Republic of Iraq
Ministry of Higher Education
and Scientific Research
University of Babylon
College of Science for Women
Department of Computer

Developing OCR and QR Code Based System for Organizational Archive Management

A Research

Submitted to the Council of College of Science for Women-University of Babylon in Partial Fulfillment of the Requirements for the Degree of Higher Diploma in Computer Science

By

Rosa Kahtan Ibrahim

University of Babylon

Supervised by

Dr. Mahdi Abed Salman

2021 A.D. 1443 A.H.


بسم الله الرحمن الرحيم
قَالُواْ سُبْحَانَكَ لاَ عِلْمَ لَنَا إِلاَّ مَا عَلَّمْتَنَا
إِنَّكَ أَنتَ الْعَلِيمُ الْحَكِيمُ
صَدَقَ اللهُ العَلِيُّ العَظيم
سورة البقرة(٣٢)

Supervisor Certification

I certify that the project entitled " **Developing OCR and QR Code Based System for Organizational Archive Management** " was prepared at the Department of computer Sciences/ College of Science / University of Babylon, by **Rosa Kahtan Ibrahim** as partial fulfillment of the requirements for the degree of higher diploma in Computer Science for Women.

Signature:
Name: Dr. Mahdi Abed Salman
Date: / /

The Head of the Department Certification

In view of the available recommendations, I forward the research entitled "**Developing OCR and QR Code Based System for Organizational Archive Management**" for debate by the examination committee.

Signature:

Name : Dr. Farah Mohammed Al-Shrafy

Date: / / 2022

Address: Head of Computer Science Department / College of Science for Women

Certification of the examination Committee

We are the member of the examine committee, certify that we have read this project entitled (**Developing OCR and QR Code Based System for Organizational Archive Management**) presented by the student (**Rosa Kahtan Ibrahim**) in its content at 27/12/2021, and that in our opinion it is accepted as a project for the degree of higher diploma in science\computer science with a degree (very good).

Committee Chairman:

Signature:

Name: Ahmed Saleem Abbas

Scientific order: Assist. Prof. PhD

Date: / 1 /2022

Committee Member:

Signature:

Name: Muhammed Abaid Mahdi

Scientific order: Assist. Prof. PhD

Date: / 1 /2022

Committee Member (Supervisor):

Signature:

Name: Mahdi Abed Salman

Scientific order: Assist. Prof. PhD

Date: / 1 /2022

Deanship authentication of college of science for women.

Approved for the college committee of grade studies.

Signature:

Name: Faez Ali Rashid

Scientific order: Prof. PhD

Address: Dean of College Science for Women

Date: / 1 /2022

Acknowledgments

First and foremost, I am grateful to Allah, our Lord and Cherisher, for His Inspiration, and Guidance throughout all my life and study. Peace and favors descend in perpetuity on our beloved Prophet Muhammad al-Mustafa and Ahlul-Bait, whose are mercies for all the worlds.

Foremost, I would like to express my sincere gratitude to my advisor **Dr. Mahdi Abed Salman** for his continuous support in my project.

Last but not the least; I would like to thank my family, my husband & my daughters. This project not have been complete without your unwavering support and their prayers and well wishes.

Abstract

Despite the development of information technology systems and the way in which official documents in paper format are produced, it is still difficult to organize and archive these documents on the computer, as most of the current archiving systems load the scanned document as images in a database that provides some details that are set manually by the responsible user. However, these documents may not have the ability to search in their contents and this is what we see clearly in the scanned documents. Here comes the role of OCR, which has effectively contributed to the development of the document management process by converting images into text files that enable the user to search within their contents. OCR performs an image analysis process to read text from images based on character classification on A set of libraries that can be included in most programming languages. This is what was used in this project. we will read the texts from the images and then store it in the database so that the content of the book is searchable within it. In order to develop this process, we will convert the documents content into a QR code, which is an encoding technique used to represent the text as a code that is easy to read by code scanners, which facilitates the process of sending content and reducing the size of the data sent.

List of Contents

List of Tables

List of Figures

LIST OF ABBREVIATIONS

Abbreviation	Full statement
OCR	Optical character recognition
QR	Quick response code
NLP	Neuro-linguistic programming
RDBMS	Relational data base management system
DPI	Dots per inches

Chapter 1 Introduction

==========

1.1 Introduction

The spotlight on converting images to texts began in 1950. The implementation of such process revolutionizing the document and document management process. Optical character recognition of scanned documents has made it more than just images, which in turn is reflected in the ability to make the content of documents searchable as programs and applications recognize it. [1]

OCR extracts the text and sends it automatically to the database instead of the traditional method that requires rewriting the text manually. OCR classified as a broad field under which a variety of applications falls under it, such as bill imaging, banking services, etc.[1]

OCR is also used to secure sites as we see in the Captcha, in addition, it is used in digital libraries of educational institutions, smart warehouses and documenting car violations through automatic recognition of the license plate numbers of the violating cars.

An additional technology is used in this project in order to improve performance, and this technology is a QR code, which is a code that a cell phone or any reader can easily read it. A specific amount of information is sent through it and when a QR code is scanned using a spacing combination as a barcode matrix (a two-dimensional bar code) the code is decoded and its content displayed using any reader or mobile device.

==========

1.2 Problem statement

Finding data in printed documents and entering it into the database and issuing data in an encoded form from the database to printed texts to facilitate its retrieval and search for its details.

1.3 Project Aim

This project aims to develop a system that has the ability of converting an Image to Text document by using OCR, which will enable us to scan documents, store them in both image and text format and retrieve them using text match searching and QR scan.

The system will be used in managing the archive of document process which will contain four main steps:

1- Uploading the document and feed the system with the necessary data to be classified later.
2- Saving the content of the document and all the other data to a database.
3- Generating a QR code for each document and save it with the related record to be easy retrieved later
4- Retrieving data from the database by using words or QR code.

All the users will have specific authorization to do their tasks with the system depending on its role. This work proposes that each QR code will contain information of a single document.

==========

1.4 Related works

The related works included in this section is related to OCR and QR encoding in different platforms

In reference [2] , Proposed a protection method against phishing attacks. Using the technology of optical character recognition (OCR), by reading the banners on the website and comparing them to the website's URL, the attach can be distinguished. A prototype showed a high detection accuracy in the initial experiments.

In reference [3] , the authors explained how IoT, OCR, and blockchain technologies can be used to create a wine tracking system in a real-world environment. The research focuses on the digital transformation of the traditional wine supply chain after reading existing serial numbers that have been tagged on bottles, to uniquely identify wine bottles and track the life cycle of an item.

In reference [4], building a project to use QR code labels as a tag to guide the preprocessing steps of the captured image, then traditional OCR is used to extract the data in the image. From that data, the system can decide whether the capture image is tilted or not. This work shows that with this technology it is possible to process the image and achieve higher resolution.

In reference [5], they propose two visual challenges with enough randomness to deny any replay attacks: Plain text is recognized via (OCR) and QR codes. In both cases, the validator generates a random string of inputs and it is displayed on the screen as a plain text image or as a QR

code after the random string is encoded. To extract the random string from the scanned image, the receiver uses an OCR or QR code then check if the string in the scanned image is the same (or close enough in the case of OCR) as the one sent as a challenge.

In reference [6], they aimed to build an automated course grading system for students with OCR to fully automate the course registration process. The developed system is able to generate a unique QR code, which contains student information such as name, ID number and course title. generated The QR code by the system is printed along with the student's information and the user's schedule to fill in the course mark manually.

In reference [7], several algorithms were used for pre-processing, which includes image modification, document and table layout analysis, in order to improve the accuracy of OCR. The work of the algorithms verified using the famous OCR program Tesseract. The researcher concluded that the used ways can process document images accurately with different layouts and arbitrary angles of rotation.

In reference [8], the researchers seek to automate the classification of records using the open source library (PyTesseract), the Google Tesseract-OCR engine shell. Where documents are first converted to a digital view (scan) and then the text is recognized and extracted using the PyTesseract library. With the help of this system, keeping and retrieval of records in this way saves the trouble and effort on the records officer.

In reference [9], the authors have found a new way to recognize Arabic calligraphy and Persian calligraphy based on transforming the resizing

feature, and what distinguishes this system from other systems is that it does not need to remove noise or any pre-processing except for the low-quality image. The result was impressive, after testing they found that 1,400 text images give almost 100% discrimination.

In reference [10], the researchers designed and implemented a new system that recognizes Arabic letters without pre-segmentation based on character descriptions in terms of shape substitutions and the use of mathematical morphology, and the results showed an accuracy rate of 99.4% for noise-free text and 73% for scanned text.

In reference [11], the researchers tried to deal with old documents that already contain noise and studied them in order to improve them. The study found that the best way to improve a document is to change the brightness of the scans.

==========

1.5 Document organization

The rest chapters will include:

Chapter two: State of art: a details about the problem and what is the existing tools that solve the problem with the referring to the advantage and disadvantage.

Chapter three: solution methodology (proposed system): here the researcher introduces the proposed system in details with diagrams and any flowcharts which can help in understanding the idea and what tools used in developing the system.

Chapter four: results: this chapter includes all the Screen shots of the system with a Performance evaluation for it.

Chapter five: conclusions and perspectives: presenting the conclusion and the future work

2.1 Introduction

Computers play an important role in automation of various process and industries, and electronic archiving is one of them. The improvement of work inside offices is the goal of all the government and private institutions. For that, many developers involved in developing some software's which can help in that improvement. The problem is that during the day, many documents and numerous letters are received and generated in offices then stored in hard files and folders in offices. When we want to search inside that files and folders it takes a lot of time and the process can be worse when the employee doesn't remember the date or title of this document / letter but just the source name or some of the content lines of the document. Moreover, in some cases the misplacement of these documents can happen too.

Regarding the missioned problem, it will be efficient that we can save the image as text to make the content searchable, in order to overcome all the previous problems. What this will achieve is to allow an efficient and easy search of documents by just typing source name, title or any other word.

2.2 What is OCR?

Optical Character Recognition or OCR is a technology that enables scanned documents to be more than just images; it can transform into completely searchable documents with textual content recognized by any reader. OCR extracts letters and words and enters it into the database instead of the traditional method of manually retyping text [12].

Many applications fall under the umbrella of optical character recognition, such as invoice imaging applications, legal systems, banking services, and others. OCR is also used widely in many different fields such as Captcha, institutional repositories and digital libraries, automatic number plate recognition and handwriting recognition. OCR used several Phases to make the scanned image readable as explained in table (2.1) [12][13]

Table 2.1: Phases of OCR [13]

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

In the fifth phase (Classification), there are number of approaches can be used as it is listed below:

1- Matrix Matching
2- Fuzzy Logic
3- Feature Extraction
4- Structural Analysis
5- Neural Networks

2.3 What is QR Code

QR Code is a brand of matrix-type barcode, produced by the Japanese company **Denso Wave**. A QR code has a set of characteristics that include data encryption in acceptable quantities, relative damage resistance, high-speed reading, small space, and the ability to read from different angles.[3][12]

QR code has become the major code in many fields due to the many advantages of it like increased capacity, reduced size and there is versatility and the possibility of future development. In addition to all what has been mentioned, a QR code can encrypt the same amount of data that a barcode encodes do but with an area of 1/10 of the space of a traditional barcode. Information such as a website link, text messages and contact information as well as plain text can also be included in QR codes.[15]

2.4 Architecture and Coding

A QR code is a two-dimensional code look like a matrix symbol that is engineered to appear as a cell arranged in a square shape. Figure 2.1 shows the structure of the QR code. QR codes have a different part that are reserved for specific purposes.

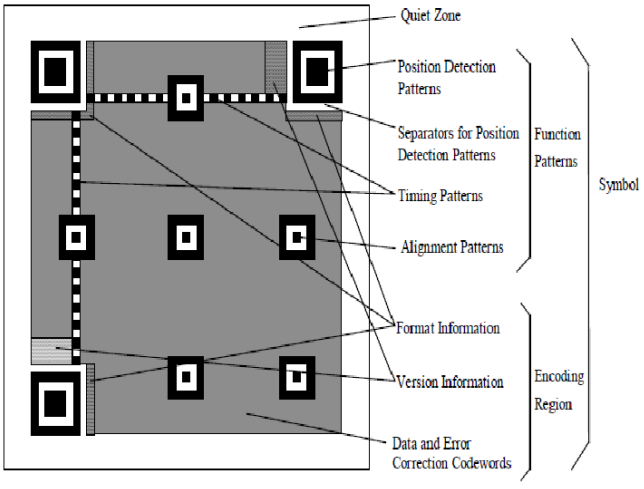

Figure (2.1): QR code architecture [15]

The finder, interval, timing patterns, and alignment patterns consist of job patterns. Function patterns may not be used to encrypt data. The search tool patterns in the

three corners of the icon are meant to help make it easier to determine its position, size, and inclination. The QR Code encoding procedure includes the following steps. First, the input data is encoded according to the most efficient method and modulated bit streams. Bitstreams are broken down into encrypted words.[15][16]

Then the ciphers are broken into blocks, and error-correcting ciphers are added to each block. All these code words are placed in an array and masked with a mask pattern. Finally, the job styles are added to the QR code. QR code has been made

2.5 Existing Tools

- Tesseract: is a tool that used by most of the OCR applications. It represents the optical character recognition engine for those applications and can be used in various operating systems. In addition, we most mention that it is free and developed by HP.
- ABBYY FineReader PDF: is an (OCR) application developed by ABBYY, and it runs under Windows 7 or later, and mac OS 10.12 Sierra or later.
- e-Aksharayan: is an OCR engine for Indian languages. Some of research work from e-Aksharayanhas have been published in different conferences and journals.
- AspriseOCR: is a commercial (OCR) and also a library for barcode recognition which has the ability to recognize text and barcodes from scanned documents and output in text formats.

- AnyDoc Software: is a company that developing, soling, installing, and supporting enterprise content management software which extract data from scanned documents and convert it readable text for office applications and content/document management systems.
- CuneiForm: is a Cognitive Open OCR system, developed by Russian Software Company.
- Dynamsoft Corp: is a multinational software development company. It provides a SDK for the text capture and barcode.
- OmniPage: is an (OCR) application available from Kofax Incorporated.
- Microsoft OneNote: is a program designed by Microsoft for the purpose of writing notes and collecting information such as graphics, screenshots and audio comments through collaboration among a group of users.
- GOCR (or JOCR): is a free OCR program. Used to convert scanned image files into text files.
- Ocrad: is a free software licensed under the GNU GPL that is used for optical character recognition.

2.6 Advantage and disadvantage of OCR

2.6.1 Advantages (OCR):

- It reads scanned images or documents with a high degree of accuracy. Part of the credit for this accuracy is due to modern scanners, which possess a large degree of accuracy and clarity.
- Speeding in data processing, which gives OCR the advantage if it is compared with manual entry methods.

2.6.2 Disadvantages (OCR):

- Problems often arise with handwritten documents due to the difficulty of OCR to extract letters.
- There is no OCR program that works with 100% quality because it is difficult to recognize letters in some cases, especially in low quality documents.

2.7 Summary

OCR is a technique for extracting texts from images and pdf files that is of great importance in improving searches within organizations by storing the extracted texts in databases. There are many tools that do this work, which were referred to above. This technology has a number of benefits, the most important of which is the possibility of making the extracted data readable. There are also some limitations and problems. For example, the accuracy of the extracted texts is affected by the quality of the image and not efficient with images whose content is handwritten. User intervention is always required for having accurate results.

We also want to mention that all the researchs we reviewed used document written using English language where our research use both Arabic and English. Farther more most the researches we mentioned in the related works implemented on labels or specific strings while we implemented directly to a real life organizational scanned document. For that, the difference I accuracy will be clear.

3.1 Introduction

Optical character recognition (OCR) distinguishes the printed or handwritten text characters inside digital images that have to scan by optical scanning devices and kind of software. By OCR, software allows a computer to read Scanned images of text and make it editable and searchable data. In addition, we use QR code which will to enhance the retrieving of information from the system.

3.2 Proposed project

The proposed project has three main functions:

- Extract Data, Generate QR Code and save all the used and generated data in the database.
- View General Reports, which contain all the data that had been saved in the previous function.
- Searching for Specific Entry by writing a any word in the search bar, and the system will display the data if there is any match with contents words

In addition, the proposed system able to scan not only images but also PDF files with taking care to the content of the files. For example, some file contains both English and Arabic words. In this case, we implement Arabic OCR as primary OCR and English OCR as secondary OCR.

3.3 Block diagram

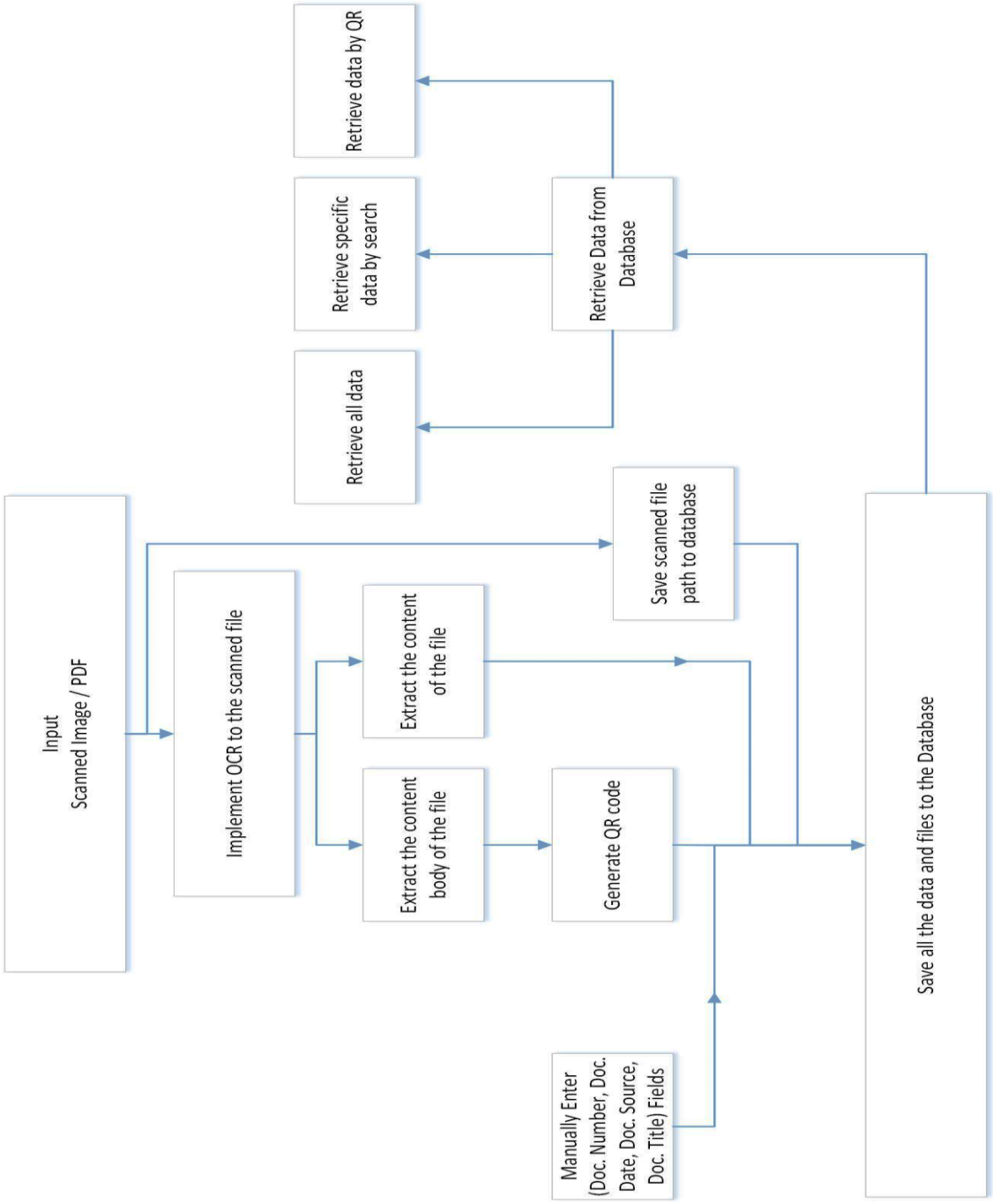

Figure (3.1): the block diagram of the proposed system

3.4 Role and expected result of each block

As it is mentioned before there are three main function but the first one contains a set of roles that will illustrated below

a) Manually Enter (Doc. Number, Doc. Date, Doc. Source, Doc. Title) Fields: these fields usually written by hand for that the OCR in most times not read it correctly so that it is preferred to enter it manually.

b) Input Scanned Image / PDF: First, the researcher start process by input a scanned document with image or PDF extension.

c) OCR the body of the file: At the same time, where the document is saved, the Optical character recognition is implemented on the document to extract the text of the body from it.

d) Extract the body content of the file: because of the low capacity of the QR Code, the researcher found that it is not possible to add all the file data in it. For that, the proposed system extracts the body content of the page only (as it appears in figure (3.2)) without the head and tail of the document and convert it to QR code. The extraction of the body content is extracted done by the following code

```
var ContentArea = new System.Drawing.Rectangle() { X = 0, Y = 250,
Height = 500, Width = 1100 };
```

The function *System.Drawing.Rectangle(),* start calculating a Rectangle area with a dimensions that had provided with.

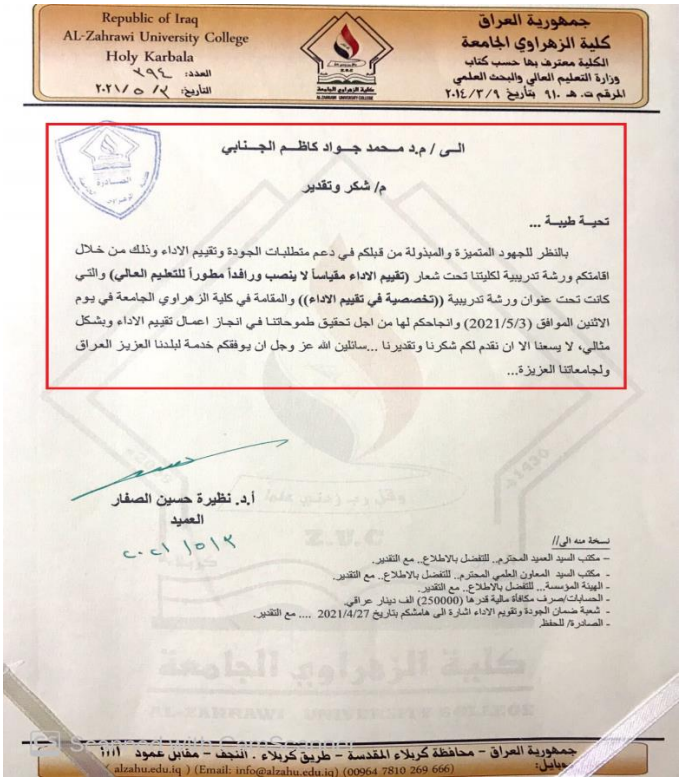

Figure (3.2): The extracted area of the document.

e) QR Generator: By this step, the text is converted, which represents the body of the document to QR code.

f) Saving all data to Database: after completing the scanning converting and manually entering, all the data will be saved in the database.

g) Retrieving Data & Print Out: finally, we will have the ability to retrieve data in so many forms for example:

- Retrieving all the data in the database.
- Retrieving a specific data depending on keywords.

3.5 Implementation and programming

The developed project is using several tools as listed below:

1- Microsoft Visual Studio: is an (IDE) developed by Microsoft company. It is Used to build software, websites, and another services and apps. In this project, VS 2019 is used because of many new features like Easier to Launch Your Code, Simplified UI, A Better Search and the AI-Assisted IntelliCode.

2- ASP.NET Core: is a replacement to ASP.NET, developed by Microsoft. It is free to use in developing .NET applications and the cross-stage .NET. The system is a finished change that joins the beforehand independent ASP.NET MVC and ASP.NET Web API into a solitary programming model. It is used in the project because most of the efficient OCR libraries have developed to run over .net Core 3 and above.

3- Model-View-Controller (MVC): is a designed pattern that separate the developed application into three fundamental parts: model, view, and Controller.

4- C#: is a universal worldview programming language. C# envelops static composing, solid composing, lexically perused, basic, explanatory, utilitarian, conventional, object-arranged (class-based), and part situated programming disciplines

5- HyperText Markup Language (HTML): Is the standard markup language used to build webpages that are displayed in a web browser. The HTML alone is not enough to build the webpage for that we need assisted technologies such

as Cascading Style Sheets (CSS) and client-side scripting languages such as JavaScript

6- Cascading Style Sheets (CSS): used for designing the presentation of a webpage which written using HTML. CSS is used along with HTML in building the page that runs over the net.

7- TESSERACT-OCR: is an OCR engine available for various operating platforms. With other tools like Leptonica Image Processing library, it can read various image formats and convert them into text format Although. It is used because of it is free cost and efficiency.

8- IronOCR: is a C # software library that enables .NET platform software to recognize and read text from scanned documents and PDF documents. It is a pure .NET OCR library uses the most advanced Tesseract engine known.

9- QRCoder: is a library that built with C# which enables developers to create QR codes. Available as a .NET Framework and .NET Core PCL version at Newgate.

10- Microsoft SQL Server: it is RDBMS developed by Microsoft. its main function, is to store data and retrieve it as requested by other users which may run either locally or remotely.

All the mentioned tools are used together to implementing and programming the system.

In the first step, the aim was to get the text from the scanned document and save the text in the database using OCR (Figure (3.3)).

Figure 3.3: The Main idea of the OCR process

The first challenge was the system ability to work over images files and PDF files, and here we imbedded two block of codes, the first one to extracting from Image where the second is for extracting from PDF recognizing the file by extension.

The second challenge in the extraction was how to make the system able to extract Arabic and English words from the same file because most of the OCR libraries support English words or Arabic words and this problem was solved by making the system use two languages, Arabic as primary language and English as secondary one (figure 3.4).

```
var Ocr = new IronTesseract();
Ocr.Language = OcrLanguage.Arabic;
Ocr.AddSecondaryLanguage(OcrLanguage.English);
```

Figure 3.4: Display the code of including extraction Languages

Now is time to generate the QR code, in this step The QrCode() function has two inputs. At first, the string which is needed to be convert to a QR code, and the second input is error correction level (ECCLevel) and here there are four levels, the levels are:

1- L (7%).

2- M (15%).

3- Q (25%).

4- H (30%).

The percentage indicates how much of the code we want to convert can be hidden/destroyed until the error correction algorithm can not recreate the original message that was encoded in the QR code. In other words, when the percentage increased, the accuracy is decreased. The choose of the levels depend on the amount of data needed to be convert to QR Code and there is also a limitation for the data (Table (3.1))

Table 3.1: Specifications of QR Codes

Maximum data capacity	**Numeric**	**7089 characters**
	Alphanumeric	**4296 characters**

After completing the extraction of the code, the extracted text with the manually entered fields are saved in a table which was designed for this purpose (Figure (3.5))

Column Name	Data Type
id	int
DocNo	varchar(50)
DocDate	smalldatetime
DocTitle	nvarchar(50)
DocSource	nvarchar(50)
OCR	nvarchar(MAX)
FilePath	varchar(200)
QR	image

Figure (3.5): Display the design mode

==========

4.1 Introduction

In this chapter, the proposed system results and performance will be discussed. As it was mentioned, two technologies in the project which are OCR and QR.

Both technologies have advantage and drawbacks that listed in chapter 2, but here the discussion will be focusing on the practical that is faced in the implementation of the project

4.2 Experimental Results

The system consists of a set of interfaces with functions that were mentioned in Chapter 3. Figure (4.1) clearly shows the main form of the proposed system which contain the data that required for each document.

Figure (4.1): Illustrate the main page of the System

==========

After filling the field with data and implement, the operations which listed as a buttons in the bottom side of the form, the results data will be displayed (figure (4.2)).

Figure (4.2): Display the main form after execute all the operations

In the first box from the right side in figure (4.2), the scanned file is listed, followed by the extract data of all the file, and last box contain the extracted data of the body content only which will be converted to QR Code.

By any QR code reader, the content of the code can be read (figure (4.3)) which represents the body content of the extracted document

Figure (4.3): Display the content of the QR Code by QR reader.

All the mentioned data, including the path of the document, as well as the QR code, are stored in the database prepared for this purpose in order to retrieve data when needed to facilitate the process of searching, sorting, organizing and extracting data in the form of reports. Figure (4.4) represents one of the reports through which all the documents information is displayed, including the document QR code.

==========

Figure (4.4): General Report

Figure (4.5) represents a report dedicated to display the document and that are being searched by entering word or words from the document content in the search box. All document information whose content or part of its content matches the text entered in the search box are displayed.

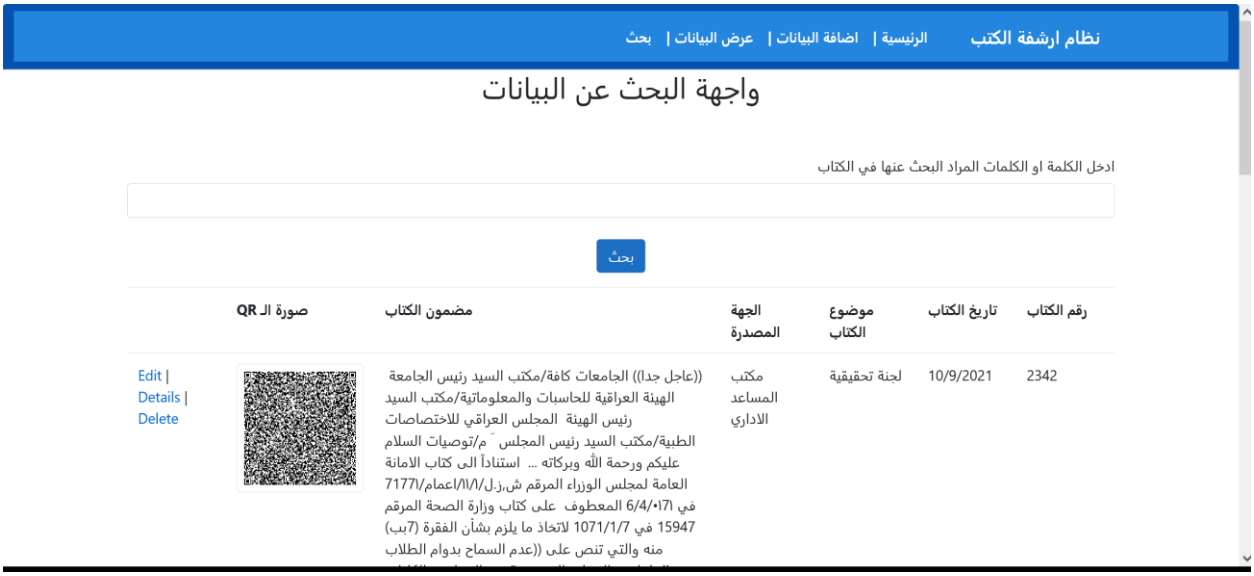

Figure (4.5): Search of specific entries.

==========

4.3 Experiment accuracy

Experiment was done by using 100 document with scanning the same document six times each time with different dpi (100,200,300,400,500,600). In addition, two accuracy factors were used, **Character Accuracy and Word Accuracy.** The result was that the accuracy of the OCR with engine tesseract have a good performance when the dpi is between 200 and 300. Figure (4.6)

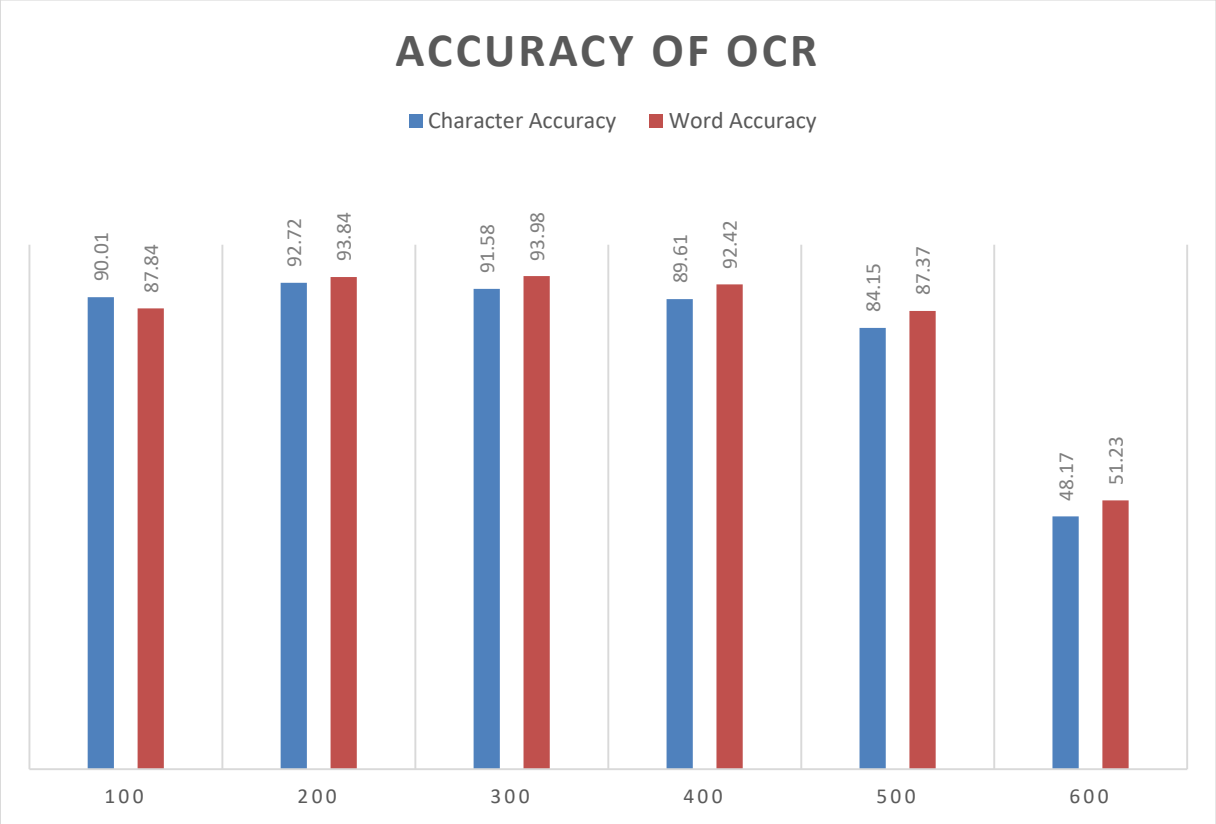

Figure (4.6): The accuracy of OCR with different dpi

The accuracy calculated by counting the number of missed or wrong characters and subtracted it from the number of the character in the original scanned file and the same thing done in the word case. By implementation we find that keeping DPI lower than 200 will give unclear and incomprehensible results while keeping the

DPI above 600 will unnecessarily increase the size of the output file without improving the quality of the file. Thus, a DPI of 300 works best for this purpose.

4.4 Factors that affect the OCR accuracy

In the experiment, the researcher found that there are several factors that affect the accuracy of the OCR accuracy and the factors are:

1- The original document Quality: The document can be damaged, torn, aged, faded, and discolored. Also it was found that ink color is affect the accuracy, for example colors like green, blue, yellow and red have a low contrast where is black ink have a highest contrast. In addition, the human handwriting and using specific types of paper that decrease crispness and contrast between the background and foreground in the resulting scan.

2- The scan quality: In the experiment, used different dpi in scanning the document. found that the better dpi to use OCR scanning is between 200 and 300 and the worst one is 600.

5.1 Conclusions

Optical character Recognition OCR is an image analyzing process to read text from images that is available as embedded libraries for programmers within most programming languages. QR code is an encoding technique used to represent text as a figure which is easy to be read by code scanner devices providing new produced document which contains QR code representing its text content to facilities retrieving process. Most of current archiving systems upload scanned document as images in a database or file system provide some details that are set manually by the responsible user and this process make it difficult to organize and archive such papers on computer storage.

In this project, OCR and QR code techniques are used for making the image with its content searchable and readable by the application. Also we used different tools to develop the application and got good results and solve a number of challenges which includes (extracting text from different types of file, different content language and the capacity of the QR code).

Farther more, we study the resulted accuracy in different dpi's and the best one was 300 where the worst was 600.

5.2 Future Work

Analyzing text using NLP tools to get database fields directly from OCR output and Reprinting QR code on original document for fast retrieval without need to database.

References

==========

[1] D. Berchmans and S. S. Kumar, “Optical character recognition: An overview and an insight,” *2014 Int. Conf. Control. Instrumentation, Commun. Comput. Technol. ICCICCT 2014*, no. June, pp. 1361–1365, 2014, doi: 10.1109/ICCICCT.2014.6993174.

[2] Y. Wang and I. Duncan, “A novel method to prevent phishing by using OCR technology,” *2019 Int. Conf. Cyber Secur. Prot. Digit. Serv. Cyber Secur. 2019*, pp. 1–5, 2019, doi: 10.1109/CyberSecPODS.2019.8885101.

[3] S. Cakic *et al.*, “Digital Transformation and Transparency in Wine Supply Chain Using OCR and DLT,” *2021 25th Int. Conf. Inf. Technol. IT 2021*, no. February, pp. 16–20, 2021, doi: 10.1109/IT51528.2021.9390117.

[4] S. L. Lai, B. Y. Ooi, and Y. L. Chen, “Using QR Code Labels to Enhance OCR for Capturing Legacy Machines’ Data,” *Proc. - 2019 Int. Symp. Intell. Signal Process. Commun. Syst. ISPACS 2019*, pp. 18–19, 2019, doi: 10.1109/ISPACS48206.2019.8986403.

[5] J. Valente and A. A. Cárdenas, “Using visual challenges to verify the integrity of security cameras,” *ACM Int. Conf. Proceeding Ser.*, vol. 7-11-Decem, pp. 141–150, 2015, doi: 10.1145/2818000.2818045.

[6] Z. Saad, M. S. Sulaiman, R. Seman, Z. Hisham, and C. Soh, “Smart Autogate Using Optical Character Recognition (Ocr) and Color Detection,” *PalArch’s J. Archarology Egypt/Egyptology*, vol. 17, no. 10, pp. 946–956, 2020.

[7] A. Phan Việt, P. Viet Anh, N. Duy Tung Khanh, T. Manh Dat, and P. Van Dan, “Improved Ocr Quality for Smart Scanned Document Management System,” *J. Sci. Tech. Quy Don Tech. Univ.*, vol. 210, no. February 2021, 2020, [Online]. Available: https://www.researchgate.net/publication/348959820.

[8] J. M. Jayoma, E. S. Moyon, and E. M. O. Morales, “OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga, Philippines,” *2020 IEEE 12th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag. HNICEM 2020*, 2020, doi: 10.1109/HNICEM51456.2020.9400000.

[9] M. Zahedi and S. Eslami, “Farsi/Arabic optical font recognition using SIFT features,” *Procedia Comput. Sci.*, vol. 3, pp. 1055–1059, 2011, doi: 10.1016/j.procs.2010.12.173.

[10] B. Al-Badr and R. M. Haralick, “Segmentation-free word recognition with application to Arabic,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 1, no. September 1995, pp. 355–359, 1995, doi: 10.1109/ICDAR.1995.599012.

[11] P. Herceg, B. Huyck, C. Johnson, L. Van Guilder, and A. Kundu, “Optimizing OCR accuracy for bi-tonal, noisy scans of degraded Arabic documents,” *Vis. Inf. Process. XIV*, vol. 5817, p. 179, 2005, doi: 10.1117/12.606447.

References

==========

[12] A. Singh, K. Bacchuwar, and A. Bhasin, “A Survey of OCR Applications,” *Int. J. Mach. Learn. Comput.*, vol. 2, no. 3, pp. 314–318, 2012, doi: 10.7763/ijmlc.2012.v2.137.

[13] J. Barwick, “Building an institutional repository at Loughborough University: Some experiences,” *Program*, vol. 41, no. 2, pp. 113–123, 2007, doi: 10.1108/00330330710742890.

[14] A. G. Menon, A. V Mohanan, and G. V. Jaison, “ATM Custodian: A New Type Of Authentication For ATM’s,” *Int. J. Comput. Appl.*, vol. 6, no. 2, pp. 100–106, 2016.

[15] D. Moorty, M. Ruman, N. Irfan, and K. R. Aishwarya, “QR Code Based Text To Speech Conversion,” vol. VI, no. May, pp. 66–69, 2017.

[16] K. H. Pandya and H. J. Galiyawala, “A Survey on QR Codes: in context of Research and Application,” *Int. J. Emerg. Technol. Adv. Eng. Website www.ijetae.com ISO Certif. J.*, vol. 4, no. 3, pp. 258–262, 2014.

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة بابل
كلية العلوم للبنات
قسم علوم الحاسبات

تطوير نظام يعتمد الـOCR وQR Code لادارة الارشفة الالكترونية في المؤسسات

البحث

مقدم إلى مجلس كلية العلوم للبنات ـ جامعة بابل وهي جزء من متطلبات نيل درجة الدبلوم العالي في العلوم/علوم الحاسوب

من قبل الطالبة

روزة قحطان ابراهيم

باشراف

أ.م.د مهدي عبد سلمان

2021 A.D. 1443 A.H.

الملخص

على الرغم من تطور أنظمة تكنولوجيا المعلومات والطريقة التي يتم بها إنتاج الوثائق الرسمية ذات التنسيق الورقي ، إلا أنه لا يزال من الصعب تنظيم هذه الوثائق وأرشفتها على الكمبيوتر، حيث تقوم معظم أنظمة الأرشفة الحالية بتحميل المستند الممسوح ضوئيًا كصور في قاعدة بيانات توفر بعض التفاصيل التي يتم تعيينها يدويًا بواسطة المستخدم المسؤول. ومع ذلك ، قد لا تتوفر في تلك الوثائق القابلية على البحث في محتوياتها وهذا ما نراه واضحا في المستندات الممسوحة ضوئيا. وهنا ياتي دور التعرف الضوئي على الحروف الـ OCR الذي اسهم بشكل فاعل في تطوير عملية ادارة الوثائق من خلال تحويل الصور الى ملفات نصية تمكن المستخدم من البحث داخل محتوياتها حيث يقوم التعرف الضوئي على الحروف OCR بعملية تحليل للصور لقراءة النص من الصور معتمدا في تصنيف الحروف على مجموعة من المكتبات التي يمكن تضمينها في معظم لغات البرمجة. وهذا ما تم استخدامه في المشروع ، حيث سنقوم بقراءة النصوص من الصور ومن ثم تخزينها في قاعدة البيانات ليصبح محتوى الكتاب قابل للبحث داخله. ومن اجل تطوير هذه العملية تمت اضافة امكانية تحويل محتوى الكتاب الى رمز الاستجابة السريعة QR والذي هو عبارة عن تقنية ترميز تُستخدم لتمثيل النص كرمز يسهل قراءته بواسطة أجهزة الماسح الضوئي للرموز والذي من شانه تسهل عملية ارسال المحتوى وتقليص حجم البيانات المرسلة.