OCR BASED IMAGE TEXT TO SPEECH CONVERSION

SATHISH KUMAR K R[1], SANTHOSH S[2], SANJAY V[3], SANTHOSH KUMAR V[4]
1 ASSISTANT PROFESSOR, 2,3,4 UG STUDENTS
DEPARTMENT OF COMPUTER SCIENCE
MAHENDRA ENGINEERING COLLEGE, TAMILNADU, INDIA

Abstract

Optical Character Recognition (OCR) is used for the purpose of extracting text from an image. The main agenda of an OCR is to make easy viewed & editable documents from existing paper documents or image files. That can be used to translate images of any format to the text format. This website helps one to convert the texts in image files into editable text files. So that we have to convert bulk of pages converted into editable text in a simple way.

Keywords: OCR Convertion, Text Converter

1. INTRODUCTION

1.1 PURPOSE

The main purpose of Optical Character Recognition (OCR) system based on a grid infrastructure is to perform Document Image Analysis, document processing of electronic document formats converted from paper formats more effectively and efficiently. This improves the accuracy of recognizing the characters during document processing compared to various existing available character recognition methods. Here OCR technique derives the meaning of the characters, their font properties from their bit-mapped images. ¬ The primary objective is to speed up the process of character recognition in document processing. As a result the system can process huge number of documents with-in less time and hence saves the time. 3 ¬ Since our character recognition is based on a grid infrastructure, it aims to recognize multiple heterogeneous characters that belong to different universal languages with different font properties and alignments.

1.2 SCOPE

The scope of our product Optical Character Recognition on a grid infrastructure is to provide an efficient and enhanced software tool for the users to perform Document Image Analysis, document processing by reading and recognizing the characters in research, academic, governmental and business organizations that are having large pool of documented, scanned images. Irrespective of the size of documents and the type of characters in documents, the product is recognizing them, searching them and processing them faster according to the needs of the environment

1.3 DESCRIPTION

In the running world there is a growing demand for the users to convert the printed documents in to electronic documents for maintaining the security of their data. Hence the basic OCR system was invented to convert the data available on papers in to computer process able documents, So that the documents can be editable and reusable. The existing system/the previous system of OCR on a grid infrastructure is just OCR without grid functionality. That is the existing system deals with the homogeneous character recognition or character recognition of single languages.

2. LITERATURE SURVEY

2.1 INTENDED AUDIENCE

In this section, we identify the audience who are interested with the product and are involved in the implementation of the product either directly or indirectly. As from our research, the OCR system is mainly useful in R&D at various scientific organizations, in governmental institutes and in large business organizations, we identify the following as various interested audience in implementing OCR system. The scientists, the research scholars and the research fellows in telecommunication institutions are interested in using OCR system for processing the word document that contains base paper for their research. The Librarian to manage the information contents of the older books in building virtual digital library requires use of OCR system. Various sites that vendor e-books have a huge requirement of this OCR system inorder to scan all the books in to electronic format and thus make money. The Amazon book world is largely using this concept to build their digital libraries.

2.2 READING SUGGESTIONS

Now we present the reading suggestions for the users or clients through which the user can better understand the various phases of the product. These suggestions may be effective and useful for the beginners of the product rather than the regular users such as research scholars, librarians and administrators of various web-sites. With these suggestions, the user need not waste his time in scrolling the documents up and down, browsing through the web, visiting libraries in search of different books and The following are the various reading suggestions that the user can follow in-order to completely understand about our product and to save time.

4. PROPOSED SYSTEM

The proposed system is a Text Recognition using Optical Character Recognition (OCR) named Tesseract, which is open-source software released under the Apache License, Version 2.0 and development has been sponsored by Google since 2006.

4.1 BENEFIT OF PROPOSED SYSTEM

The benefit of proposed system that overcomes the drawback of the existing system is that it supports multiple functionalities such as editing and searching. It also adds benefit by providing heterogeneous characters recognition.

5. SYSTEM ARCHITECTURE DIAGRAM

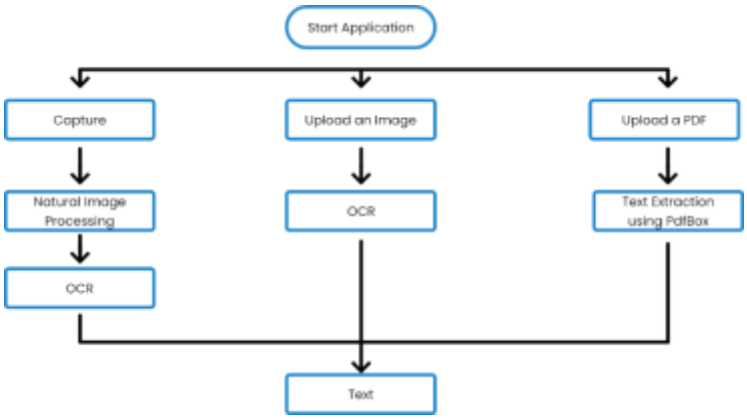


FIG 1- SYSTEM ARCHITECTURE DIAGRAM

5.1 DATA FLOW DIAGRAM

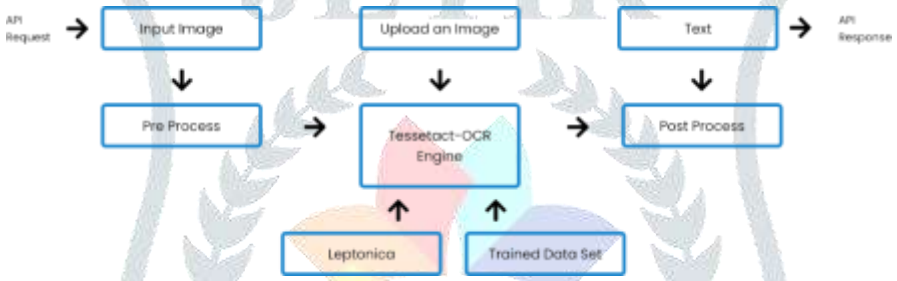


FIG 2 – DATA FLOW DIAGRAM

6 SOFTWARE REQUIREMENT ANALYSIS

6.1 PROBLEM STATEMENT

The problem here is for the software systems to recognize characters in computer system when information is scanned through paper documents as we know that we have number of newspapers and books which are in printed format related to different subjects. Whenever we scan the documents through the scanner, the documents are stored as images such as jpeg, gif etc., in the computer system. These images cannot be read or edited by the user. But to reuse this information it is very difficult to read the individual contents and searching the contents form these documents line-by-line and word-by-word. These days there is a huge demand in "storing the information available in these paper documents in to a computer storage disk and then later editing or reusing this information by searching process".

6.2 MODULES AND THEIR FUNCTIONALITIES

Our software system Optical Character Recognition on a grid infrastructure can be divided into five modules based on its functionality. The modules classified are as follows.

Document Processing Module

System Training Module

Document Recognition Module

Document Editing Module and

Document Searching Module.

6.3 DOCUMENT PROCESSING MODULE

This module is accessed by administrator whose role in our application is a librarian.This module perform certain activities such as scanning documents, storing them as images, recognizing characters in images to transfer them into word format. During the recognition process, this module uses the OCR methodology in support of grid infrastructure datastructure. The module supports the following services.

Scanning printed documents.

Storing the documents as snapshots or images.

Processing those image-based documents.

Converting these image-based documents into e-documents(also called structured documents).

Recognizing the characters in documents.

Generating grid infrastructure data structure.

6.4 DOCUMENT EDITING MODULE

This module can be accessed by both the administrator and the end-user during document editing to implement the character recogniton process. Once the scanned documents are stored, they reside in computer memory. This data resides in the form of an image that is just viewable in an image viewer. Hence, the document is first coverted into a form such that it is editable. The desired form of the document may be MS-Word, Text,… as specified by the user. The objective of this module is to let the user perform.

6.5 DOCUMENT SEARCHING MODULE

This module can be accessed by both the administrator and the end-user during the search of the user required document to implement the character recogniiton process on it. The user requests the system to search for a particular document. Then the system finds the documents based on OCR methodology and returns the result of the search to the user.

7. CONCLUSION

What does the future hold for OCR? Given enough entrepreneurial designers and sufficient research and development dollars, OCR can become a powerful tool for future data entry applications. However, the limited availability of funds in a capital-short environment could restrict the growth of this technology. But, given the proper impetus and encouragement, a lot of benefits can be provided by the OCR system. They are:-

The automated entry of data by OCR is one of the most attractive, labor reducing technology.

The recognition of new font characters by the system is very easy and quick.

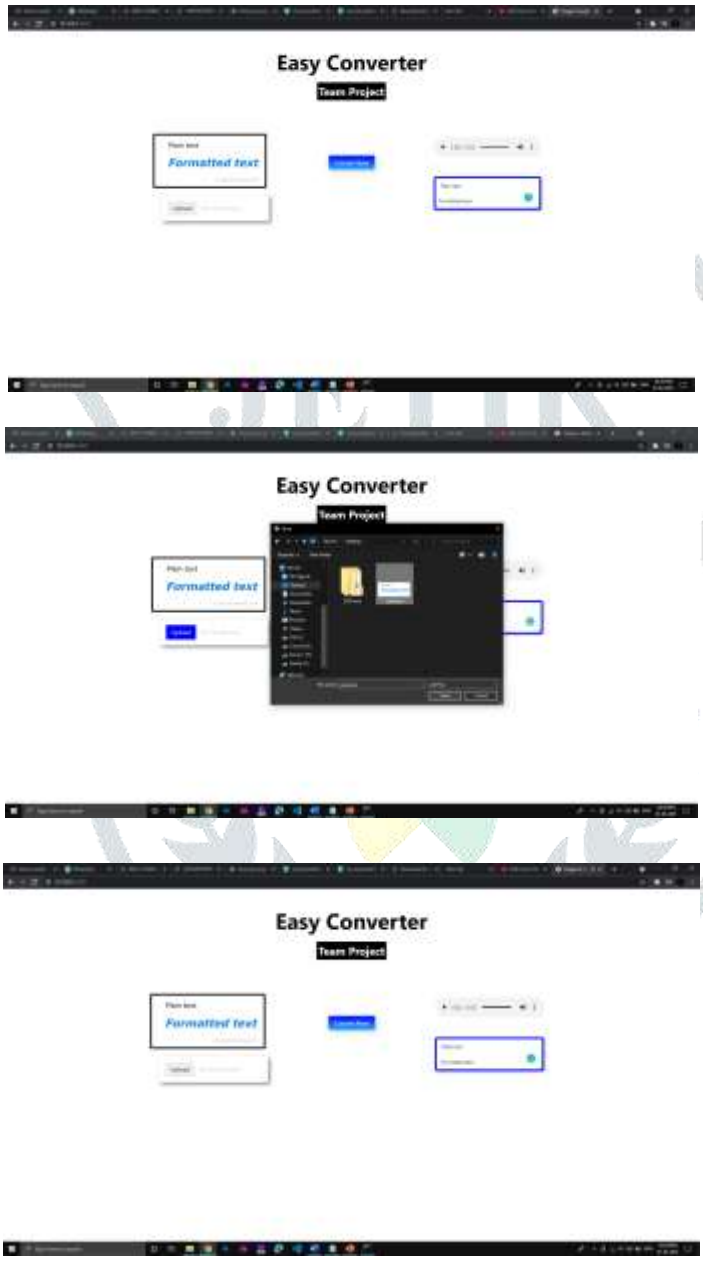
We can edit the information of the documents more conveniently and we can reuse the edited information as and when required.

The extension to software other than editing and searching is topic for future works. The Grid infrastructure used in the implementation of Optical Character Recognition system can be efficiently used to speed up the translation of image based documents into structured documents that are currently easy to discover, search and process.

8. RESULTS

Observed outcome of project : Text is extracted from the image and converted to block of text. It recognizes both capital as well as small letters.

9. SCREENSHOTS



10. REFERENCES

http://alvinalexander.com/java/java-image-how-to-crop-image-in-java

http://kalanir.blogspot.in/2010/02/how-to-split-image-into-cunks-java.html

http://www.voicerss.org/tts/

http://www.comsys.net/technology/speechframe/text-to-speechtts.html