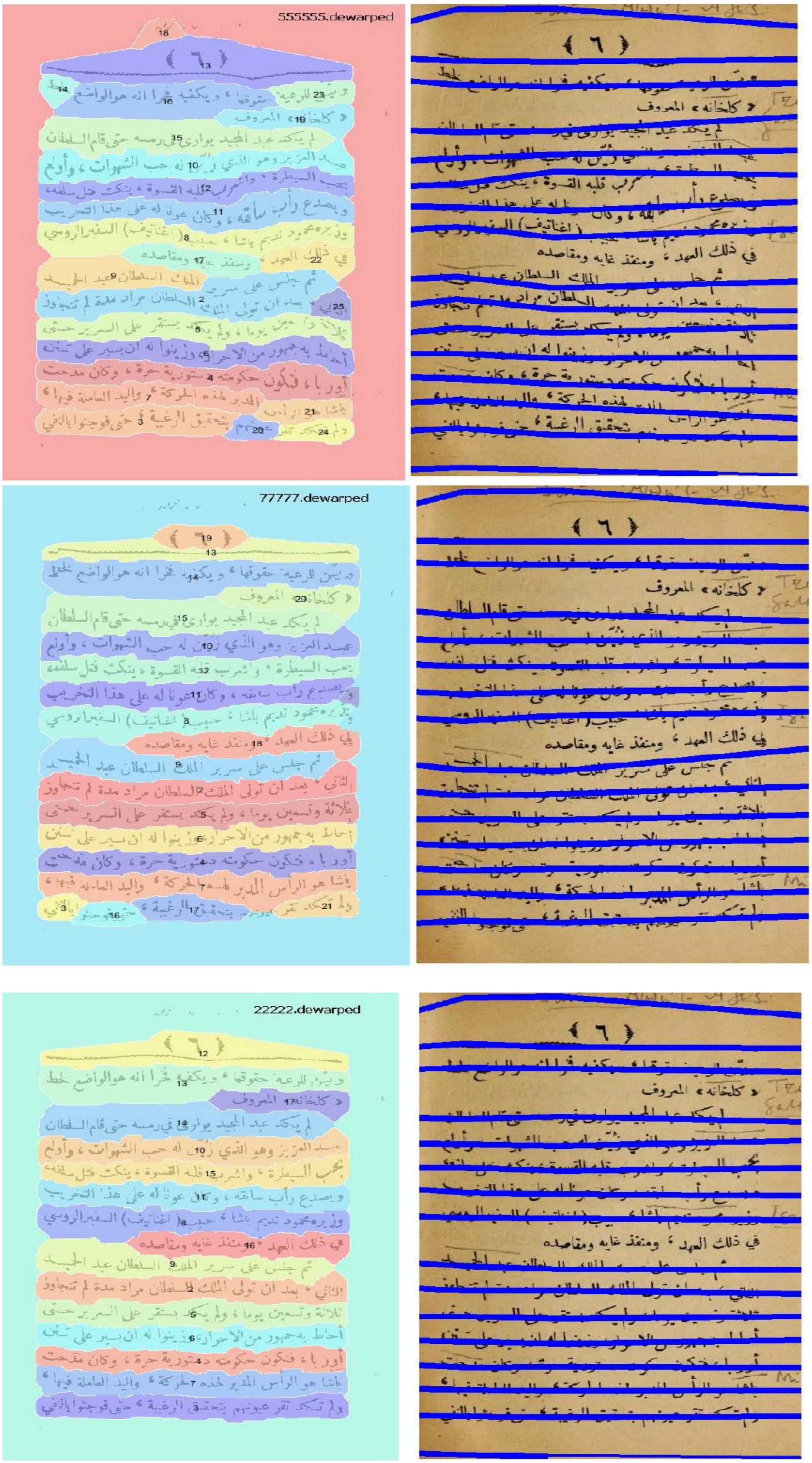


Result of detected text lines in the normal image

Appendix C Result of text line segmentation

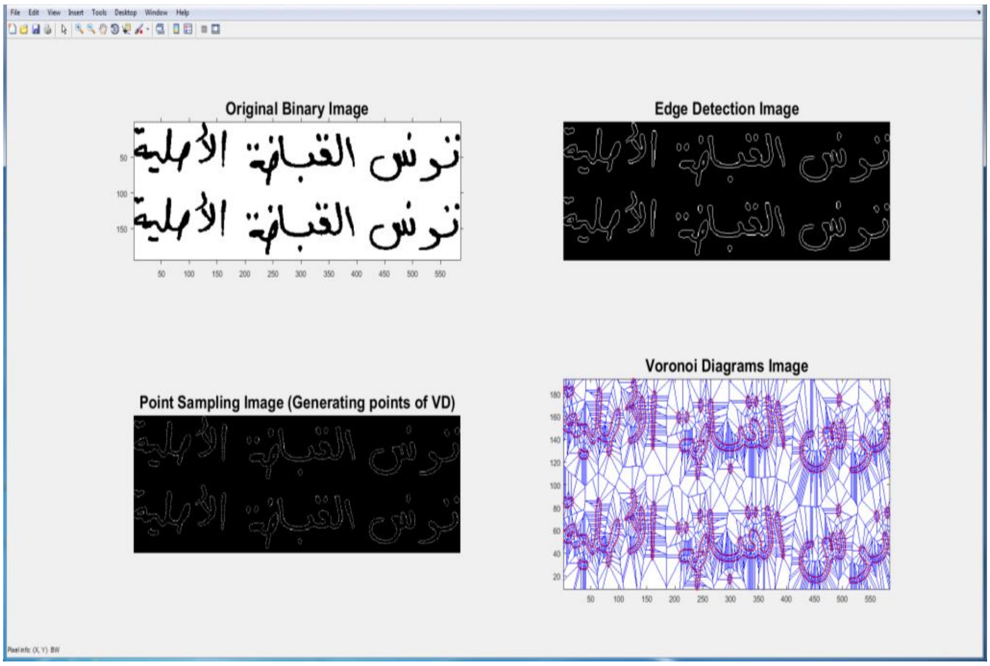


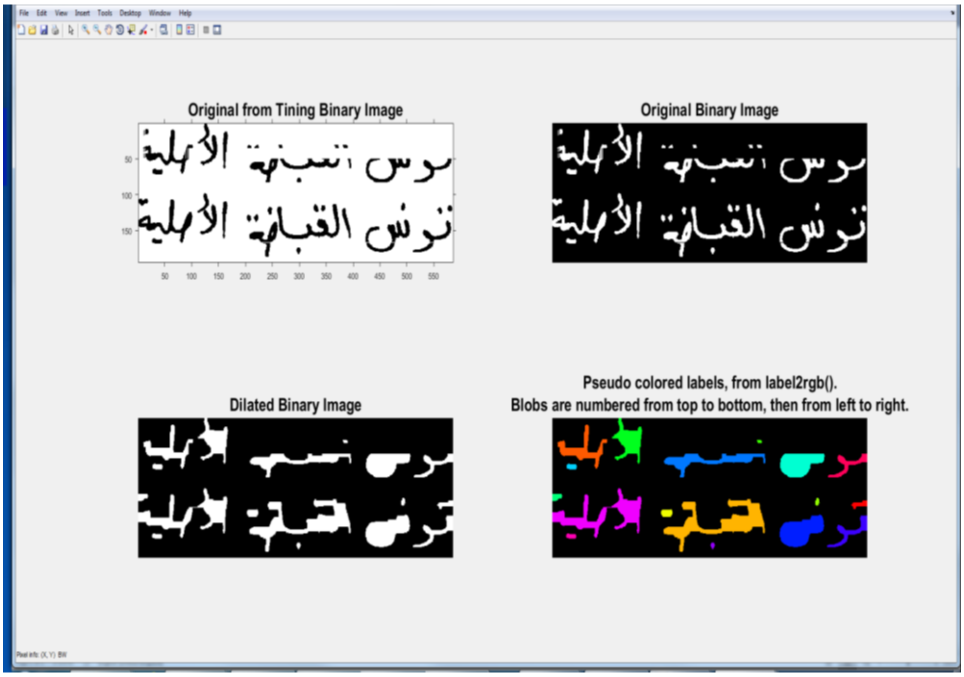
Result of Seam carving for text line extraction on different warp samples



Result of watershed transform and Projection-based method on different warp samples

Appendix E Results of baseline estimation process stages


File Edit View Insert Tools Desktop Window Help
Original Binary Image
Edge Detection Image
Point Sampling Image (Generating points of VD)
Voronoi Diagrams Image
Pixel info: (X, Y) BW


File Edit View Insert Tools Desktop Window Help
Original from Tining Binary Image
Original Binary Image
Dilated Binary Image
Pseudo colored labels, from label2rgb().
Blobs are numbered from top to bottom, then from left to right.
Pixel info: (X, Y) BW

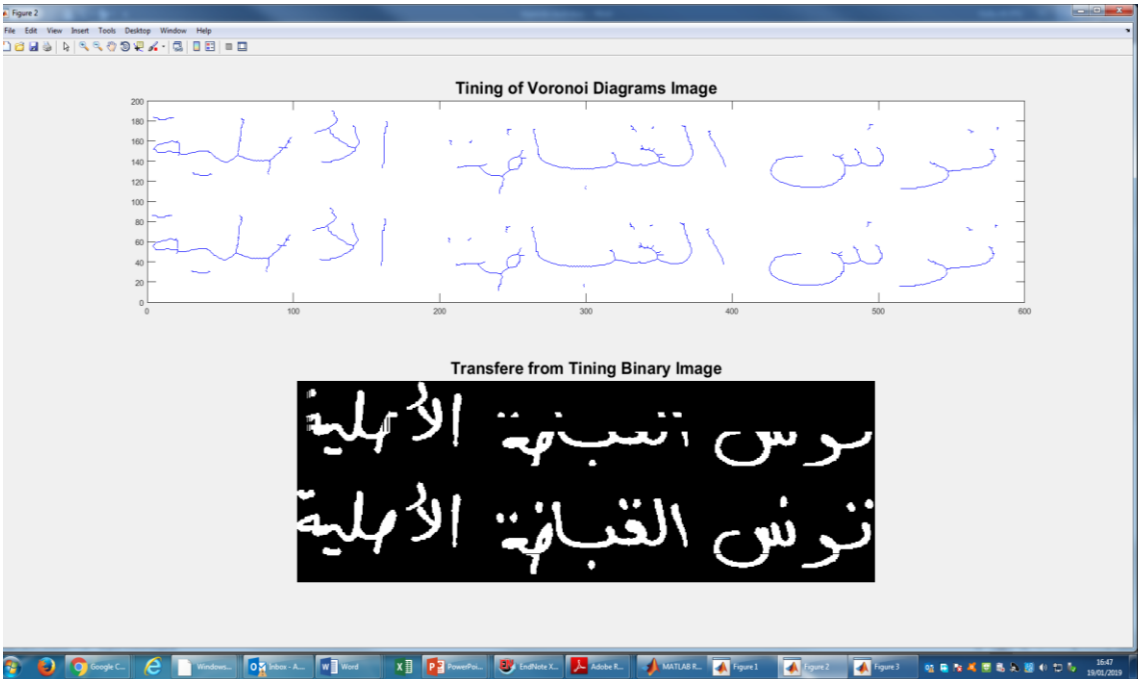
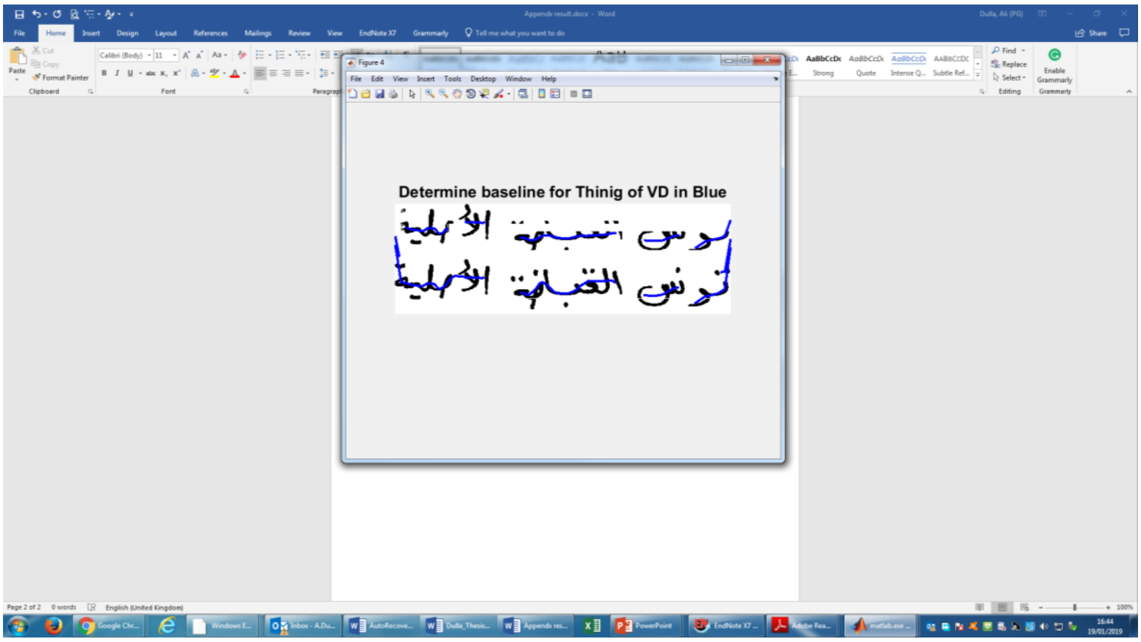


Figure 2
File Edit View Insert Tools Desktop Window Help
Tining of Voronoi Diagrams Image
0
20
40
60
80
100
120
140
160
180
200
100
200
300
400
500
600
Transfere from Tining Binary Image


Appendx result.docx - Word
File Home Insert Design Layout References Mailings Review View EndNote X7 Grammarly
Figure 4
File Edit View Insert Tools Desktop Window Help
Determine baseline for Thinig of VD in Blue
Page 2 of 2 0 words English (United Kingdom)

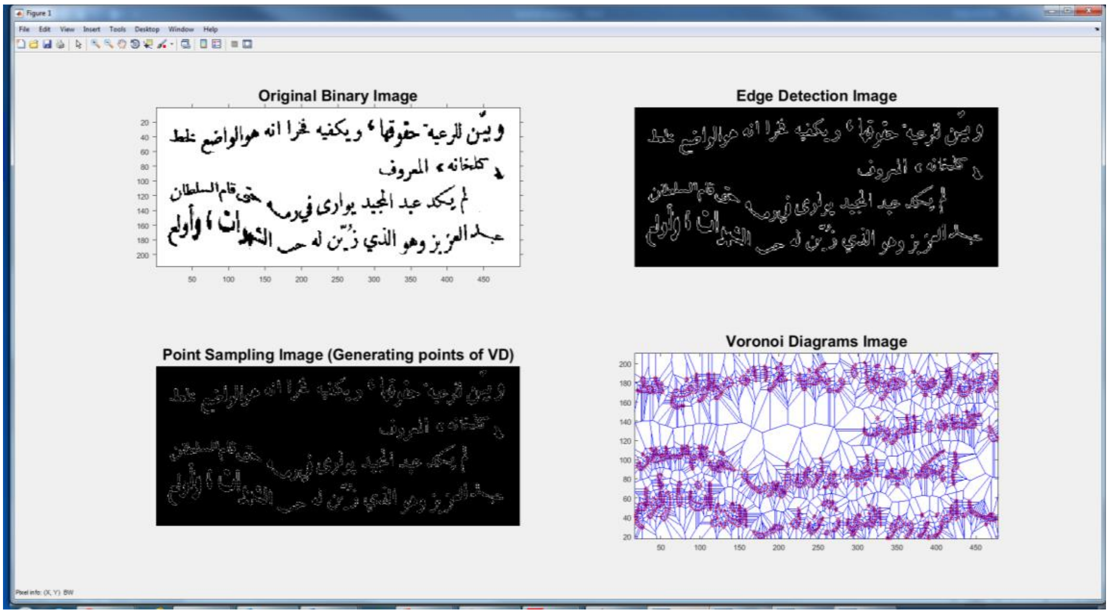


Figure 1
File Edit View Insert Tools Desktop Window Help
Original Binary Image
Edge Detection Image
Point Sampling Image (Generating points of VD)
Voronoi Diagrams Image
20 40 60 80 100 120 140 160 180 200
50 100 150 200 250 300 350 400 450
Pixel info: (X, Y) BW

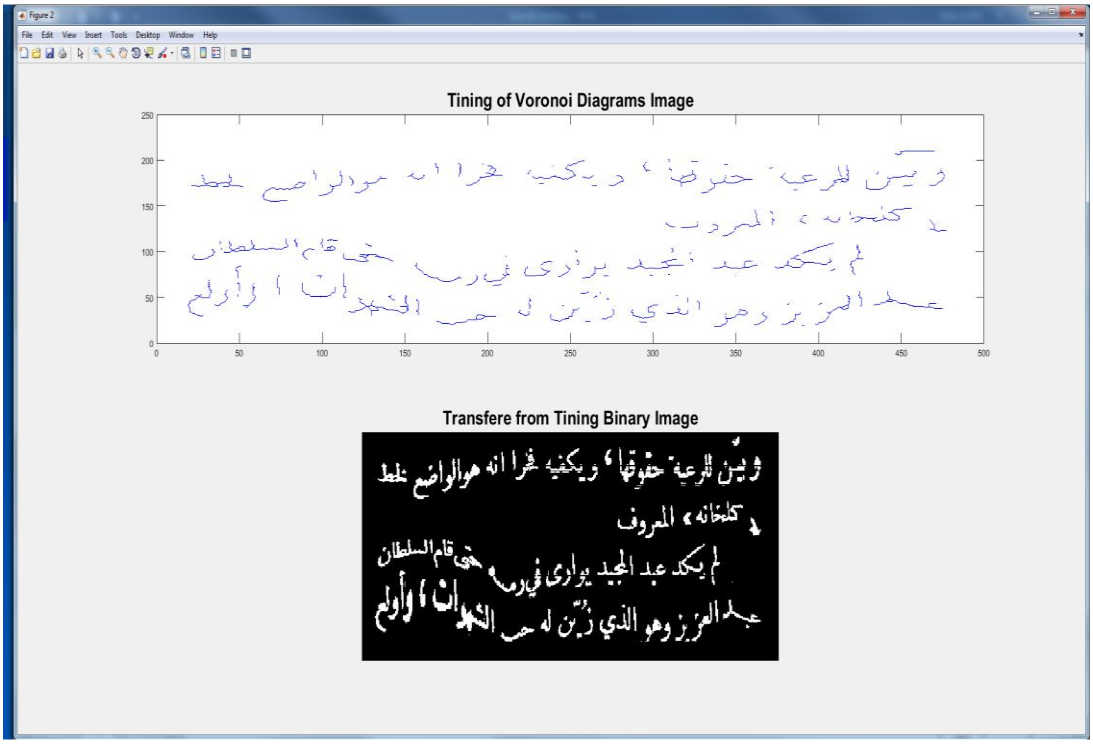


Figure 2
File Edit View Insert Tools Desktop Window Help
Tining of Voronoi Diagrams Image
0 50 100 150 200 250
0 50 100 150 200 250 300 350 400 450 500
Transfere from Tining Binary Image

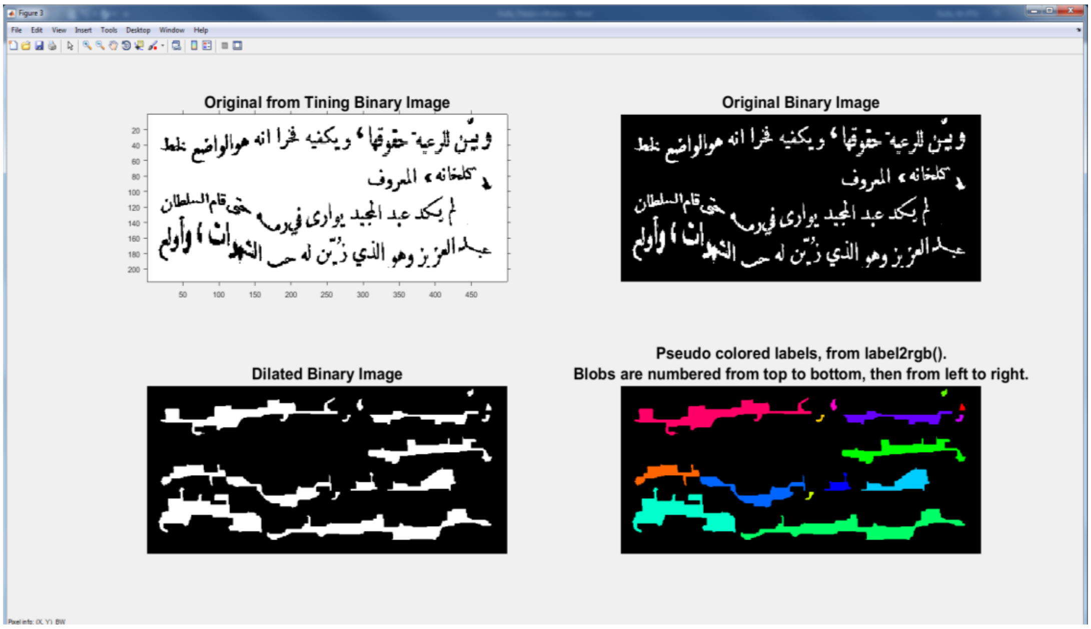


Figure 3
File Edit View Insert Tools Desktop Window Help
Original from Tining Binary Image
Original Binary Image
Dilated Binary Image
Pseudo colored labels, from label2rgb().
Blobs are numbered from top to bottom, then from left to right.
Pixel info: (X, Y) BW

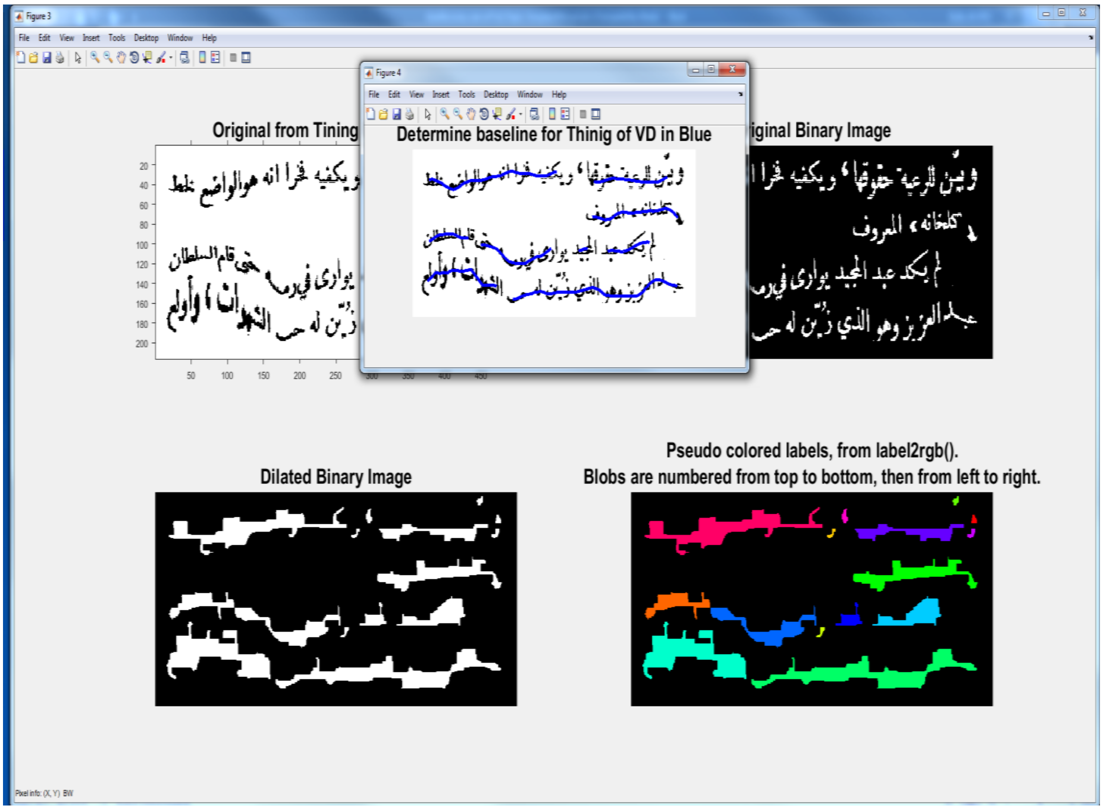
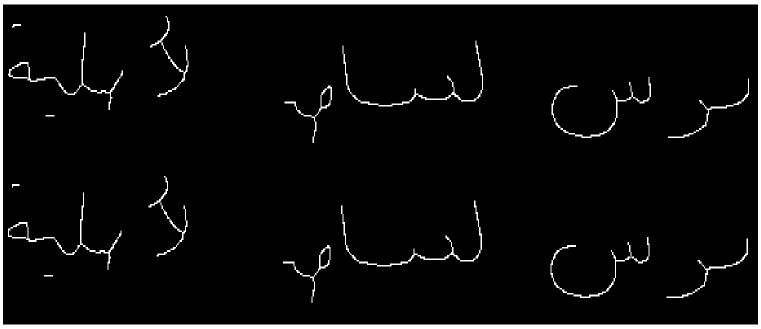


Figure 3
File Edit View Insert Tools Desktop Window Help
Figure 4
Determine baseline for Thinig of VD in Blue
Original from Tining
iginal Binary Image
Dilated Binary Image
Pseudo colored labels, from label2rgb().
Blobs are numbered from top to bottom, then from left to right.
Pixel info: (X, Y) BW

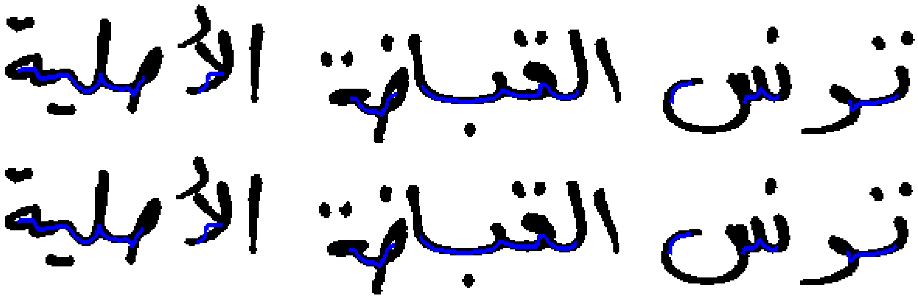
Binarized Image

Left and Right of Image Cropped

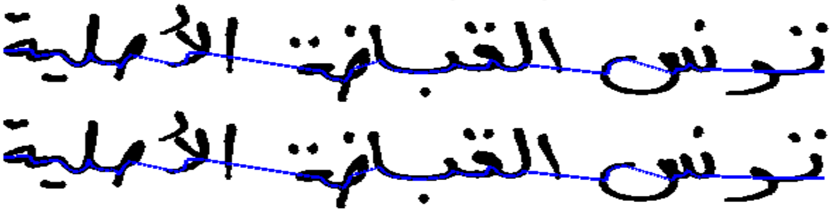
Filtered Skeleton Image



Detected Baseline



Baseline Over Original Image



Appendix F MATLAB coding

Source code

```
Clear
close all
warning ('off','all');
clc
filename='2.tif';
I=imread(filename);
%I=rgb2gray(I);
BW=im2bw(I);
th=5; %size of the structuring element for the dilation
BW = padarray(BW,[th th],1,'both'); % padding the image with white pixels on both
sides
BW=~bwareaopen(~BW,5); % Remove components having area less than 5 pixels)
imshow(BW); title('Original binrized Image') % displaying the image
y1=fix(size(BW,1)/3);
x1=fix(size(BW,2)/3);
portion=BW(y1:y1*2,x1:x1*2); % cropping the middel portion of the image
SW=LenghtWidthCC(portion); % estimating the strok width from this portion
SizeOfDiacratics=(SW*4)*3; % estimation of the size of diacratics
BW=~bwareaopen(~BW,SizeOfDiacratics); %Removing diacratical points
NewLabel=bwlabeln(~BW,8); %find connected components
newImg=ones(size(BW)); %new white image which will hold the dialated components
Boxs = regionprops(NewLabel,
'BoundingBox','Orientation','ConvexImage','Image');%Connected components
k=1;
for i=1:length(Boxs)
   widthOfComponent(k)=Boxs(i).BoundingBox(3); % storing the width of each
components
   k=k+1;
   se = strel('line',th,Boxs(i).Orientation); % a line structuring element with 5pixels width
and orientation of the current component
   BW2 = padarray(Boxs(i).ConvexImage,[th th],0,'both'); % padding the convex image
in both sides
   BW2 = ~imdilate(BW2,se); % application of morphological dialation to the convex
image
   x1=fix(Boxs(i).BoundingBox(1))-th;
   y1=fix(Boxs(i).BoundingBox(2))-th;
   x2=fix(x1+Boxs(i).BoundingBox(3))-1+2*th;
   y2=fix(y1+Boxs(i).BoundingBox(4))-1+2*th;
   newImg(y1:y2,x1:x2)=newImg(y1:y2,x1:x2)&BW2; %pasting the dialated
component to the new image
end
%estimation of the space between words as the mean of the width of all
```

```
%components multiplied by 2

widthOfComponent=fix(mean(widthOfComponent)*2);
NewLabel=bwlabeln(~newImg,8);%find connected components of the new image
Boxs = regionprops(NewLabel,
'BoundingBox','Orientation','ConvexImage','Image','Centroid');%Connected component
figure,
imshow(newImg); title('Image that contains filled convex areas of all components ')
hold on,
k=1;
for i=1: length(Boxs)
    %for each component search for the neighboring components in its left
    %and right base of the function of field of view (VisionOfCC)
    %linkThem is a matrix of two columns the first one is the current
    %component and second one is the neighboring component (left or right)
    ListCCInVision = VisionOfCC(NewLabel,Boxs(i),widthOfComponent,'LEFT');
    if length (ListCCInVision)>1
        linkThem(k,1)=i;
        linkThem(k,2)=ListCCInVision(2);
        k=k+1;
    end
    ListCCInVision = VisionOfCC(NewLabel,Boxs(i),widthOfComponent,'RIGHT');
    if length (ListCCInVision)>1
            linkThem(k,1)=i;
        linkThem(k,2)=ListCCInVision(2);
            k=k+1;
    end
end

%___________________________________________recursive function igure,
imshow(BW);  title('Original Image after removing small components')
hold on,
global ListNoeud; % global vaiabl that hold list of component of current line
ListNoeud=[];
AllTheNodes=[];
k=1;
LinkedImage=BW;% a new image that will contain text + lower contours + links between
lower contours
for i=1: length(linkThem)
        % if the current couple (component and its neighbor) are not
        % allready visited,
        if nnz(find(AllTheNodes==linkThem(i,1)))==0 &&
nnz(find(AllTheNodes==linkThem(i,2)))==0
            getAllLinks(linkThem(i,1),1,linkThem);
            getAllLinks(linkThem(i,1),2,linkThem);
            getAllLinks(linkThem(i,2),1,linkThem);
            getAllLinks(linkThem(i,2),2,linkThem);
            ListN=unique(ListNoeud);

            AllTheNodes=[AllTheNodes,ListN];
```

```
      % extract coordinate of lower contour of all components in the
      % current line
      [x,y]=NextDiacraticsWithLower(ListN,Boxs,BW);
      %plot(x,y,'b-','LineWidth',2);

      % smoothing the curve (lower contour) to estimate the base line
      Yhat = baselinetest(x,y,2);
      plot(x,Yhat,'m-','LineWidth',2);

      Baselines=ones(size(BW));% an image that contains only the lower contour
      % for all pixels in lower contour fill pixels in the two image
      % LinkedImage and Baselines
      for j=1: length(x)
        LinkedImage(fix (Yhat(j)),x(j))=0;
        Baselines(fix (Yhat(j)),x(j))=0;
      end

      %=========== Connect baselines ==================
      NewLabel2=bwlabeln(~Baselines,8);%find connected components of the image containing only lower contour
      Boxs2 = regionprops(NewLabel2, 'PixelList');%Connected component
      for j=1:length(Boxs2)-1
        X0=fix(Boxs2(j).PixelList(size(Boxs2(j).PixelList,1),2));
        Y0=fix(Boxs2(j).PixelList(size(Boxs2(j).PixelList,1),1));
        X1=fix(Boxs2(j+1).PixelList(1,2));
        Y1=fix(Boxs2(j+1).PixelList(1,1));
        % draw a line between connected components
        LinkedImage = func_DrawLine(LinkedImage, X0, Y0, X1, Y1, 0);
      end
      %=============================
      ListNoeud=[];
      k=k+1;
    end
end

figure,
imshow(LinkedImage), title('LinkedImage')
th=100; % Threshold to remove remaining small components
LinkedImage=~bwareaopen(~LinkedImage,th); % Remove components having area less than 5 pixels)
figure,
imshow(LinkedImage), title('LinkedImage after removing small components')

%Apply watershed transform on the linked image in order to extract line
%regions
L = watershed(LinkedImage);
Lrgb = label2rgb(L, 'jet', 'w', 'shuffle');
figure
imshow(L)%I
```

```
hold on
imshow(Lrgb);title('Result of Watershed Transform on the LinkedImage')
alpha(0.3)
figure
imshow(L,[])
hold on,
XML_content = makeXML(L,filename);
if exist(strcat(filename,'.xml'), 'file')==2
  delete(strcat(filename,'.xml'));
end
fid=fopen(strcat(filename,'.xml'),'w');
fprintf(fid,XML_content);
fclose(fid);
fid=fopen(strcat(filename,'.xml'));
fclose(fid);
close all
clear
clc
```

References

IMPACT: Improving Access to Text [Online]. European Union 7th Framework.

ABU-AIN, T., ABDULLAH, S. N. H. S., BATAINEH, B., OMAR, K. & ABU-EIN, A. 2013a. A novel baseline detection method of handwritten Arabic-script documents based on sub-words. *Soft Computing Applications and Intelligent Systems.* Springer.

ABU-AIN, W., ABDULLAH, S. N. H. S., BATAINEH, B., ABU-AIN, T. & OMAR, K. 2013b. Skeletonization algorithm for binary images. *Procedia Technology,* 11**,** 704-709.

ABUHAIBA, I. S. 2003. A discrete Arabic script for better automatic document understanding.

ABULHAB, S. D. 2007. Roots of Modern Arabic Script: From Musnad to Jazm.

AL-SHATNAWI, A. 2010. *A Non-Iterative Thinning Method Based on Exploited Vertices of Voronoi Diagrams.* PhD Thesis

AL-SHATNAWI, A. 2016. A Novel Baseline Estimation Method for Arabic Handwritten Text Based on Exploited Components of Voronoi Diagrams. *International Arab Journal of Information Technology (IAJIT),* 13.

AL-SHATNAWI, A. & OMAR, K. Detecting arabic handwritten word baseline using voronoi diagram. Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on, 2009. IEEE, 18-22.

AL-SHATNAWI, A. M., ALFAWWAZ, B. M., OMAR, K. & ZEKI, A. M. Skeleton extraction: Comparison of five methods on the arabic ifn/enit database. 2014 6th International Conference on Computer Science and Information Technology (CSIT), 2014. IEEE, 50-59.

AL AGHBARI, Z. & BROOK, S. 2009. HAH manuscripts: A holistic paradigm for classifying and retrieving historical Arabic handwritten documents. *Expert Systems with Applications,* 36**,** 10942-10951.

ALAEI, A., NAGABHUSHAN, P. & PAL, U. 2011. Piece-wise painting technique for line segmentation of unconstrained handwritten text: a specific study with Persian text documents. *Pattern Analysis and Applications,* 14**,** 381-394.

ALGINAHI, Y. M. 2013. A survey on Arabic character segmentation. *International Journal on Document Analysis and Recognition (IJDAR),* 16**,** 105-126.

ALSHAHRANI, A. A. 2008. Arabic Script and the Rise of Arabic Calligraphy. *Online Submission*.

ANTONACOPOULOS, D. A. 1995. Introduction to Document Image Analysis.

AOUADI, N., AMIRI, S. & ECHI, A. K. 2013. Segmentation of Connected Components in Arabic Handwritten Documents. *Procedia Technology,* 10**,** 738-746.

AOUADI, N. & KACEM, A. 2016. A proposal for touching component segmentation in Arabic manuscripts. *Pattern Analysis and Applications***,** 1-23.

ARVANITOPOULOS, N. & SÜSSTRUNK, S. Seam carving for text line extraction on color and grayscale historical manuscripts. Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on, 2014. IEEE, 726-731.

ATALLAH, A.-S. & OMAR, K. 2008. Methods of arabic language baseline detection–The state of art. *IJCSNS,* 8**,** 137.

ATALLAH, A.-S. & OMAR, K. A comparative study between methods of arabic baseline detection. Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on, 2009. IEEE, 73-77.

AZMI, A. & ALSAIARI, A. 2010. Arabic typography: a survey. *International Journal of Electrical & Computer Sciences,* 9**,** 1.

BAIRD, H. S., BUNKE, H. & YAMAMOTO, K. 2012. *Structured document image analysis,* Springer Science & Business Media.

BAUMANN, R., BLACKWELL, C. & SEALES, W. B. 2012. Automatic Perspective Correction of Manuscript Images. *The Outreach of Digital Libraries: A Globalized Resource Network.* Springer.

BELAÏD, A. & OUWAYED, N. 2012. Segmentation of ancient Arabic documents. *Guide to OCR for Arabic Scripts.* Springer.

BERMAN, P., KAHNG, A. B., VIDHANI, D., WANG, H. & ZELIKOVSKY, A. Optimal phase conflict removal for layout of dark field alternating phase shifting masks. Proceedings of the 1999 international symposium on Physical design, 1999. ACM, 121-126.

BOOKSTEIN, F. L. 1997. Morphometric Tools for Landmark Data. *Morphometric Tools for Landmark Data, by Fred L. Bookstein, pp. 455. ISBN 0521585988. Cambridge, UK: Cambridge University Press, June 1997.***,** 455.

BOUKERMA, H. & FARAH, N. A novel Arabic baseline estimation algorithm based on sub-words treatment. Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on, 2010. IEEE, 335-338.

BOULID, Y., SOUHAR, A., AMEUR, E. & OUAGAGUE, M. M. 2017. Watershed transform for text lines extraction on binary handwrite documents.

BROWN, M. S. & PISULA, C. J. Conformal deskewing of non-planar documents. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005. IEEE, 998-1004.

BROWN, M. S. & SEALES, W. B. Document restoration using 3D shape: a general deskewing algorithm for arbitrarily warped documents. Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, 2001. IEEE, 367-374.

BROWN, M. S. & SEALES, W. B. 2004. Image restoration of arbitrarily warped documents. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 26**,** 1295-1306.

BROWN, M. S., SUN, M., YANG, R., YUN, L. & SEALES, W. B. 2007. Restoring 2D content from distorted documents. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 29**,** 1904-1916.

BROWN, M. S. & TSOI, Y.-C. 2006. Geometric and shading correction for images of printed materials using boundary. *Image Processing, IEEE Transactions on,* 15**,** 1544-1554.

BUKHARI, S. S., SHAFAIT, F. & BREUEL, T. M. Segmentation of curled textlines using active contours. Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on, 2008. IEEE, 270-277.

BUKHARI, S. S., SHAFAIT, F. & BREUEL, T. M. Dewarping of document images using coupled-snakes. Proceedings of Third International Workshop on Camera-Based Document Analysis and Recognition, Barcelona, Spain, 2009. Citeseer, 34-41.

BUKHARI, S. S., SHAFAIT, F. & BREUEL, T. M. 2013. Coupled snakelets for curled text-line segmentation from warped document images. *International Journal on Document Analysis and Recognition (IJDAR),* 16**,** 33-53.

BURROW, P. 2004. Arabic handwriting recognition. *Report of Master of Science School of Informatics, University of Edinburgh*.

CANNY, J. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence***,** 679-698.

CHUA, K. B., ZHANG, L., ZHANG, Y. & TAN, C. L. A fast and stable approach for restoration of warped document images. Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on, 2005. IEEE, 384-388.

CLAUSNER, C., ANTONACOPOULOS, A. & PLETSCHACHER, S. A robust hybrid approach for text line segmentation in historical documents. Pattern Recognition (ICPR), 2012 21st International Conference on, 2012. IEEE, 335-338.

CLAUSNER, C., PLETSCHACHER, S. & ANTONACOPOULOS, A. Aletheia-an advanced document layout and text ground-truthing system for production environments. Document Analysis and Recognition (ICDAR), 2011 International Conference on, 2011a. IEEE, 48-52.

CLAUSNER, C., PLETSCHACHER, S. & ANTONACOPOULOS, A. Scenario driven in-depth performance evaluation of document layout analysis methods. Document Analysis and Recognition (ICDAR), 2011 International Conference on, 2011b. IEEE, 1404-1408.

COHEN, R., KEDEM, K., DINSTEIN, I. & EL-SANA, J. Occluded character restoration using active contour with shape priors. Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on, 2012. IEEE, 497-502.

DALDALI, M. & SOUHAR, A. 2018. Handwritten Arabic Documents Segmentation into Text Lines using Seam

Carving. *International Journal of Interactive Multimedia and Artificial Intelligence*.

DAS, S., MISHRA, G., SUDHARSHANA, A. & SHILKROT, R. The Common Fold: Utilizing the Four-Fold to Dewarp Printed Documents from a Single Image. Proceedings of the 2017 ACM Symposium on Document Engineering, 2017. ACM, 125-128.

DULLA, A. 2018. A dataset of Warped Historical Arabic Documents.

DUMMER, J. 2004. A simple time-corrected verlet integration method. *Game Developer*.

EPPSTEIN, D. 1992. The farthest point Delaunay triangulation minimizes angles. *Computational Geometry,* 1**,** 143-148.

ESTEBAN-LANSAQUE, A., SÁNCHEZ, C., BORRÀS, A., DIEZ-FERRER, M., ROSELL, A. & GIL, D. Stable Anatomical Structure Tracking for Video-Bronchoscopy Navigation. Workshop on Clinical Image-Based Procedures, 2016. Springer, 18-26.

FAROOQ, F., GOVINDARAJU, V. & PERRONE, M. Pre-processing methods for handwritten Arabic documents. Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on, 2005. IEEE, 267-271.

FELIX, A. Y., JESUDOSS, A. & MAYAN, J. A. Entry and exit monitoring using license plate recognition. Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2017 IEEE International Conference on, 2017. IEEE, 227-231.

GACEK, A. 2009. *Arabic manuscripts: a vademecum for readers*, Brill.

GATOS, B., PRATIKAKIS, I. & NTIROGIANNIS, K. Segmentation based recovery of arbitrarily warped document images. Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, 2007. IEEE, 989-993.

GATOS, B., PRATIKAKIS, I. & PERANTONIS, S. J. 2006. Adaptive degraded document image binarization. *Pattern Recognition,* 39**,** 317-327.

GOLD, C. & SNOEYINK, J. 2001. A one-step crust and skeleton extraction algorithm. *Algorithmica,* 30**,** 144-163.

GONZALEZ, R. C. 2016. Digital image processing. Prentice hall.

HARTLEY, R. & ZISSERMAN, A. 2003. *Multiple view geometry in computer vision,* Cambridge university press.

HOUGH, P. V. 1962. Method and means for recognizing complex patterns.

IMPACT. 2012 *IMPACT: Improving Access to Text, European Union 7th Framework* [Online]. Available: http://www.impact-project.eu/ [Accessed 29/10/2018.

JÄHNE, B. 2002. Texture. *Digital Image Processing.* Springer.

JAMAL, A. T. 2015. *End-Shape Analysis for Automatic Segmentation of Arabic Handwritten Texts.* Concordia University Montreal, Quebec, Canada.

KANUNGO, T., MARTON, G. A. & BULBUL, O. OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products. Electronic Imaging'99, 1999. International Society for Optics and Photonics, 109-120.

KASSIS, M., ABDALHALEEM, A., DROBY, A., ALAASAM, R. & EL-SANA, J. 2017. VML-HD: The Historical Arabic Documents Dataset for Recognition Systems.

KEFALI, A., SARI, T. & SELLAMI, M. Evaluation of several binarization techniques for old Arabic documents images. The First International Symposium on Modeling and Implementing Complex Systems MISC, 2010. 88-99.

KHAYYAT, M., LAM, L., SUEN, C. Y., YIN, F. & LIU, C.-L. Arabic handwritten text line extraction by applying an adaptive mask to morphological dilation. Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, 2012. IEEE, 100-104.

KIM, B. S., KOO, H. I. & CHO, N. I. 2015. Document dewarping via text-line based optimization. *Pattern Recognition,* 48**,** 3600-3614.

KOVESI, P. 1999. Phase preserving denoising of images. *signal,* 4**,** 1.

LAVIALLE, O., MOLINES, X., ANGELLA, F. & BAYLOU, P. Active contours network to straighten distorted text lines. Image Processing, 2001. Proceedings. 2001 International Conference on, 2001. IEEE, 748-751.

LEEDHAM, G., YAN, C., TAKRU, K., TAN, J. H. N. & MIAN, L. Comparison of some thresholding algorithms for text/background segmentation in difficult document images. null, 2003. IEEE, 859.

LI, Q. & XIE, Y. 2003. Randomised hough transform with error propagation for line and circle detection. *Pattern Analysis & Applications,* 6**,** 55-64.

LI, Z. 2008. *A unified framework for document image restoration.*

LIKFORMAN-SULEM, L., HANIMYAN, A. & FAURE, C. A Hough based algorithm for extracting text lines in handwritten documents. Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, 1995. IEEE, 774-777.

LIKFORMAN-SULEM, L., ZAHOUR, A. & TACONET, B. 2007. Text line segmentation of historical documents: a survey. *International Journal of Document Analysis and Recognition (IJDAR),* 9**,** 123-138.

LIU, C., ZHANG, Y., WANG, B. & DING, X. 2015. Restoring camera-captured distorted document images. *International Journal on Document Analysis and Recognition (IJDAR),* 18**,** 111-124.

LOULOUDIS, G., GATOS, B., PRATIKAKIS, I. & HALATSIS, C. 2008. Text line detection in handwritten documents. *Pattern Recognition,* 41**,** 3758-3772.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision,* 60**,** 91-110.

LU, S., CHEN, B. M. & KO, C. C. 2005. Perspective rectification of document images using fuzzy set and morphological operations. *Image and Vision Computing,* 23**,** 541-553.

LU, S., CHEN, B. M. & KO, C. C. 2006. A partition approach for the restoration of camera images of planar and curled document. *Image and Vision Computing,* 24**,** 837-848.

LU, S. & TAN, C. L. Document flattening through grid modeling and regularization. Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, 2006a. IEEE, 971-974.

LU, S. & TAN, C. L. 2006b. The restoration of camera documents through image segmentation. *Document Analysis Systems VII.* Springer.

LUND, W. B. 2014. Ensemble Methods for Historical Machine-Printed Document Recognition.

LUONG, Q.-T. & FAUGERAS, O. 2001. The geometry of multiple images. *MIT Press, Boston,* 2**,** 4.5.

MADDOURI, S. S., SAMOUD, F. B., BOURIEL, K., ELLOUZE, N. & EL ABED, H. Baseline extraction: Comparison of six methods on ifn/enit database. The 11th International Conference on Frontiers in Handwriting Recognition, 2008. Citeseer.

MENG, G., PAN, C., XIANG, S., DUAN, J. & ZHENG, N. 2012. Metric rectification of curved document images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 34**,** 707-722.

MENG, G., WANG, Y., QU, S., XIANG, S. & PAN, C. Active Flattening of Curved Document Images via Two Structured Beams. Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, 2014. IEEE, 3890-3897.

MENG, G., XIANG, S., PAN, C. & ZHENG, N. 2016. Active Rectification of Curved Document Images Using Structured Beams. *International Journal of Computer Vision***,** 1-27.

MENG, G., XIANG, S., ZHENG, N. & PAN, C. 2013. Nonparametric illumination correction for scanned document images via convex Hulls. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 35**,** 1730-1743.

MOTAWA, D., AMIN, A. & SABOURIN, R. Segmentation of Arabic cursive script. Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on, 1997. IEEE, 625-628.

MOUSA, I. S. 2001. The Arabs in the first communication revolution: the development of the Arabic Script. *Canadian Journal of Communication,* 26.

NAFCHI, H. Z., MOGHADDAM, R. F. & CHERIET, M. 2014. Phase-based binarization of ancient document images: Model and applications. *IEEE transactions on image processing,* 23**,** 2916-2930.

NAZ, S., HAYAT, K., ANWAR, M. W., AKBAR, H. & RAZZAK, M. I. Challenges in baseline detection of cursive script languages. Science and Information Conference (SAI), 2013, 2013. IEEE, 551-556.

NIBLACK, W. 1986. *An Introduction to Digital Image Processing*, Prentice-Hall.

OKABE, A., BOOTS, B., SUGIHARA, K. & CHIU, S. N. 2009. *Spatial tessellations: concepts and applications of Voronoi diagrams*, John Wiley & Sons.

OTSU, N. 1975. A threshold selection method from gray-level histograms. *Automatica,* 11**,** 23-27.

ØYE, V. 2015. *Accelerating nonlinear image transformations with OpenGL ES: A study on fish-eye undistortion.*

PAPADOPOULOS, C., PLETSCHACHER, S., CLAUSNER, C. & ANTONACOPOULOS, A. The IMPACT dataset of historical document images. Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, 2013. ACM, 123-130.

PARHAMI, B. & TARAGHI, M. 1981. Automatic recognition of printed Farsi texts. *Pattern Recognition,* 14**,** 395-403.

PARVEZ, M. T. & MAHMOUD, S. A. 2013. Offline arabic handwritten text recognition: a survey. *ACM Computing Surveys (CSUR),* 45**,** 23.

PECHWITZ, M. & MARGNER, V. Baseline estimation for Arabic handwritten words. Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on, 2002. IEEE, 479-484.

PILU, M. Undoing paper curl distortion using applicable surfaces. Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, 2001. IEEE, I-67-I-72 vol. 1.

PU, Y. & SHI, Z. 1998. A natural learning algorithm based on Hough transform for text lines extraction in handwritten documents.

PUGLIESE, A., POMES, S., FERILLI, S. & REDAVID, D. 2014. A novel model-based dewarping technique for advanced Digital Library systems. *Procedia Computer Science,* 38**,** 108-115.

RAHNEMOONFAR, M. 2010. *Correction of arbitrary geometric artefacts in historical documents.* Salford: University of Salford.

RAHNEMOONFAR, M. & ANTONACOPOULOS, A. Restoration of arbitrarily warped historical document images using flow lines. Document Analysis and Recognition (ICDAR), 2011 International Conference on, 2011. IEEE, 905-909.

RAMDAN, J., OMAR, K. & FYZULL, M. 2016. Segmentation of Arabic VVords Using Area Voronoi Diagrams and Neighbours Graph. *International Journal of Soft Computing,* 11**,** 282-288.

RESTORER, B. 2018. *Heritage Digitization* [Online]. Available: https://www.i2s.fr/en/heritage-digitization/our-markets 2018].

ROMEO-PAKKER, K., MILED, H. & LECOURTIER, Y. A new approach for Latin/Arabic character segmentation. Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, 1995. IEEE, 874-877.

SAFABAKHSH, R. & ADIBI, P. 2005. Nastaaligh handwritten word recognition using a continuous-density variable-duration HMM. *Arabian Journal for Science and Engineering,* 30**,** 95-120.

SAÏDANI, A., ECHI, A. K. & BELAID, A. Identification of machine-printed and handwritten words in Arabic and Latin scripts. Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, 2013. IEEE, 798-802.

SAIDANI, A., KACEM, A. & BELAID, A. Co-occurrence Matrix of Oriented Gradients for word script and nature identification. Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, 2015. IEEE, 16-20.

SALVI, D., ZHENG, K., ZHOU, Y. & WANG, S. Distance Transform Based Active Contour Approach for Document Image Rectification. Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on, 2015. IEEE, 757-764.

SARKIS, K. 21 March 2012 The Influences of Greta Arabic.

SEZGIN, M., SANKUR, B 2004. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging,* 13**,** 146-168.

SHAKOORI, R. A method for text-line segmentation for unconstrained Arabic and Persian handwritten text image. Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on, 2014. IEEE, 338-344.

SHAMGHOLI, M., KHOSRAVI, H. & RIAZI, S. 2014. DOCUMENT IMAGE DEWARPING BASED ON TEXT LINE DETECTION AND SURFACE

MODELING (RESEARCH NOTE). *International Journal of Engineering-Transactions C: Aspects,* 27**,** 1855.

SHAMQOLI, M. & KHOSRAVI, H. Warped document restoration by recovering shape of the surface. Machine Vision and Image Processing (MVIP), 2013 8th Iranian Conference on, 2013. IEEE, 262-265.

SHAPIRO, L. & STOCKMAN, G. C. 2001. Computer vision. 2001. *Ed: Prentice Hall*.

SHARMA, P. & SHARMA, S. Image processing based degraded camera captured document enhancement for improved OCR accuracy. Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference, 2016. IEEE, 441-444.

SHEN, M., GIANNAROU, S. & YANG, G.-Z. 2015. Robust camera localisation with depth reconstruction for bronchoscopic navigation. *International journal of computer assisted radiology and surgery,* 10**,** 801-813.

SLIMANE, F., INGOLD, R., KANOUN, S., ALIMI, A. M. & HENNEBERT, J. Impact of character models choice on arabic text recognition performance. Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on, 2010. IEEE, 670-675.

SONG, J. & LYU, M. R. 2005. A Hough transform based line recognition method utilizing both parameter space and image space. *Pattern recognition,* 38**,** 539-552.

SONKA, M., HLAVAC, V. & BOYLE, R. 2014. *Image processing, analysis, and machine vision*, Cengage Learning.

STAMATOPOULOS, N., GATOS, B. & PRATIKAKIS, I. A methodology for document image dewarping techniques performance evaluation. Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, 2009. IEEE, 956-960.

STAMATOPOULOS, N., GATOS, B., PRATIKAKIS, I. & PERANTONIS, S. J. A two-step dewarping of camera document images. Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on, 2008. IEEE, 209-216.

STAMATOPOULOS, N., GATOS, B., PRATIKAKIS, I. & PERANTONIS, S. J. 2011a. Goal-oriented rectification of camera-based document images. *Image Processing, IEEE Transactions on,* 20**,** 910-920.

STAMATOPOULOS, N., GATOS, B., PRATIKAKIS, I. & PERANTONIS, S. J. 2011b. Goal-oriented rectification of camera-based document images. *IEEE Transactions on Image Processing,* 20**,** 910-920.

SULAIMAN, A., OMAR, K. & NASRUDIN, M. F. A database for degraded Arabic historical manuscripts. Electrical Engineering and Informatics (ICEEI), 2017 6th International Conference on, 2017. IEEE, 1-6.

TAN, C. L., ZHANG, L., ZHANG, Z. & XIA, T. 2006. Restoring warped document images through 3d shape modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 28**,** 195-208.

TANG, Y. Y. & SUEN, C. Y. 1993. Image transformation approach to nonlinear shape restoration. *IEEE transactions on Systems, Man, and Cybernetics,* 23**,** 155-172.

TIMSARI, B. & FAHIMI, H. Morphological approach to character recognition in machine-printed Persian words. Electronic Imaging: Science & Technology, 1996. International Society for Optics and Photonics, 184-191.

TSOI, Y.-C. & BROWN, M. S. Geometric and shading correction for images of printed materials: a unified approach using boundary. null, 2004. IEEE, 240-246.

XU, C. & PRINCE, J. L. 1998. Snakes, shapes, and gradient vector flow. *Image Processing, IEEE Transactions on,* 7**,** 359-369.

YANG, P., ANTONACOPOULOS, A., CLAUSNER, C. & PLETSCHACHER, S. Grid-based modelling and correction of arbitrarily warped historical document images for large-scale digitisation. Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, 2011. ACM, 106-111.

YANG, P., ANTONACOPOULOS, A., CLAUSNER, C., PLETSCHACHER, S. & QI, J. 2017. Effective Geometric Restoration of Distorted Historical Document for Large-Scale Digitization. *IET Image Processing*.

YANG, P., CLAPWORTHY, G., DONG, F., CODREANU, V., WILLIAMS, D., LIU, B., ROERDINK, J. B. & DENG, Z. 2016. GSWO: A programming model for GPU-enabled parallelization of sliding window operations in image processing. *Signal Processing: Image Communication,* 47**,** 332-345.

ZAHOUR, A., LIKFORMAN-SULEM, L., BOUSSALAA, W. & TACONET, B. Text line segmentation of historical arabic documents. Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, 2007. IEEE, 138-142.

ZAHOUR, A., TACONET, B., LIKFORMAN-SULEM, L. & BOUSSELLAA, W. 2009. Overlapping and multi-touching text-line segmentation by Block Covering analysis. *Pattern analysis and applications,* 12**,** 335.

ZAHOUR, A., TACONET, B., MERCY, P. & RAMDANE, S. Arabic hand-written text-line extraction. Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on, 2001. IEEE, 281-285.

ZAKI, A. 2008. Segmentation of Arabic characters using Voronoi diagrams. *University Kebangsaan Malaysia, Malaysia*.

ZEKI, A. M. The segmentation problem in arabic character recognition the state of the art. Information and Communication Technologies, 2005. ICICT 2005. First International Conference on, 2005. IEEE, 11-26.

ZHANG, L., YIP, A. M., BROWN, M. S. & TAN, C. L. 2009. A unified framework for document restoration using inpainting and shape-from-shading. *Pattern Recognition,* 42**,** 2961-2978.

ZHANG, L., YIP, A. M. & TAN, C. L. A restoration framework for correcting photometric and geometric distortions in camera-based document images. Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007. IEEE, 1-8.

ZHANG, L., ZHANG, Z., TAN, C. L. & XIA, T. 3D geometric and optical modeling of warped document images from scanners. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005. IEEE, 337-342.

ZHANG, Y., LIU, C., DING, X. & ZOU, Y. Arbitrary warped document image restoration based on segmentation and Thin-Plate Splines. Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, 2008. IEEE, 1-4.

ZHANG, Y. & TAN, C. L. 2008. An improved physically-based method for geometric restoration of distorted document images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 30**,** 728-734.

ZHANG, Z. Flexible camera calibration by viewing a plane from unknown orientations. Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, 1999. IEEE, 666-673.

ZHANG, Z., LIM, C. & FAN, L. Estimation of 3D shape of warped document surface for image restoration. Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, 2004a. IEEE, 486-489.

ZHANG, Z. & TAN, C. L. Recovery of distorted document images from bound volumes. Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on, 2001a. IEEE, 429-433.

ZHANG, Z. & TAN, C. L. Restoration of images scanned from thick bound documents. Image Processing, 2001. Proceedings. 2001 International Conference on, 2001b. IEEE, 1074-1077.

ZHANG, Z. & TAN, C. L. Straightening warped text lines using polynomial regression. Image Processing. 2002. Proceedings. 2002 International Conference on, 2002. IEEE, 977-980.

ZHANG, Z. & TAN, C. L. Correcting document image warping based on regression of curved text lines. Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on, 2003. IEEE, 589-593.

ZHANG, Z., TAN, C. L. & FAN, L. Restoration of curved document images through 3D shape modeling. Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, 2004b. IEEE, I-10-I-15 Vol. 1.

ZOLAIT, A. H. S. 2013. *Technology Diffusion and Adoption: Global Complexity, Global Innovation: Global Complexity, Global Innovation*, IGI Global.