Sudan University of Science and Technology

College of Graduate Studies

College of Computer Science and Information Technology

Enhancing the Accuracy of Optical Character Recognition (OCR) for short text of Arabic language Image by implementing an algorithm to improve the Quality of Images

تحسين دقة التعرف الضوئي على الحرف للنصوص القصيرة في اللغة العربية بواسطة بناء خوارزمية لتحسين جودة الصور

A Thesis Submitted in Partial Fulfillment of the Requirements for

MSc

degree in Computer Science

By:

Ahmed Suliman Albashir Mohamed

Supervisor:

Dr. Mohammed Hamouda Karboos Hamid

May 2022

Declaration

I hereby declare that the work reported in this MSc thesis titled "**Enhancing the Accuracy of Optical Character Recognition (OCR) for short text of Arabic language Image by implementing an algorithm to improve the Quality of Images**" submitted for the Sudan University of Science and Technology, is an authentic record of my work carried out under the supervision of **Dr.** Mohammed Hamouda Karboos Hamid. And never has been submitted elsewhere for other degrees.

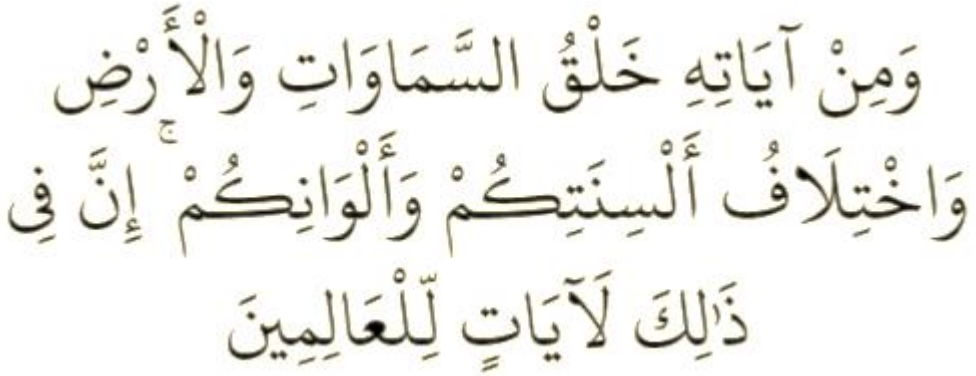
Ahmed Suliman Albashir Mohamed

Dr. Mohammed Hamouda Karboos Hamid

Supervisor

Introductive

الأية



سورة الروم الاية 22

Dedication

I extend the highest expressions of gratitude and appreciation to my family, my compassionate mother, my father, the educator, my brothers and sisters, my friends, my companions in my master's studies, my teachers, who stood by me and provided me with support and advice.

Sincerely grateful …

Ahmed Suliman Albashir Mohamed

Acknowledgment

I wish to express my deepest appreciation to all those who helped me, in one way or another, to complete this project. First and foremost I thank Allah almighty who provided me with strength, purpose, and success throughout this endeavor. Special thanks to my supervisor Dr. Mohammed Hamouda Karboos Hamid for all his patience, expert guidance, and support during the execution of this research, And also Dr. Nesren Alass for her consultation and advice. My thanks also go to staff of **Sudan University of Science and Technology** for making their resources available for researchers and students.

Abstract

Optical Character Recognition (OCR) plays a major role in understanding, learning, and recognizing the language in the era of communication. OCR helps non-native speakers and even non-humans to understand the language and recognize its texts, words, phrases, and structures. Although, Optical Character Recognition provides more accurate way to recognize texts, but there is a lack of sufficient interest and support for Arabic languages in this field compared to other languages, especially English.

This research aims to implement an algorithm for enhancing the accuracy of Arabic text recognition through improving image quality. This can be conduct via image processing which performs a set of image processing operations, and repeating the process several times to achieve maximum accuracy, so that text recognition software can easily detect texts. This could be done through a direct application in an experimental environment.

The average similarity rate of the original images without modification was (0.50) to (1). The average similarity rate of texts for images after improving was reached (0.91) to (1) which is a much better result. The results showed that many future improvements can be made to obtain a greater similarity rate by improving images and using artificial intelligence.

المستخلص

في عصر التواصل ، يلعب التعرف على النصوص دورا رئيسيا في فهم وتعلم والتعرف على اللغة خاصة لغير المتحدثين بها وحتى لغير البشر حتى يستطيع الحاسوب فهم اللغة والتعرف على نصوصها وكلماتها وعباراتها وتراكيبها. والأهم من ذلك أنه يقدم طريقة اكثر دقة للتعرف على النصوص. لكن المشكلة الجلية في هذا المجال هي عدم وجود إهتمام ودعم من قبل المبرمجين والمطورين العرب للغة العربية بالقدر الكافي مقارنة باللغات الاخرى خاصة اللغة الانجليزية.

يهدف هذا البحث لتطبيق خوارزمية تعمل على تحسين دقة التعرف على النصوص للغة العربية عن طريق تحسين جودة الصور بواسطة معالجة الصور وذلك من خلال عمليات معالجة الصور لتحسينها وتكرار العملية عدة مرات للوصول الي اقصى دقة ، لكي تتمكن برمجية التعرف على النصوص من اكتشاف النصوص بسهولة ، وتم ذلكَ بالتطبيق التجريبي المباشر للخوارزمية.

بلغ متوسط معدل الصور الاصلية بدون تعديل (0.50) الي (1) تم الوصول الي متوسط معدل التشابه للنصوص للصور بعد تحسيناها (0.91) الي (1) وهي نتيجة افضل بكثير، أظهرت النتائج أنه يمكن القيام بالعديد من التحسينات المستقبلية للوصول الي معدل تشابه اكبر وذلك بتحسين الصور واستخدام تقنيات الذكاء الاصطناعي.

Table of contents

List of figures

List of Tables

List of abbreviations

TERM	MEANING
OCR	Optical Character Recognition
JND	Just Noticeable Difference
OpenCV	Open Source Computer Vision Library
IQA	Image Quality Assessment
AI	Artificial Intelligent
DSP	Digital Signal Processor
LCS	Longest Common SubString

CHAPTER ONE

INTRODUCTION

1. CHAPTER I

1.1. Introduction

Image quality directly affects the accuracy of Optical Character Recognition (OCR), As long as the image quality is excellent, the results will be more accurate, especially if it has been developing to discover patterns and printed letters by using the concept of Optical Character Recognition (OCR) in the Arabic language, regarding the difficulty of installing Arabic language and the patterns of different letter shapes and letters attached and separated.

This research aims to implement an algorithm to improve the accuracy of the pattern detection tool (Tesseract OCR Tool) using the standard of similarity in texts as a concept for calculating accuracy, A set of operations to the images will be conducted to measure the effectiveness of performance in discovering the changes that occurred in the same images and calculating the proportions in disparity.

Perceptual image quality assessment plays an important role in digital image technology, such as the development and optimization of image compression and transmission schemes. Subjective quality assessment is considered to be the most reliable way to evaluate the quality of image presentations, but it is time-consuming. Over the years, some researchers have contributed significant research in the design of image quality assessment algorithms, claiming to have made headway in their respective domains (Junyong You, Andrew Perkis 2010).

1.2. Problem Statement

Nowadays, the challenges that face Arabic OCR systems stem from the cursive and continuous nature of Arabic scripts. It difficult task for many reasons such as low scanning and printing quality Thus, lead to bad results for text recognition. This difficulty is increased in case of recognition of a high inflected language such as Arabic language, due to the morphological and script characteristics of Arabic language

There are very few tools for recognizing Arabic texts provided by the Arab developer community. On the other hand, many tools recognizing texts in other languages, including Arabic.

1.3. Research Questions

Regarding the problem statement mentioned in previous section, the research question can be formulated as follow: how can the best result of accuracy and efficiency of a tool be obtained in context Arabic recognition after enhancing image quality?

1.4. Objectives of the research

The main objective of this research is to develop a framework to get the best results by improve the quality of images by image processing, which will be achieved by:

- Collecting images that reflect and represent the differences and overlap of Arabic letters.
- Building an algorithm that deals with images and enhance it.
- Applying the algorithm to images and calculate the similarity rate
- Comparing and evaluating the results, to fulfill the aim of this research.

1.6. Research methodology

This research focuses on implement an algorithm to enhance the quality of images in order to get the best results from reading characters; the similarity rate is used to compare results with the predicted and detected text.

To implement an algorithm set of operations or processes will be done to the images, to increasing the quality of images using image processing (Rescaling, Binarisation, Noise Removal, Dilation, Erosion, Rotation, Deskewing, Borders, and Transparency).

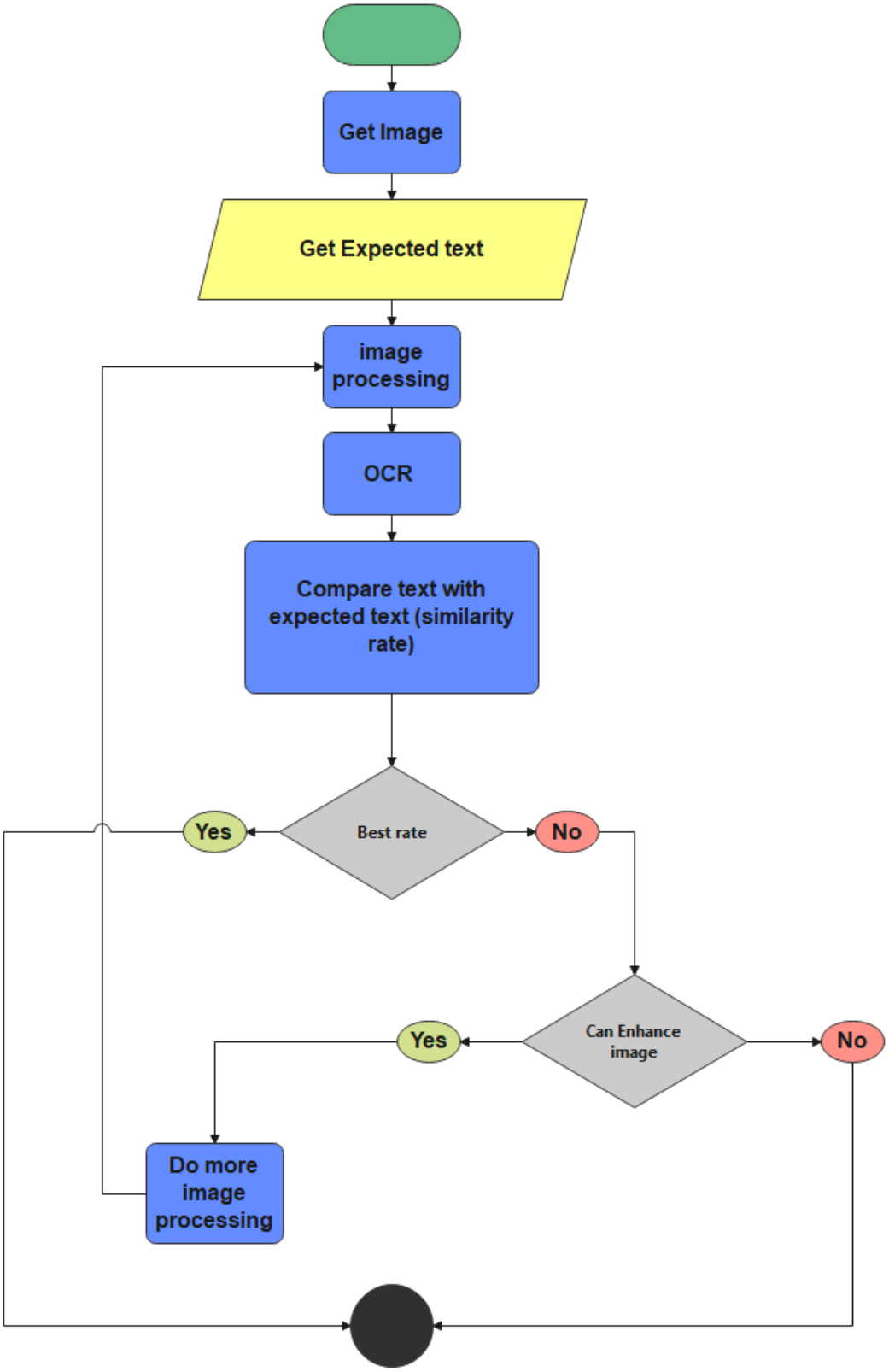


Figure 1.2: research schema

1.7. Scope of the research

This research aims to implement an algorithm of OCR tool in short text of Arabic language to enhance the accuracy of OCR by improving the quality of the images to get the best results from the tool to discover patterns and extract text from images.

1.8. Thesis organization

This research organized as follows:

Chapter I contains the research problem statement and objectives.

Chapter II discusses the literature review and related work.

Chapter III describes the research methodology and the implementation of the techniques used

Chapter IV presents the Experiments and results. Finally.

Chapter V concludes this research and presents Recommendations for future works.

CHAPTER TWO

LITERATURE REVIEW

2. CHAPTER II

2.1. Introduction

This chapter provides the problem background and some ideas, reviews of OCR technologies, explaining some of the older technologies and the studies that done in the Arabic language and other languages. It also includes insights into the architecture, key components, and functionality of optical character recognition and image processing.

Alongside several linguistic information about the Characteristics of Arabic language, are also discussed lastly it mentions the literature and related work in Optical character recognition and gives a summary for all of them.

2.2. Problem Background

Arabic is the official language of 21 countries in the Middle East and North Africa, from Oman in the east to Mauritania in the west. This includes Israel, where Arabic is, after Hebrew, the second official language. Significant Arab minorities exist in Iran, Turkey, Chad, and Nigeria, as well as in Western Europe and the Americas. With approximately 280 million native speakers, Arabic is by far the largest living representative of the Semitic language family. Because it is the language of the Koran and thus the liturgical language of Islam, Arabic also plays an important role for more than 1 billion Muslims worldwide (Procházka 2006).

OCR technology is considered a challenging research area in the field of pattern recognition and artificial intelligence. Many studies have been proposed for Latin and non-Latin scripts. However, the development of OCR applications for Latin script is easier than that of Arabic scripts because of Arabic writing characteristics. Figure 2.1 illustrates some characteristics of the printed Arabic script that contribute to the challenges in Arabic OCR evolution. Compared to printed Latin script, Arabic script is written cursively from right to left and contains dots and diacritics. Also, a character of an Arabic script may have four dissimilar shapes concerning its location in an Arabic word (Teahan 2016).

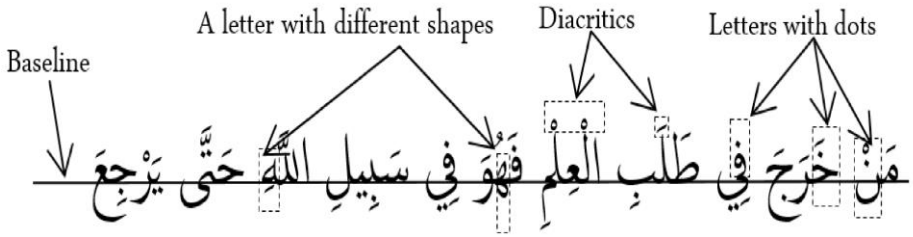


Figure 2.1: Arabic script characteristics

Tesseract is an open-source OCR engine that was developed at HP between 1984 and 1994. Like a super-nova, it appeared from nowhere for the 1995 UNLV Annual Test of OCR Accuracy, shone brightly with its results, and then vanished back under the same cloak of secrecy under which it had been developed. Now for the first time, details of the architecture and algorithms can be revealed (Smith 2005).

They are for the research question can be formulated as to how to get the best results of accuracy and efficiency of the tool on the context Arabic recognition after improving image quality.

2.3. Optical character recognition

Optical Character Recognition (OCR) is a process of converting a machine-printed or handwritten text image into a digital computer format that can be editable (Teahan 2016).

Optical character recognition (OCR) plays an important role in transforming printed materials into digital text files. These digital files can be very helpful to kids and adults who have trouble reading. That’s because digital text can be used with software programs that support reading in a variety of ways.

Character recognition is not a new problem but its roots can be traced back to systems before the inventions of computers. The earliest OCR systems were not computers but mechanical devices that we're able to recognize characters, but very slow speed and low accuracy. In 1951, M. Sheppard invented a reading and robot GISMO that can be considered as the earliest work on modern OCR(Islam, Islam, and Noor 2017).

2.4. Optical character recognition techniques

OCR consists of many phases such as Pre-processing, Segmentation, Feature Extraction, Classification, and Recognition Table 2.1: describe these phases. The task of preprocessing relates to the removal of noise and variation in handwritten. Several areas where OCR used including mail different bank processing, reading a document and postal address recognition require offline handwriting recognition systems, pattern recognition (Sarika Pansare and Sinhgad 2017).

The figure below show the components of OCR system

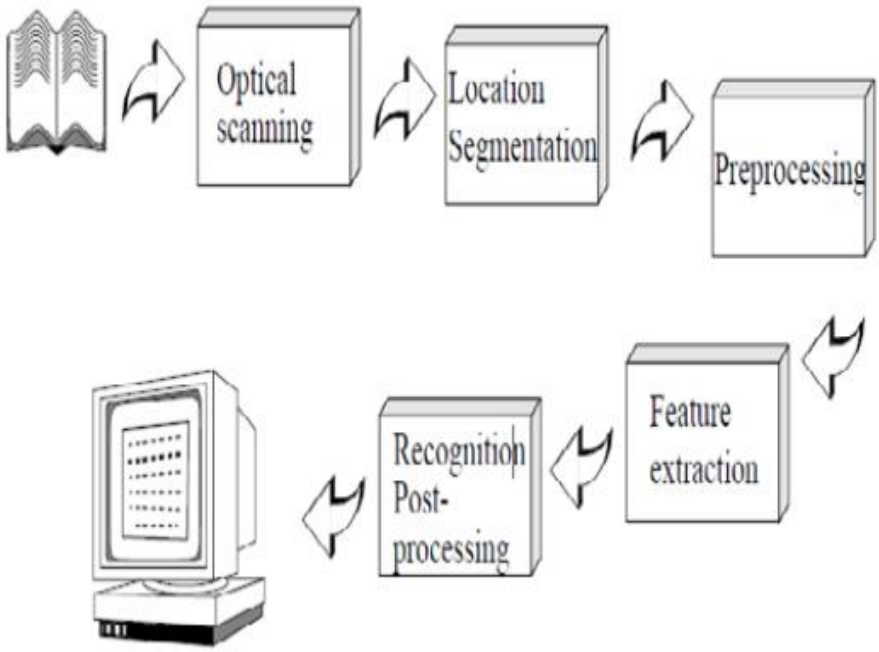


Figure 2.2: components of OCR system

Phase	Description	Approaches
Acquisition	The process of acquiring image	Digitization, binarization, compression
Pre-processing	To enhance quality of image	Noise removal, Skew removal, thinning, morphological operations

Segmentation	To separate image into its constituent characters	Implicit Vs Explicit Segmentation
Feature Extraction	To extract features from image	Geometrical feature such as loops, corner points Statistical features such as moments
Classification	To categorize a character into its particular class	Neural Network, Bayesian, Nearest Neighborhood
Post processing	To improve accuracy of OCR results	Contextual approaches, multiple classifiers, dictionary based approaches

Table 2.1: major phases of OCR system

- Acquisition

  Earlier, flat-bed scanners were used for obtaining clear images of standard quality and suitable for recognition. Using a scanner has benefits such as low noise levels, virtually no blurring and low text skewing.(R.Kiran 2015)

- Pre-processing

  The scanned image contains a certain amount of noise. The characters may be smeared or broken which is depend on the resolution on the scanner and the success of the applied technique for thresholding. In preprocessor some of these defects, which may later cause poor recognition rates, can be eliminated to smooth the digitized characters.(Modi and C. 2017)

- Segmentation

  Is one of the most important phases in OCR development? It directly affects the efficiency of any OCR. So a good segmentation technique can

increase the performance of OCR. Segmentation is process of extracting the basic constituent symbols of the script, which are indi- vidual characters. It is necessary to segment the script at character level, as classifier works at character level only. Preprocessing in form of text digitization and skew correction is performed before applying segmentation.(Yadav, Sánchez-Cuadrado, and Morato 2013)

- Feature Extraction

  In this phase, features of individual character are extracted. The performance of an each character recognition system that depends on the features that are extracted. The extracted features from input character should allow classification of a character in a unique way. We used diagonal features, intersection and open end points features, transition features, zoning features, directional features, parabola curve fitting–based features, and power curve fitting–based features in order to find the feature set for a given character.(Tomar and Kishore 2018)

- Classification

  OCR systems broadly utilize the methodologies of pattern recognition, which assigns each example to a predefined class. Classification is the procedure of distributing inputs with respect to detected information to their comparing class in order to create groups with homogeneous qualities, while segregating different inputs into different classes.(Hamad and Kaya 2016)

- Post processing

  A final step of post-processing is necessary to aggregate these characters into words and separate them with spaces to generate a meaningful text document.(Osman et al. 2020).

2.5. Image processing

Image processing is a method to perform some operations on an image, to get an enhanced image, or to extract some useful information from it. It is a type of signal processing in which input is an image and output may be an image or characteristics/features associated with that image. Nowadays, image processing is among rapidly growing technologies. It forms a core research area within engineering and computer science disciplines too.

Image processing is a method to convert an image into digital form and perform some operations on it, in order to get an enhanced image or to extract some useful information from it. It is a type of signal dispensation in which the input is an image and the output is an image or characteristics associated with that image.(Covington 2009)

Image processing is a form of signal processing in which the input is an image such as a photograph or video frame, the output is an image or set of characteristics related to image.(Naveenkumar and Ayyasamy 2016)

2.6. Image quality

Image quality measurement is very important for various image processing applications such as recognition, retrieval, classification, compression, restoration, and similar fields. The images may contain different types of distortions like blur, noise, contrast change, etc. So it is essential to rate the image quality appropriately. Traditionally subjective rating methods are used to evaluate the quality of the image, in which humans rated the image quality based on time requirements. This is a costly process and it needs experts for evaluating image quality. Nowadays many image quality assessment algorithms are available for finding the quality of images. These are mainly based on the properties of the human visual system (George and Livingston 2013).

Historically, image quality is described in terms of the visibility of the distortions in an image, such as cooler shifts, blurriness, Gaussian noise, and blockiness. The most common way of modeling an image quality metric is therefore by quantification of the visibility of these distortions. For example, the Just Noticeable

Difference (JND) model has been designed by Sarnoff predicts the subjective rating of an image by examining the visibility of distortions (Thung 2009).

There are basically two types of image quality assessment (IQA) techniques, namely the subjective method, which involve human beings to evaluate the quality of the images, and the objective method, which compute the image quality automatically (please refer to Figure. 2.2) (Thung 2009)

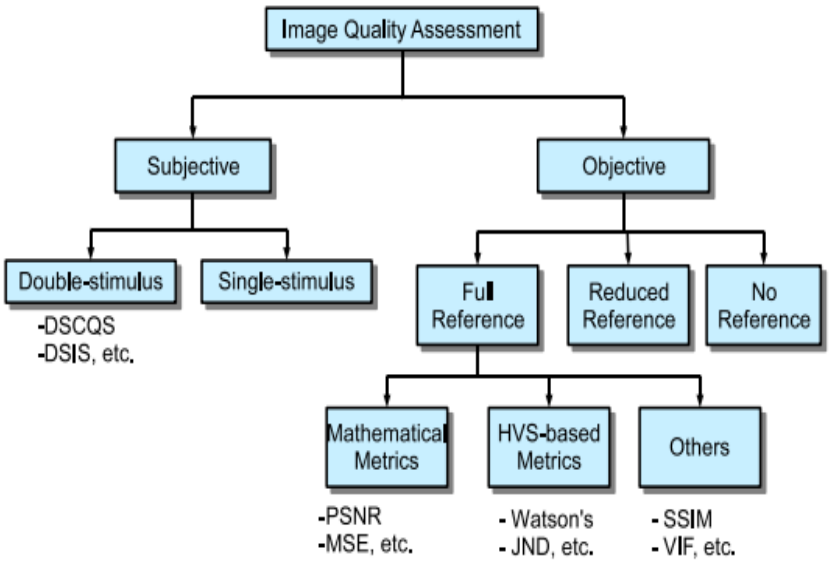


Figure 2.3: image quality assessment

2.7. Characteristics of Arabic text

An Arabic text is composed by placing linearly character blocks of varying sizes from right to left. The peculiar characteristic of the Arabic text is that the shape of characters may significantly vary within a word. This variation depends on: the position of the character within a word and its adjacent characters. On the basis of this characteristic we divide Arabic character shapes into two categories: The primary category that consists of twenty eight characters (Table 2.2) and the secondary category that consists of eight characters (Table 2.3). There are four characters that do not fit into this categorization, therefore, they form a special category (Table 2.4)(Ahmed and Al-Ohali 2000).

SN	Character name	Shape position			
		End	Middle	Begn.	Isold.
1	Alif	ـا			ا
2	Ba	ـب	ـبـ	بـ	ب
3	Ta	ـت	ـتـ	تـ	ت
4	Tha	ـث	ـثـ	ثـ	ث
5	Jim	ـج	ـجـ	جـ	ج
6	HHA	ـح	ـحـ	حـ	ح
7	KHA	ـخ	ـخـ	خـ	خ
8	Dal	ـد			د
9	Thal	ـذ			ذ
10	Ra	ـر			ر
11	Zai	ـز			ز
12	Sin	ـس	ـسـ	سـ	س
13	Shin	ـش	ـشـ	شـ	ش
14	Sad	ـص	ـصـ	صـ	ص

SN	Character name	Shape position			
		End	Middle	Begn.	Isold.
15	Dhad	ـض	ـضـ	ضـ	ض
16	Tua	ـط	ـطـ	طـ	ط
17	Zua	ـظ	ـظـ	ظـ	ظ
18	Ain	ـع	ـعـ	عـ	ع
19	Gain	ـغ	ـغـ	غـ	غ
20	Fa	ـف	ـفـ	فـ	ف
21	Qaf	ـق	ـقـ	قـ	ق
22	Kaf	ـك	ـكـ	كـ	ك
23	Lam	ـل	ـلـ	لـ	ل
24	Mim	ـم	ـمـ	مـ	م
25	Non	ـن	ـنـ	نـ	ن
26	Ha	ـه	ـهـ	هـ	ه
27	Waw	ـو			و
28	Ya	ـي	ـيـ	يـ	ي

Table 2.2: the primary category

SN	Character name	Shape position		
		End	Middle	Isolated
1	Hamzah			ء
2	Alif-Muqsorah	ـى		ى
3	Ta-Marbotah	ـة		ة
4	Hamzah-Waw	ـؤ		ؤ
5	Hamzah-Ya	ـئ	ـئـ	ئ
6	Hamzah-Alif	ـأ		أ
7	Hamzah-Alif	ـإ		إ
8	Alif-Mad	ـآ		آ

Table 2.3: the secondary category

SN	Character name	Shape position	
		End	Isolated
1	Lam-Alif	ﻠﺎ	ﻻ
2	Lam-Alif	ﻠﺎ	ﻻ
3	Lam-Alif	ﻠﺎ	ﻻ
4	Lam-Alif	ﻠﺎ	ﻻ

Table 2.4: special category character set

2.8. Related optical character recognition researches

This research conducted an investigative search for scientific papers related to visual character recognition problems in Arabic, especially problems related to image quality.

In (Rawat, Sharma, and Gusain 2021) : They investigate the usage of Tesseract OCR by applying images preprocessing for Hindi, in extracting the text from the sampled images of Garhwali textbooks. They analyzed the effect of various image preprocessing techniques to improve the OCRed results, and compare the performance of these techniques using quantitative measures.

This paper is agree with this research in many sides of image preprocessing but it applied only for Hindi

In (Aliwy and Al-sadawi 2021) : it suggested a post-processing of OCR outputs with natural language processing techniques and reduce the noise in the combined text of Corpus files , The proposed system is based on dictionary and N-gram language model LM constructed from the huge corpus.

It proposed AOCR post-processing but what if the quality of images is low, many lack results will occurred.

In (Francisca O Nwokoma et al. 2021) : reviews the various factors that increase the computational difficulties of Camera-Based OCR, and made some recommendations as per the best practices for Camera-Based OCR systems.

It give many ideas, techniques and practices that help in ORC systems to get best results.

In (Li et al. 2020) : They proposes a text based on Bezier curve correction method, through the experiment, can be good in Chinese and English text image to correct, after correction of OCR recognition rate is correct before have greatly improved, and the single image correction takes 8.9s, can meet the demand of daily OCR transformation, to further the system of packaging, to meet the demand of print text conversion.

It deal with only Bezier curve correction method applied to Chinese and English language not all texts are in books.

In (Alhomed and Jambi 2018) : The paper collected scientific papers related to the Arabic language, the efforts made in developing the visual discovery of the texts and techniques presented by each paper, the challenges faced by the papers, the advantages, and disadvantages that help in adaptation or extension of these systems to fit the recent demands.

It agree with this research in many ways, it poses new challenges and showed good possibilities. This is because most of the recent trends can used existing OCR systems with some common image processing techniques for resizing, classification and image enhancement that has been used to get best results.

A new method to restore characters quality in weak resolution images before the OCR step. has been presented by (Ugale and Joshi 2017), The paper purposed forth methods to deal with improved performance of OCR : (M1)The first method used in this paper uses bicubic interpolation. The second preprocessing method is an image sharpening method. (M2) The main reason for applying this method is to enhance the contrast between edges. (M3) The third proposed method is image blurring. It reduces high-frequency information and removes noise from the images which can cause a lower OCR rate. They proposed an equation for this process as follow:

$$h(i,j) = \frac{1}{81} \sum_{k=i-4}^{i+4} \sum_{l=j-4}^{j+4} f(k,l)$$

(M4) The fourth preprocessing method is applied to the cases when the image has a colorful background. The idea is to separate the text from the background.

It agree with this research in many ways, but not all techniques that can help ORC systems for extract best results from low quality images.

In (Habeeb et al. n.d.2014) the study has presented a new method to restore characters quality in weak resolution images before the OCR step, the paper proposed six steps : (S1) extract words' images from document image and store in an array, (S2) pass each word image in the array sequentially to four processes, (S3) perform cleaning, restoration, and resample on each word image in any of the four processes based on different conditions, (S4) each OCR engine will receive words images from one process in sequence, (S5) apply a procedure to select the best word resulting from the four OCR engines, and (S6) compile all words in one output text. Figure 2.3 shows the proposed method (Habeeb et al. 2014.).

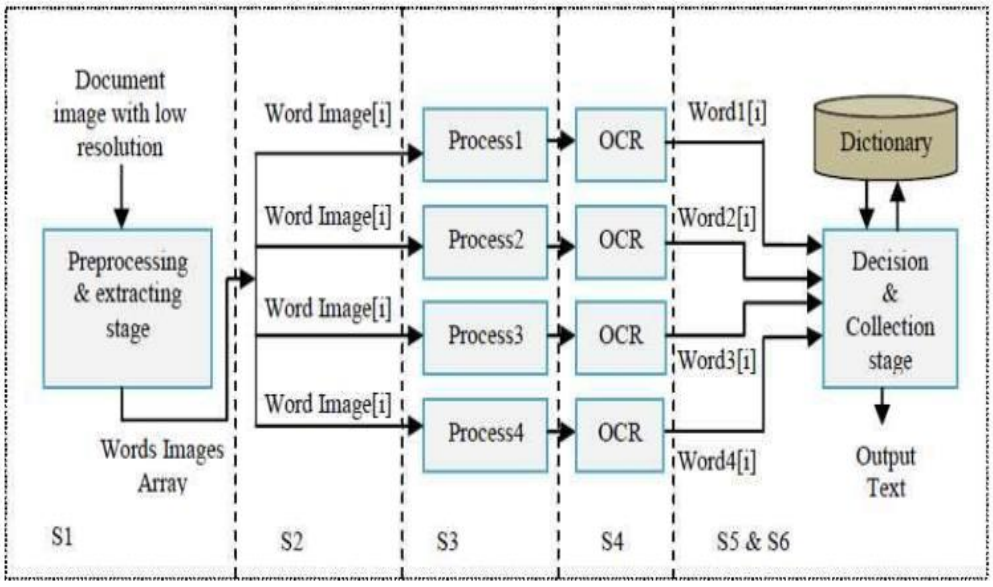


Figure 2.4: the proposed method framework in habeeb et al. 2014

The paper extract words' images from document image and store in an array this may take a time to get best results.

2.9. Summary of literature

AUTHOR, YEAR	TECHNOLOGIES	STRENGTH POINTS	WEAK POINTS
(Rawat et al. 2021)	Otsu, adaptive Thresholding and ImageMagick	Applied multi image processing methods	Manually corrected the text files
(Aliwy and Al-sadawi 2021)	Corpus-based error correction	Very good improvement in correction of errors of AOCR systems.	It may get lack results in low resolution and bad quality
(Francisca O Nwokoma et al. 2021)	Camera-Based OCR systems	Focus on Camera-based scene text detection and recognition technique	cameras usually have low resolutions
(Li et al. 2020)	Bezier curve correction method For English and Chinese language	Effective way to convert paper text into digital problem , especially books	It focus only on curve correction method and skew correction
(Alhomed and Jambi 2018)	Various aspects of the current OCR systems for Arabic language	It discussed pre-processing and post processing operations to enhance the produced results.	It done in past the new challenges has been occur
(Ugale and Joshi 2017)	Four different preprocessing methods for Improving Tesseract OCR performance on	It focus in low quality images techniques has been applied.	Low results has showed and It done only for English language.

	images grabbed from STBs are proposed.		
(Habeeb et al. n.d.2014)	A new method which can restore characters' quality in weak resolution images before passing them to OCR engines.	The method excludes traditional alignment among resulting texts used by related methods; and also no training is needed on errors before executing it	It correct words one by one this take a time for long statements and done only for English language.

Table 2.5: summary of literature

2.10. Summary

This chapter reviewed surveys and journal papers about optical character recognition and mentioned most of the techniques related to it. The survey showed that the field of optical character recognition is broadly investigated in the Arabic Language. Moreover, the literature review showed that a lot of work need to be done in the field of optical character recognition particularly for the Arabic language, therefore this research adopts a methodology based on image processing to design optical character recognition framework for the Arabic language which is going to be explained in the next chapter.

CHAPTER THREE

METHODOLOGY & IMPLEMENTATION

3. CHAPTER III

3.1. Introduction

This chapter is carried out to emphasize and describe the methodology used to fulfill the sheer objective of this research. The endower is to develop an optical character recognition framework for the Arabic language, the review of most used techniques in the field of optical character recognition was investigated both their implementation and complications alongside the datasets used respectively with each research mentioned in chapter two.

The methodology was attained by contemplating each of the objectives thoroughly as showed in Figure 1.2, from collecting representable images for the Arabic language, applying several stages of OCR and image processing of the collected images, and designing a framework using image processing by implementing (Rescaling, Binarisation, Noise Removal, Dilation, Erosion, Rotation, Deskewing, Borders, and Transparency).

3.2. Images collection

In order to apply the algorithm test, images were collected from the Internet, which are snippets of texts in Arabic that contain colored backgrounds, various distortions, and patterns that affect the discovery technology, and the purpose of this is to put all the hypotheses to test the algorithm by applying the highest degree of difficulty.

Figures (3.1 - 3.2 - 3.3 - 3.4) respectively are samples uses in this work

Figure 3.1: Image A

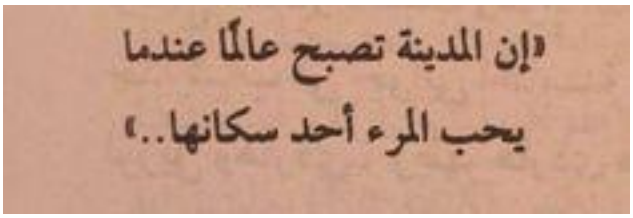


Figure 3.2: Image B

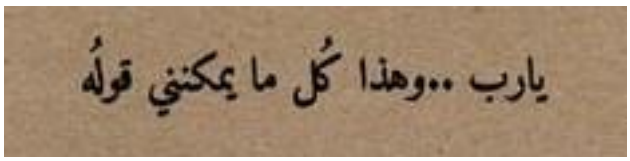


Figure 3.3: Image C

Figure 3.4: Image D

3.3. Images preparing & Algorithm

After obtaining the images, in next stage, the images are prepared and previewed to obtain the expected text and save it to compare it with the discovered texts. After each stage of image modification, the detected text is compared with the original text and the similarity rate is calculated. In the event that the detected similarity rate is further to the expected text, an adjustment is made again to the image, and the process is repeated until the similarity rate closest to the expected text is reached.

The algorithm has been tested on many images to find out the appropriate pattern to get the best results and similarity rates.

3.4. OpenCV operations of Images

At this stage, a set of image processing operations are applied to the selected image, these operations include:

- Gray Scale:

  Transformations within RGB space like adding/removing the alpha channel, reversing the channel order, conversion to/from 16-bit RGB color (R5:G6:B5 or R5:G5:B5), as well as conversion to/from grayscale using (Opencv.org 2022a):

  $$RGB[A]\ to\ Gray: Y \leftarrow 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$$

  And

  $$Gray\ to\ RGB[A]: R \leftarrow Y, G \leftarrow Y, B \leftarrow Y, A \leftarrow max(ChannelRange)$$

- Remove Noise:

  The function smoothes an image using the median filter with the $ksize \times ksize$ aperture. Each channel of a multi-channel image is processed independently. In-place operation is supported(Opencv.org 2021).

- Image Thresholding

  The function applies fixed-level Thresholding to a multiple-channel array. The function is typically used to get a bi-level (binary) image out of a grayscale image (compare could be also used for this purpose) or for removing a noise, that is, filtering out pixels with too small or too large values(Opencv.org 2022c).

  $$dst(x,y) = \begin{cases} maxValue\ \ if\ src(x,y) > T(x,y) \\ 0\ otherwise \end{cases}$$

- Dilation

  The function dilates the source image using the specified structuring element that determines the shape of a pixel neighborhood over which the maximum is taken:

$$sdst(x,y) = \max_{(x',y') : element(x',y') \neq 0} src(x+x', y+y')$$

- Canny

The function finds edges in the input image and marks them in the output map edges using the canny algorithm. The largest value is used to find initial segments of strong edges(Opencv.org 2022b).

- Invert image

  Bitwise operations helps in image masking. Image creation can be enabled with the help of these operations. These operations can be helpful in enhancing the properties of the input images.(Geeksforgeeks.org 2021)

- Image blurring

  Image blurring is achieved by convolving the image with a low-pass filter kernel. It is useful for removing noise. It actually removes high frequency content (e.g. noise, edges) from the image. So edges are blurred a little bit in this operation (there are also blurring techniques which don't blur the edges). OpenCV provides four main types of blurring techniques.(Opencv.org 2021)

- Linear image

  Linear image processing is based on the same two techniques as conventional DSP: convolution and Fourier analysis. Convolution is the more important of these two, since images have their information encoded in the spatial domain rather than the frequency domain. Linear filtering can improve images in many ways: sharpening the edges of objects, reducing random noise, correcting for unequal illumination, deconvolution to correct for blur and motion, etc.(Steven W. Smith 2021)

- Resize image

  Increasing the image size shows some details that may not be visible at normal size.

- Enhance image

  Image enhancement aims to improve image quality in terms of colors, brightness, and contrasts. Earlier methods are mainly based on histogram equalization and gamma correction. Although these methods are simple

and fast, their performance are limited by the fact that individual pixels are enhanced without consideration to contextual information(Vu et al. 2018).

- Contrasted , Brightness image

  Contrast enhancement in digital images is an important technique in image processing. This image enhancement technique helps human vision to better value details, Moreover, it is used as a preprocessing for other applications that need to enhance image quality.(Wei et al. 2021).

- Colored image

  Colorization is, in its essence, a process of assuming color Information where it is absent. In a technical sense, it is a challenging process of assigning three-dimensional RGB (Red, Green, and Blue) color information to every pixel with respect to intensity of a grayscale image in a visually acceptable, plausible way(Žeger et al. 2021).

- Sharper image

  Image sharpening is a powerful tool for emphasizing texture and drawing viewer focus. It's also required of any digital photo at some point -whether you're aware it's been applied or not. Digital camera sensors and lenses always blur an image to some degree, for example, and this requires correction. However, not all sharpening techniques are created equal.(P.Sathya, M.Bhuvaneswari, G.Kesavaraj 2016)

- Remove gray

  Colors make images more vivid. They can now be recorded with a point-and-shoot camera easily. However, people often want to add colors to old monochrome photos, and pictures are sometimes shot with severely wrong white balance settings, in such a case, a possible remedy is to keep only the captured intensities and transfer colors from another source to it.(Li, Hao, and Zhang 2008)

3.5. Similarity rate stage

After applying the image processing operations, the next stage is to calculate the similarity rate of the texts. In the event that the last similarity rate is the largest, the algorithm is stopped and the process is not repeated again to get the best similarity

rate. After obtaining the best similarity rate, the results are displayed and the image that appears will be obtained and also the last highest similarity rate.

In this research the approach that used is Longest Common SubString (LCS) of String-based that will discuss in section 3.8

After discussing the methodology, next section will be about the techniques and tools that help to implement.

3.6. Tesseract OCR engine

Tesseract is an open source optical character recognition engine. Tesseract began as a PhD research project in HP Labs, Bristol. It was further developed at HP in between 1984 to1994. It was modified and improved in 1995 with greater accuracy. In late 2005, HP released Tesseract for open source and now is available.(Nair 2017)

3.7. The Python programming language

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. Tesseract is an open source optical character recognition engine.(Covington 2009)

3.8. Text Similarity

Measuring the similarity between words, sentences, paragraphs and documents is an important component in various tasks such as information retrieval, document clustering, word-sense disambiguation, automatic essay scoring, short answer grading, machine translation and text summarization.(H.Gomaa and A. Fahmy 2013).

They are many approaches to measure text similarity but the best way is to choose the approach that can get best results, Different approaches have been promoted to measure the similarity between one texts with another.

The similarity method is divided into four major groups, String-based, Corpus-based, Knowledge-based, and Hybrid text similarities; as shown in Figure 3.1.(Prasetya, Wibawa, and Hirashima 2018)

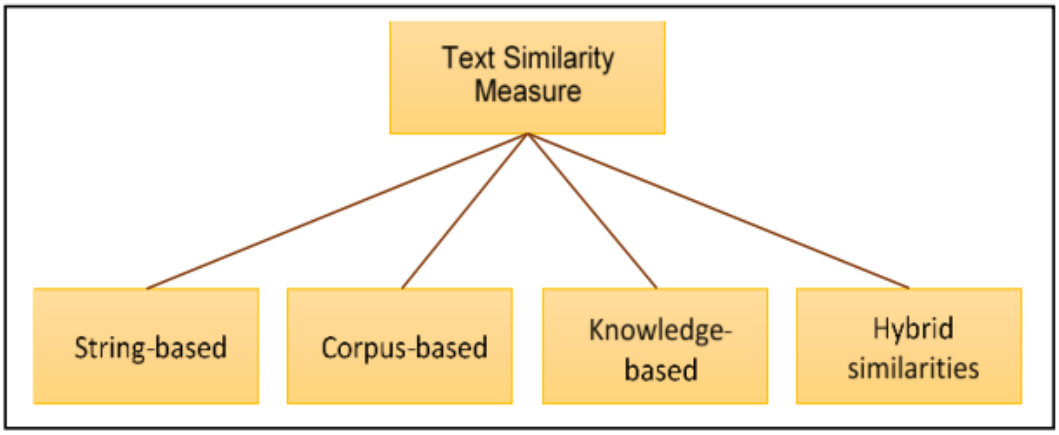


Figure 3.5: Four major groups of text similarity methods and algorithms

- **String-based**

  String similarity measures operate on string sequences and character composition. A string metric is a metric that measures similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison it use two approaches:

  1. Character-Based Similarity:

     It content many algorithms but is this research used Longest Common SubString (LCS) algorithm considers the similarity between two strings is based on the length of contiguous chain of characters that exist in both strings

  2. Term-based Similarity Measures.

.(H.Gomaa and A. Fahmy 2013).

- **Corpus-based**

  Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. (H.Gomaa and A. Fahmy 2013).

- **Knowledge-based**

  Knowledge-Based Similarity is one of semantic similarity measures that bases on identifying the degree of similarity between words using information derived from semantic networks. (H.Gomaa and A. Fahmy 2013).

- **Hybrid text similarities**

  Hybrid methods use multiple similarity measures, these previous method were evaluated separately, then they were combined together. The best performance was achieved using a method that combines several similarity metrics into one. (H.Gomaa and A. Fahmy 2013).

3.9. Summary

In this chapter, the methodology that was followed is described, from the first step of collecting images and building the software that processes and applies the algorithm from processing images, repeating the operations related to them, calculating the similarity rate and comparing it with the expected text entered manually, As long as the rate can be improved, the images are repeatedly improved until obtaining At the best similarity rate and then stop. Then the techniques used, software and tools that helped in applying the methodology were discussed.

CHAPTER FOUR

EXPERIMENTS & RESULTS

4. CHAPTER IV

4.1. Introduction

In this chapter, the proposed algorithm is tested to obtain the best similarity rate compared to the expected text by applying the image processing operations described in the methodology of this research.

4.2. Empirical Implementation

To verify the proposed methodology, the direct application and actual testing of the conditions and the difference in similarity rates for each stage of the proposed image processing operations are carried out.

4.2.1. Applying methodology

The application method executed on environment Windows 10 Pro 64-bit (10.0, Build 19042) with processor Intel(R) Core(TM) i7-4510U CPU @ 2.00GHz (4 CPUs), ~2.6GHz with memory 8192 MB RAM and embedded graphic card: Intel(R) HD Graphics Family with display memory: 2160 MB dedicated memory 112 MB with shared memory: 2048 MB and integrated graphic card NVIDIA GeForce 840M with display memory: 6073 MB dedicated memory: 2010 MB shared memory 4063 MB.

Initially, to implementation method begin by choose an image to be applied to the algorithm, OpenCV operations are performed on the image mentioned in the third chapter, where an image is saved for each operation, the optical text recognition process is performed on the processed images, the text is extracted and the texts are then compared That was extracted with the entered text to calculate the similarity rate, the highest similarity rate is selected, then the image with the highest similarity rate is selected and then the steps are applied again until the highest similarity rate is obtained.

Figure 4.1 shows a sample of images that were processed test the proposed method.

Figure 4.1: Image selected to apply methodology

Images below figure 4.2 were obtained for the first phase after the image processing

Figure 4.2: Images after processing for selected Image

The actual text is entered to compare discovered texts and calculate the similarity rate between them and the expected text

```
#17
humen = """أشتهي الأشياء التي أَعْلَمْ أنها
ستدمرني في النهاية.
"""
```

Figure 4.3: The actual text

4.2.2. Results

To measure, the accuracy and quality of the algorithm has been applied to a set of 20 images that differ in general features, content and characteristics, in order to ensure that all image processing operations are carried out. After selecting the images and preparing them for image processing operations, and then performing the optical detection of texts.

After that, similarity rates are recorded for each image according to the process that was applied, processing images and the transaction discovery process were saved for each process to compare them with the actual text.

Table 4.1 shows the results obtained from comparison processes for the selected image after processing and process repeated until the highest rate

	Stage 1	Stage 2	Stage 3
original	0.69903	0.84127	0.90598
resize	0.28261	0.59184	0.79310
gray	0.06250	0.84127	0.90598
noise	0.06250	0.84127	0.90598
threshold	0.21622	0.85484	0.83761
invert thresholding	0.57143	0.66019	0.14085
invert gray	0.06250	0.11111	0.65347
dilate	0.21622	0.85484	0.83761
canny	0.06250	0.17949	0.47423
invert	0.69903	0.11111	0.65347
blur	0.05882	0.88889	0.80000
linear	0.06250	0.84800	0.14754
enhance	0.06250	0.84800	0.88889
brightness	0.06250	0.84127	0.49573
colored	0.70588	0.84127	0.80000
contrasted	0.11765	0.84800	0.84211
sharped	0.06250	0.84800	0.86207
remove gray	0.84800	0.66019	0.19178
Large Similarity Rate	0.84800	0.88889	0.90598

Operation	Remove gray	blur	original

Table 4.1: Results for similarity category

The algorithm can applied multi time in same for image, in this image figure 4.1 the original image when applied ORC and get result with (0.69) after apply the algorithm get result (0.84) by remove gray method this for stage 1.

Stage 2 begin by put the image of stage 1 as the original image to stage after apply ORC to the image the results that get is (0.88) by blur image.

Stage 3 begin by put the image of stage 2 as the original image to stage after apply ORC to the image the results that get is (0.90) from original image.

The algorithm will stop if best similarity rate has occurs and if there low similarity rate in stage 2, it varies from image to image depending on the quality of images , it can by one stage or two stage or many stage to get best similarity rate.

Figure 4.4: Image with low quality

In figure 4.5 that shown in low quality after applied the ORC, the result of original image is (0.57) and after applied the algorithm is (0.92) by apply brightness to image, this done in one stage and the algorithm ended when low similarity rate occurred.

Figure 4.5: Image with very low quality

In figure 4.6 that shown in very low quality after applied the ORC, the result of original image is (0.56) and after applied the algorithm is (0.56) this done in first stage and the algorithm ended when same similarity rate occurred.

4.3. Discussion

Using a set of 20 images with different characteristics and specifications to test and evaluate the algorithm, in order to obtain correct and more accurate results, the algorithm was applied. The framework built using the Python language and running in a Windows operating system environment with specifications (Intel processor core i7 4 CPUs, 8GB RAM with 2 graphic cards embedded graphic card and integrated graphic card).

From Table 4.1 the similarity rate of original image in Stage 1 was (0.69) after apply many image processing we get new similarity rate (0.85) from (remove gray) operation, Next Stage begin by set (remove gray) image Stage 2 (0.89) from (blur) image and last stage is same image for beginning of stage 3 we get (0.91).

From the results, it is clear that the best results can be obtained after improving the images, and using tribal processing for the most important stage in the process of visual discovery of texts. In letter patterns not only, but the letters are also attached, and this increases the complexity, the algorithm succeeded in increasing the efficiency of the readings and results.

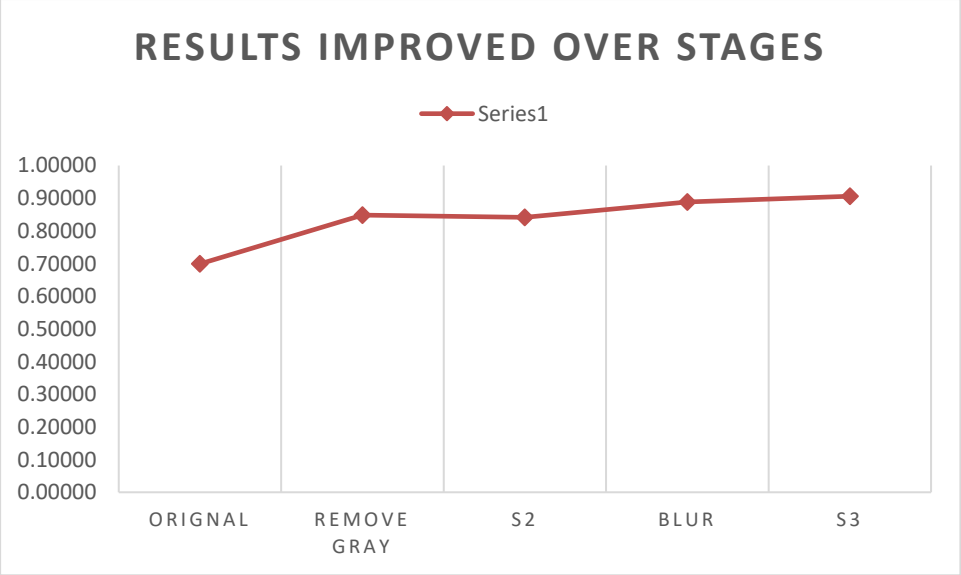


Figure 4.4: Results improved over stages

4.4. Summary

In this chapter, by apply empirical implementation to test the algorithm , directly and actual application of the proposed algorithm and methodology to improve the accuracy of optical character recognition of Arabic letters and the results were obtained and were compared and described in a graph.

CHAPTER FIVE

CONCLUSIONS & RECOMMENDATIONS

5. CHAPTER V

5.1. Conclusion

The Arabic language doesn't found care from none Arabic spoken developers. On the other hand Arabic developers build and create a few tools, frameworks, or methods that helping others none Arabic spoke. This research has focused on design an ORC framework for the Arabic language. The researcher started with the collection of related images and then went through the process of designing the framework by designing an algorithm, and then tested the images collected. And evaluating the produced results by similarity rate.

Through the journey of conducting this research, there have been a lot of obstacles to overcome, especially the two major steps (building a framework and testing the images). Manual getting expects from images for testing and gets the best results from images that were collected, testing the framework also took a long time.

OCR systems accuracy is very high in ideal conditions, but the error rate increases when the images contain noise, the scanning resolution is low, and a cursive written typed languages. This research presents a method which can restore characters' quality in weak resolution images before passing them to OCR engines. The method excludes traditional alignment among resulting texts used by related methods; and also no training is needed on errors before executing it. Furthermore, it performs a procedure to select the best among the output texts and correct wrong words if they occurred in the output texts. In addition to that, this method can be used for any language with simple modification. The experiment results show that this method will improve OCR of output text considerably.

The results, after applying the algorithm to a group of 20 images, showed the average similarity rate of the texts detected to the original images without any modification is (0.50), and the similarity rate of the texts detected for the images that have been improved and repeated image processing operations is (0.91).

5.2. Recommendations

According to the studies Arabic language is one of the languages that is developing rapidly and increasingly in recent times, so it must be taken a care by Arabic developers they must do their best to provide simple and helpful methods using technical and digital means and tools with the ability to accelerate the understanding of the language.

It worth mentioning that Optical character recognition did not find enough interest from Arabic developers therefore this domain could be consider as a hot area of research. Of course, the Arabic language will not be interested for those who do not understand it or know its rules.

The researcher recommends collecting images from many sources and designing more cases in order to get best results and to test power and strength so as to improve algorithm that used on framework.

5.3. Framework improvement

To further improve the performance of the designed framework, the following recommendation were suggested:

- Use a new version of Tesseract and OpenCV.
- Add new Functions and techniques to improve images quality and enhancements speed of OCR and detection.
- Use AI techniques for improve quality of images automatically.

5.4. References

Ahmed, Pervez and Yousef Al-Ohali. 2000. “Arabic Character Recognition: Progress and Challenges.” *Journal of King Saud University - Computer and Information Sciences* 12:85–116.

Alhomed, Lutfieh S. and Kamal M. Jambi. 2018. “A Survey on the Existing Arabic Optical Character Recognition and Future Trends.” 7(3):78–88.

Aliwy, Ahmed Hussain and Basheer Al-sadawi. 2021. “Corpus-Based Technique for Improving Arabic OCR System.” 21(1):233–41.

Covington, Michael A. 2009. “Overview of Image Processing.” *Digital SLR Astrophotography* (December):145–64.

Francisca O Nwokoma, Juliet N Odii, Ikechukwu I Ayogu, and James C Ogbonna. 2021. “Camera-Based OCR Scene Text Detection Issues: A Review.” *World Journal of Advanced Research and Reviews* 12(3):484–89.

Geeksforgeeks.org. 2021. “Bitwise Operations on Binary Images.” Retrieved (https://www.geeksforgeeks.org/arithmetic-operations-on-images-using-opencv-set-2-bitwise-operations-on-binary-images/).

George, Alphy and S. John Livingston. 2013. “A SURVEY ON FULL REFERENCE IMAGE QUALITY ASSESSMENT ALGORITHMS.” 303–7.

H.Gomaa, Wael and Aly A. Fahmy. 2013. “A Survey of Text Similarity Approaches.” *International Journal of Computer Applications* 68(13):13–18.

Habeeb, Imad Qasim, Shahrul Azmi, Mohd Yusof, Faudziah B. Ahmad, and Iraqi Commission. 2014. “Improving Optical Character Recognition Process for Low Resolution Images.” 6(3):13–21.

Hamad, Karez and Mehmet Kaya. 2016. “A Detailed Analysis of Optical Character Recognition Technology.” *International Journal of Applied Mathematics, Electronics and Computers* 4(Special Issue-1):244–244.

Islam, Noman, Zeeshan Islam, and Nazia Noor. 2017. “A Survey on Optical Character Recognition System.” *ArXiv* 10(December):1–4.

Junyong You, Andrew Perkis, Moncef Gabbouj. 2010. “IMPROVING IMAGE QUALITY ASSESSMENT WITH MODELING VISUAL ATTENTION 1 Centre for Quantifiable Quality of Service in Communication Systems (Q2S) , Norwegian University of Science and Technology , Trondheim , Norway ; 2 Tampere University of Technology , Ta.” *Science* 177–82.

Li, Jun, Pengwei Hao, and Chao Zhang. 2008. “Transferring Colours to Grayscale Images by Locally Linear Embedding.” in *BMVC 2008 - Proceedings of the British Machine Vision Conference 2008*.

Li, Yu, Furong Zou, Suiping Yang, Hanhan Liu, Yao Ding, and Kai Zhu. 2020. “Research on Improving OCR Recognition Based on Bending Correction.” 2020:833–37.

Modi, Hiral and M. C. 2017. “A Review on Optical Character Recognition Techniques.” *International Journal of Computer Applications* 160(6):20–24.

Nair, Akhil. 2017. “Overview of Tesseract OCR Engine An Overview of Tesseract OCR Engine Seminar Report Akhil S B130625CS Department of Computer Science and Engineering National Institute of Technology , Calicut Monsoon-2016.” (December 2016).

Naveenkumar, M. and V. Ayyasamy. 2016. “OpenCV for Computer Vision Applications.” *Proceedings of National Conference on Big Data and Cloud Computing (NCBDC’15)* (March 2015):52–56.

Opencv.org. 2021. “Image Blurring.” Retrieved (https://docs.opencv.org/4.5.2/d4/d13/tutorial_py_filtering.html).

Opencv.org. 2022a. “Color Conversions.” *OpenCV*. Retrieved (https://docs.opencv.org/3.4/de/d25/imgproc_color_conversions.html#color_convert_rgb_gray).

Opencv.org. 2022b. “Feature Detection.” *OpenCV*. Retrieved (https://docs.opencv.org/3.4/dd/d1a/group__imgproc__feature.html#ga04723e007ed888ddf11d9ba04e2232de).

Opencv.org. 2022c. “Miscellaneous Image Transformations.” Retrieved

(https://docs.opencv.org/3.4/d7/d1b/group__imgproc__misc.html#gae8a4a146d1ca78c626a53577199e9c57).

Osman, Hussein, Karim Zaghw, Mostafa Hazem, and Seifeldin Elsehely. 2020. “An Efficient Language-Independent Multi-Font OCR for Arabic Script.” 57–71.

P.Sathya, M.Bhuvaneswari, G.Kesavaraj, K. S. Saravana. 2016. “Research in Computer Applications and Robotics a Survey on Trust Based.” 4(4):55–58.

Prasetya, Didik Dwi, Aji Prasetya Wibawa, and Tsukasa Hirashima. 2018. “The Performance of Text Similarity Algorithms.” *International Journal of Advances in Intelligent Informatics* 4(1):63–69.

Procházka, S. 2006. “Arabic.” *Encyclopedia of Language & Linguistics* (1989):423–31.

R.Kiran, Jambekar. 2015. “A Review of Optical Character Recognition System for Recognition of Printed Text.” *IOSR Journal of Computer Engineering (IOSR-JCE)* 17(3):28–33.

Rawat, Sukhbindra Singh, Ashutosh Sharma, and Rachana Gusain. 2021. “ANALYSIS OF IMAGE PREPROCESSING TECHNIQUES TO IMPROVE OCR OF GARHWALI TEXT OBTAINED USING THE HINDI TESSERACT MODEL.” 2588–94.

Sarika Pansare, Dhanshree Joshi and Sinhgad. 2017. “A Survey on Optical Character Recognition System.” *ArXiv* 3(12):2012–14.

Smith, Ray. 2005. “An Overview of the Tesseract OCR Engine.”

Steven W. Smith, Ph. D. 2021. “Linear Image Processing.” Retrieved (https://www.dspguide.com/ch24.htm).

Teahan, William J. 2016. “Arabic OCR Evaluation Tool.”

Thung, Kim-han. 2009. “A Survey of Image Quality Measures.”

Tomar, Swati and Amit Kishore. 2018. “A REVIEW: OPTICAL CHARACTER RECOGNITION.” *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY* 7(7):233–38.

Ugale, Mahendra K. and Madhuri S. Joshi. 2017. “Improving Optical Character Recognition for Low Resolution Images.” *IJCSN International Journal of Computer Science and Network* 6(25):18–20.

Vu, Thang, Cao V Nguyen, Trung X. Pham, Tung M. Luu, and Chang D. Yoo. 2018. “Fast and Efficient Image Quality Enhancement via Desubpixel Convolutional Neural Networks.”

Wei, Yifei, Zhenhong Jia, Jie Yang, and Nikola K. Kasabov. 2021. “High-Brightness Image Enhancement Algorithm.” *Applied Sciences (Switzerland)* 11(23):1–18.

Yadav, Divakar, Sonia Sánchez-Cuadrado, and Jorge Morato. 2013. “Optical Character Recognition for Hindi Language Using a Neural-Network Approach.” *Journal of Information Processing Systems* 9(1):117–40.

Žeger, Ivana, Sonja Grgic Member, Josip Vuković Member, and Gordan Šišul. 2021. “Grayscale Image Colorization Methods : Overview and Evaluation.” XX.