ISSN 2321 – 2004

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 1, Issue 2, May 2013

Digital Document Archiving System with Optical Character Recognition

Simmi Dutta[1], Shruti[2], Haneet Kour[3], Chandani Bhagat[4]

Asstt. Prof., Computer Engg. Department, Govt. College of Engg. & Technology, Jammu, India[1]

Final year student, Computer Engg. Department, Govt. College of Engg. & Technology, Jammu, India [2]

Final year student, Computer Engg. Department, Govt. College of Engg. & Technology, Jammu, India [3]

Final year student, Computer Engg. Department, Govt. College of Engg. & Technology, Jammu, India [4]

Abstract : Computers are playing an important role in automation of various process and industries and Digital Data Archiving is one of them where in we tend to improve the working of office through some software process. Each Day numerous letters, visting cards and documents are received and generated in offices and then they are stored in files and folders in offices. To search any document takes a lot of time andthe things go more worse when we don’t remember the date or heading of this document / letter but just the sender’s name or just a line from the text of the document.. There is also chances of misplacement of these documents. So it is the need of the hour to built an application that will have capabilities to scan documents and store them in image format, extract the text out these images and store that text in database. What this will achieve is to allow an efficient and easy search of documents by just typing sender’s name or the company name.

Keywords : Optical character recognition(OCR) , Digital archive , Scanning , emguCV , WIA lib (window image acquisition library), MODI lib (microsoft office document imaging library), tesseract-ocr.

I. INTRODUCTION

This paper aims to exhibit our project that has been developed for such an Image to Text Conversion Application which will enable us to scan documents, store them in both image and text format and retrieve them using search just to reduce paper load in our offices.This project will be very helpful in business environment. In this work , we are going to make use of OCR technology. OCR is an important evolving technology that can help us in playing an important role in automation of various process. There is no doubt that with the help of OCR ,we can convert any image or scanned document into text form. Infact because of practical importance of OCR applications ,there is a lot of great research interest and tangible advances in this field[1].OCR processes any image into text in five steps – Preprocessing , Segmentation , Feature extraction, Classification and Postprocessing[2]. We can also save these resultant images. But if we have to see a particular visiting card or document and we only remember the heading or a specific line; to search that card or document we have to look at all visiting cards in our files or scanned documents . It causes a lot of wastage of time and is a very complicated process. The OCR can only convert an image into text format .There is a need for developing a technique for effective digital archiving implementation. So in this paper we present an application which has capabilities to scan documents and store them in image format, extract the text out these images and store that text in database. This will enable one to search documents easily by just typing sender’s name or the company name. This is a database application implemented with the help of OCR technology.

II. AIM AND OBJECTIVES

This paper presents a program that aims to recognize the text contained in images(image to text conversion by OCR); stores this text into respective attributes of database(which we have created in MS SQL Server).It also stores the images in database.

To search any document or visiting card, the user gives any input keyword(i.e. heading of document or any specific line) then the program performs search accordingly and displays the result.It can display result in both image format as well as in text format depending upon the choice of the user.This work aims to implement a scanning and converting Software which will significantly reduce the documents in the offices and convert it to a office with-out pile of files.Thus we can easily create digital archives[3]. In our work, we have used following libraries (of vb.net[4]):

a) EmguCV - for image processing. Emgu CV is a cross platform .Net wrapper to the OpenCV image processing library.allowing OpenCV functions to be called from .NET compatible languages. The wrapper can be compiled in Mono and run on windows , Linux Mac X OS , iphone, ipad and android devices[5].

b) WIA – for scanning . The WIA provides a platform which enables graphics applications to interact with imaging hardware and standardizes the interaction between various applications and scanners.It provides a platform for more robust , stable and reliable scanning experience by isolating the driver and the application [6].

c) TESSERACT-OCR– it is the most accurate free software ocr engine available for various operating system (Linux, Window, Mac OS X). When combined with Leptonica Image Processing library, it can read various image formats(i.e. tiff, jpeg, bmp etc.) and convert them into text format(.txt) [7]-[8]

Although commercial ocr(such as AbbuyFinereader and Maestro) are more efficient than tesseract-ocr , but they are highly expensive as compared to it. Thus, it is widely used as it is cost effective. Now emphasis is put on improving its performance [9].

d) MODI – for integration of OCR functionality into our own application. .It is a discontinued MS office application that supports editing documents .It permits users for scanning documents , recognizing images using ocr , viewing a scanned document and annotating scanned documents [10].

Although ,modi is already available to developers that have installed an MS office product yet emgucv is better and efficient than modi because emgucv recognizes the text(in images) efficiently in case if there is any folds in our document.

III. PROPOSED SYSTEM

- System will have capabilities to scan documents and import pre existing scanned or clicked images.
- After Scanning, the documents will be stored in image format
- The system will also extract text out of the document and store that in the database. This will be done using Optical Character Recognition System.
- The text extracted will be stored in a Database for easy retrieval.
- As we search using a text or phrase we can locate the document and retrieve the original image.

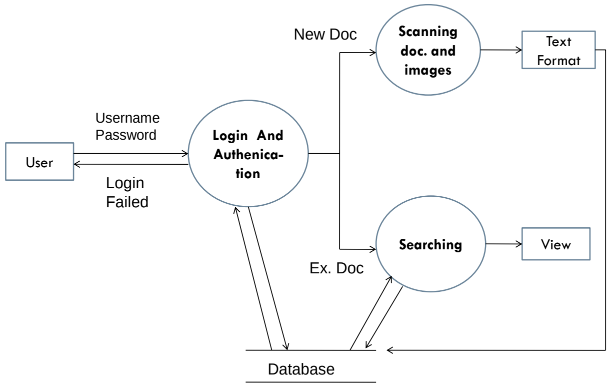

Fig 1.General Scheme for Working of Developed System

IV. IMPLEMENTATION OF MODULES AND RESULTS

The software has been designed in a modular manner. There is a separate module for the every function of this Application. These are then integrated to build an easy to use system(as shown in figure 1).Here is a brief listing of the various Design Modules of the Software:

1) *Scanning Sub-System*:

The sub module will integrate a scanning software into our application and the user will be able scan documents directly from our application and does not have to use any other software.Wia library is used for this implementation.

2) *Image to Text Conversion Sub-System*:

This module will be based on the Principal of Optical Character Recognition, where we will implement an OCR Algorithm to extract Text from Images and stores the text along with the image into the database. Modi and tesseractocr library is used for this implementation.

3) *Intelligent Search Sub-System*:

This module will enable users to search for entered keywords in the system and all documents pertaining to that keyword or containing the keyword.

4) *Login Protection*:

The application will be protected from unauthorized access using login protection so that the user has to enter username and password to access the details.

V. EXPERIMENTAL STUDY

Before the development of this technique, we need to hire employees to do the task of data entry because of which it takes a lot of time and chances of errors are very high.

But now less labour force is required. The wastage of time has been reduced very much and level of accuracy has been increased. Now Digital Archiving implementation has become very efficient.

VI. CONCLUSION

With the help of our developed application ; now we can store almost everything into electronic format instead of paper format and Users have advantage of searching through their documents and easily locating them.

ACKNOWLEDGMENT

We are thankful to our management for their continuing support and encouragement for completing this work and we are also thankful to our head of the department for her valuable suggestions.

REFERENCES

[1] Amarjot Singh, KetanBacchuwar and AkshayBhasin , "*A survey of ocr applications*" , *International Journal of Machine Learning and Computing*, vol.2, No.3 , June 2012.

[2] G.S. Lehal and ChandanSingh , Punjabi University, Patiala,Punjab,India " *Feature Extraction and classification of ocr for Gurmukhi script*".

[3] http://www.oclc.org/digital-archive.en.html.

[4] Steven Holzner –" *Visual Basics.Net Programming"* .

[5] http://www.emgu.com.

[6] http://social.msdn.microsoft.com>VisualStudioLanguagesForums>Visual Basics.

[7] http://code.google.com/p/tesseract-ocr.

[8] R.W. Smith "*The Extraction and Recognition of text from Multimedia Document Images*", Phd Thesis, University of Bristol,November 1987.

[9] http://vbridge.co.uk.

[10] MartinWelker http://www.codeproject.com>Articles>EnterpriseSystems>OfficeDevelopment>General.