Extract Text from Scanned Copy to Excel sheet by using OCR

Prof. Manisha Abhyankar*
Prof Deepali Deshmukh**

Abstract

*Today's uses of OCR are still somewhat limited to the scanning of the written word into useable computer text. The uses include word processing, mail delivery system scanning, ticket reading, and other such tasks. It is a tedious and time consuming process to list out data of each candidate, manually check them against documents submitted, resulting in delay and wastage of valuable resources. Multiple merit lists often leads to duplication of work for the institution, even candidates / guardians have to visit again and again to check the latest merit list. One of the current hopes for OCR is the chance of developing OCR software that can read compressed files. Text that is compressed into an image and saved as ASCII or Hexadecimal data could be read by OCR and transferred back into readable text. The objective of the study is developing the "**Extract Text from Scanned Copy to Excel Sheet by using OCR"** as there are many software to extract text from images. While installing a new software, use Image processing Concept which can fetch the text data from scanned copy as there is no need of Internet connection.*

Keywords: **OCR, Excel sheet**.

Introduction

Optical character recognition, usually abbreviated to **OCR**, is the mechanical or electronic translation of images of handwritten or typewritten text (usually captured by a scanner) into machine-editable text.Optical Character Recognition (OCR) is one of the most common and useful applications of machine vision technology.

OCR began as a field of research in pattern recognition, artificial intelligence, and machine vision. Academic research in the field continues, the focus on OCR has shifted to implementation of proven techniques. Optical character recognition (using optical techniques such as mirrors and lenses) and digital character recognition (scanners and computer algorithms) were originally considered separate fields. Because very few applications survive that use true optical techniques, the optical character recognition term has now been broadened to cover digital character recognition as well.

Fundamentals of OCR system

Three main components

- Image Scanner
- OCR Hardware/Software
- Output Interface

The OCR system consists of 3 main components Scanner, OCR Hardware/Software, Output Interface. The document to be digitized is taken and scanned as an image using a scanner. This image is then given to the OCR Hardware/ Software component. This image undergoes modifications in the various processes involved in this component. The output of the OCR system is digital form of he scanned document.

Computer Science Department KBP College,Vashi, Email: *msabhyankar@yahoo.com; **deepalid98@gmail.com

OCR System

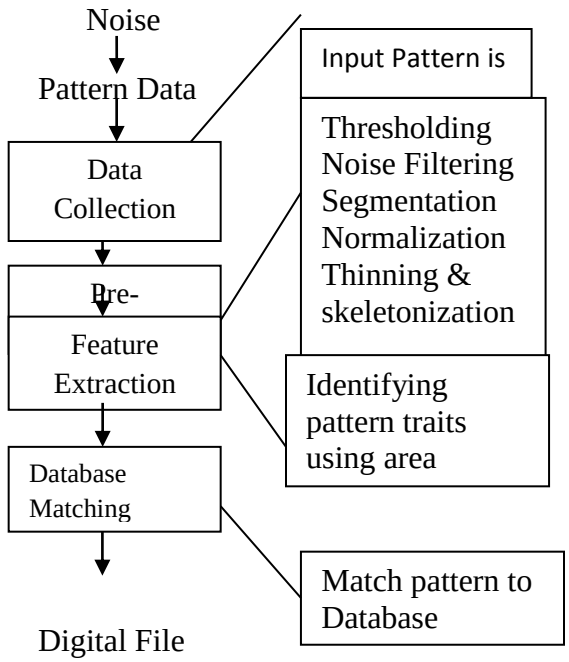
Basics for developing ocr system

TIF (Tagged Image File Format)

The document is scanned and the images is stored in Tagged Image File Format. Storing the files in .TIF format gives us the following advantage

- It is extremely mature and stable.
- It is independent of computer architecture, Operating System and graphics hardware.
- Have large options for development and organization of images.
- It is compact and can show black and white, grayscale as well as color images

SYSTEM DESIGN (conceptual model)



Methodology

- **Data collection and flow analysis**

  The document is scanned and the image obtained is stored in .TIF (Tagged Image File Format). Care should be taken while scanning of the document so that there is no dust on the scanner also that the document is not too dirty to be difficult to recognize. This image is the input data for the OCR software. This image undergoes various processes, which modify and alter the image to get the digital output. Those processes have been listed and explained in chronological order below.

- **Thresholding**

  Thresholding involves separation of the desired target object (text) from the background. A threshold value (t) is set between 0 to 255. Every pixel in the scanned document is compared with the threshold value and if its intensity is found less than the threshold value then its intensity is set to 0 else it is set to 1 in binary.

- **Noise Removal**

Noise is basically a black pixel that must be white (or vice versa). Noise can be due to dirt on the scanner lens or paper etc. Noise affects the quality of the skeleton produced, and it should be removed before extracting the character's skeleton.

- **Segmentation**

  The image after undergoing Thresholding and Noise removal undergoes a process known as Segmentation. The image undergoes vertical and horizontal segmentation. In horizontal segmentation the lines of

  characters are separated from each other in horizontal manner i.e. each row is separated. These separated rows are then made to undergo vertical segmentation. During vertical segmentation the characters in each row are separated from each other. Thus after segmentation we get individual character images.

- **Normalization**

  After the process of segmentation we get images of individual character. These characters may be of any arbitrary pixel length. We need to standardize it to a given size. Normalization is the method used to do that.It standardizes the given character image into standard pixel size of 20*20 pixels.

- **Thinning**

  Thinning is used for extracting the pattern's skeleton. Thinning operation used to remove selected foreground pixels from a binary image. It is believed that thin-line representation is closer to the human conception of the pattern. Thinning reduces the memory storage and computation needs for an image.

- **Character Representation**

  The characters present in the image are represented by Area descriptors known as moments. These help in representing the shape of the object under consideration. The moments are used to calculate centroid for the image. These are then used to calculate normalized central moments (NCM) these moments are invariant to scaling and translations. NCM is further used to find the principal angle (PA) for the image. PA has the advantage to being invariant to rotations. Thus the characters are represented by their principal angle.

- **Proposed system**

  OCR technology allows the conversion of scanned images of printed text into text or information that can be understood or editable. Optical character recognition, usually abbreviated to **OCR**, is the mechanical or electronic translation of scanned images of handwritten, typewritten, or printed text into machine-encoded text.

  OCR technology uses three steps- Scanning acquisition of printed documents as optical images. Recognition- involves converting these images to character streams representing letters of recognized words and the final element involves accessing or storing the converted text.

  Converted text is referred as extracted text. When, the user begins by capturing an image containing text of interest using the Scanner, the specified area of the image is processed on the device in order to optimize it for transfer and input to the OCR. Firstly it analyses text, transforms text into pronounceable form.

Conclusion

OCR is essential in digitizing the conventional paper work system. Our system will provide an affordable and reliable option to users.
OCR is a modern technology used for digital replication.
There are advantages and disadvantages of OCR.

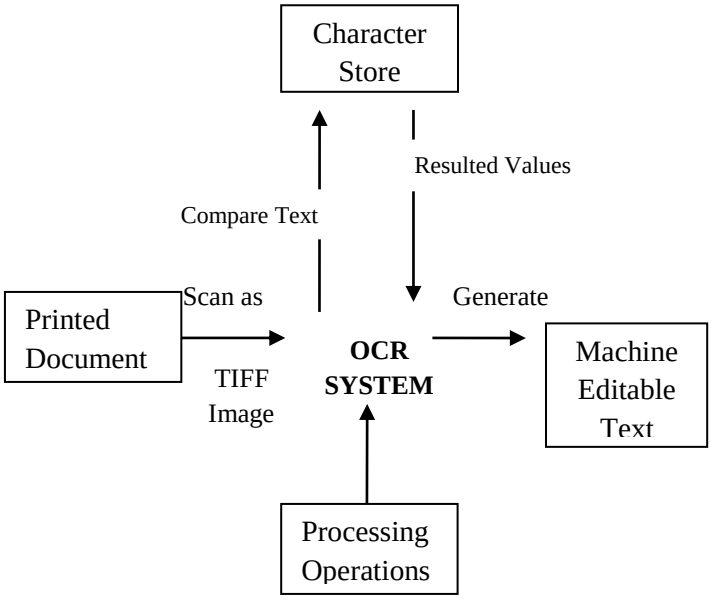


Fig 1.2: **Data Flow Diagram**

Advantage of OCR

- The Optical Character Recognition process can save both time and effort when developing a digital replica of the document.
- It provides a fast and reliable alternative to typing manually.
- All documents created through OCR program software are editable and allow you to modify the content as you see fit. If you compare the cost of OCR with the cost of manual data entry, OCR is a lot cheaper. OCR program software proved to be better than data entry service for organizations specializing in developing electronic copy out of printed books.

Future scope

Today's uses of OCR are still somewhat limited to the scanning of the written word into useable computer text. The uses include word processing, mail delivery system scanning, ticket reading, and other such tasks.One of the current hopes for OCR is the chance of developing OCR software that can read compressed files by using Mobile Application. Text that is compressed into an image and saved as ASCII or Hexadecimal data could be read by OCR and transferred back into readable text. Our next works with OCR Mobile Application will include the improvement of the

results by the use of table boundaries detection techniques and the use of text post-processing techniques to detect the noise and to correct bad-recognized words.

Appendix

OCR:- Optical character recognition, NCM:- normalized central moments, PA:-Principle angle
E-GOV :- Electronic governance,GUI:- Graphical user interface

Apache Hadoop Goes Realtime at Facebook

***Prof.Komal Shringare**

Abstract

Facebook recently deployed Facebook Messages, its first ever user-facing application built on the Apache Hadoop platform. Apache HBase is a database-like layer built on Hadoop designed to support billions of messages per day. This paper describes the reasons why Facebook chose Hadoop and HBase over other systems such as Apache Cassandra and Voldemort and discusses the applicationBs requirements for consistency, availability, partition tolerance, data model and scalability. I explore the enhancements made to Hadoop to make it a more effective realtime system, the tradeoffs we made while configuring the system, and how this solution has significant advantages over the sharded MySQL database scheme used in other applications at Facebook and many other web-scalecompanies. I discuss the motivations behind my design choices, the challenges that we face in day-to-day operations, and future capabilities and improvements still under development.I offer these observations on the deployment as a model for other companies who are contemplating a Hadoop-based solution over traditional sharded RDBMS deployments.

Keywords: Data, scalability, resource sharing, distributed file system, Hadoop, Hive, HBase, Facebook, Scribe, log aggregation, distributed systems.

Introduction

Apache Hadoop [1] is a top-level Apache project that includes open source implementations of a distributed file system [2] and MapReduce that were inspired by GoogleBs GFS [5] and MapReduce [6] projects. The Hadoop ecosystem also includes projects like Apache HBase [4] which is inspired by GoogleBs BigTable, Apache Hive [3], a data warehouse built on top of Hadoop, and Apache ZooKeeper [7], a coordination service for distributed systems.

At Facebook, Hadoop has traditionally been used in conjunction with Hive for storage and analysis of large data sets. Most of this analysis occurs in offline batch jobs and the emphasis has been on maximizing throughput and efficiency. These workloads typically read and write large amounts of data from disk sequentially. As such, there has been less emphasis on making Hadoop performant for random access workloads by providing low latency access to HDFS. Instead, I have used a combination of large clusters of MySQL databases and caching tiers built using memcached[8]. In many cases, results from Hadoop are uploaded into MySQL or memcached for consumption by the web tier.

The first set of applications requires realtime concurrent, but sequential, read access to a very large stream of realtime data being stored in HDFS. An example system generating and storing such data is Scribe [9], an open source distributed log aggregation service created by and used extensively at Facebook. Previously, data generated by Scribe was stored in expensive and hard to manage NFS servers. Two main applications that fall into this category are Realtime Analytics