OPTICAL CHARACTER RECOGNITION

A report submitted
in fulfillment of the requirements
for
Degree of
Bachelor of Technology

By
Shashank Bhardwaj(20135032)
Shailendra Azad(20135159)
Sanjay Kumar Shahu(20135013)
Suyog Bhandari(20135007)

Preface

Optical Character Recognition usually abbreviated as OCR, involves a computer system designed to translate images of typewritten text (usually captured by a scanner) into machine editable text or to translate pictures of characters into a standard encoding scheme representing them. In OCR processing, the scanned-in image or bitmap is analyzed for light and dark areas inorder to identify each alphabetic letter or numeric digit. When a character is recognized, it is converted into an ASCII code. Special circuit boards and computer chips designed expressly for OCR are used to speed up the recognition process.

OCR began as a field of research in artificial intelligence and computational vision. OCR is being used by libraries to digitize and preserve their holdings. OCR is also used to process checks and credit card slips and sort the mail. Billions of magazines and letters are sorted everyday by OCR machines, considerably speeding up mail delivery.

CONTENTS

CHAPTER 1

1. INTRODUCTION

In the running world, there is growing demand for the software systems to recognize charactersin 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. 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 reusing this information by searching process”. One simple way to store information in these paper documents in to computer systemis to first scan the documents and then store them as images. 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. The reason for this difficulty is the font characteristics of the characters in paper documents are different to font of the characters in computer system. As a result, computer is unable to recognize the characters while reading them. This concept of storing the contents of paper documents in computer storage place and then reading and searching the content is called Document Processing. Sometimes in this document processing we need to process the information that is related to languages other than the English in the world. For this document processing we need a software system called **CHARACTER RECOGNITION SYSTEM**. This process is also called DOCUMENT IMAGE ANALYSIS (DIA).

Thus our need is to develop character recognition software system to perform Document Image Analysis which transforms documents in paper format to electronic format. For this process there are various techniques in the world. Among all those techniques we have chosen Optical Character Recognition as main fundamental technique to recognize characters. The conversion of paper documents in to electronic format is an on-going task in many of the organizations particularly in Research and Development (R&D) area, in large business enterprises, in government institutions, so on. From our problem statement we can introduce the necessity of Optical Character Recognition in mobile electronic devices such as cell phones, digital camerasto acquire images and recognize them as a part of face recognition and validation.
To effectively use Optical Character Recognition for character recognition in-order to perform Document Image Analysis (DIA), we are using the information in Grid format. This system is thus effective and useful in Virtual Digital Library’s design and construction.

2. WHAT IS OCR??

Optical Character Recognition, or OCR, is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data. A scanner is not enough to make magazine article, brochure, or PDF contract information available for editing, say in Microsoft Word. All a scanner can do is create an image or a snapshot of the document that is nothing more than a collection of black and white or colour dots, known as a raster image. In order to extract and repurpose data from scanned documents, camera images or image-only PDFs, you need an OCR software that would single out letters on the image, put them into words and then - words into sentences, thus enabling you to access and edit the content of the original document.

3. 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 ofthe 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.

4. ARCHITECTURE

The Architecture of the optical character recognition system on a grid infrastructure consists ofthe three main components. They are:-

➢ Scanner

➢ OCR Hardware or Software

➢ Output Interface

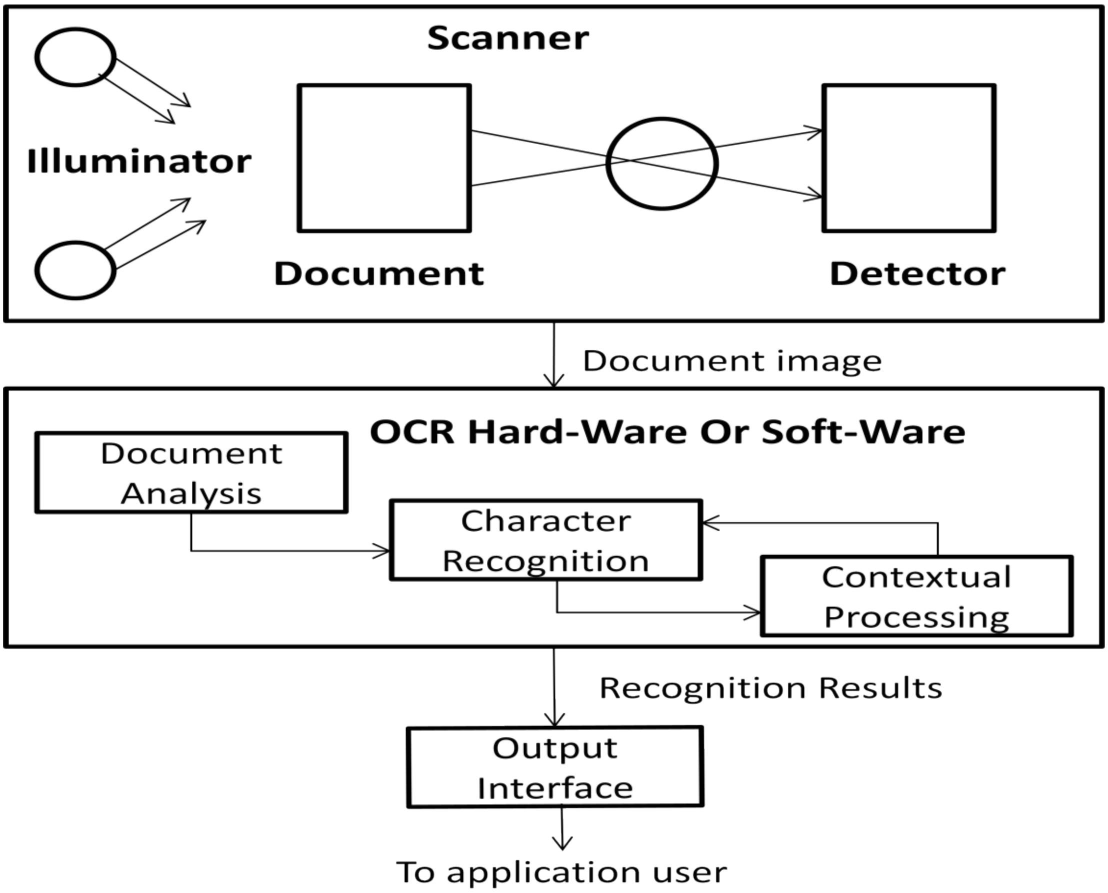

Figure.1: OCR Architecture

CHAPTER 2

1. STEPS IN OCR

A scanner generates an image of the paper document. On the scanned image, pre-processing techniques, feature extraction. classification and post-processing techniques are applied to get thedesired output.

2. PRE-PROCESSING

OCR cannot be applied without the help of Image Processing and/or Artificial Intelligence. Any OCR system goes through numerous phases including: data acquisition, pre-processing, feature extraction, classification and post-processing where the most crucial aspect is the pre-processing which is necessary to modify the data either to correct deficiencies in the data acquisition process due to limitations of the capturing device sensor, or to prepare the data for subsequent activities later in the description or classification stage. Data pre-processing describes any type of processing performed on raw data to prepare it for another processing procedure. Hence, pre-processing is the preliminary step which transforms the data into a format that will be more easily and effectively processed. Therefore, the main task in pre-processing the captured data is to decrease the variation thatcauses a reduction in the recognition rate and increases the complexities, as for example, pre-processing of the input raw stroke of characters is crucial for the success of efficient character recognition systems. Thus, pre-processing is an essential stage prior to feature extraction since it controls the suitability of the results for the successive stages.

OCR software often "pre-processes" images to improve the chances of successful recognition. Techniques include:

- De-skew – If the document was not aligned properly when scanned, it may need to be tilted a few degrees clockwise or counterclockwise in order to make lines of text perfectly horizontal or vertical.
- De-speckle – remove positive and negative spots, smoothing edges
- Binarisation – Convert an image from color or greyscale to black-and-white (called a "binary image" because there are two colours). The task of binarisation is performed as a simple way of separating the text (or any other desired image component) from the background. The task of binarisation itself is necessary since most commercial recognition algorithms work only on binary images since it proves to be simpler to do so. In addition, the effectiveness of the binarisation step influences to a significant extent the quality of the character recognition stage and the careful decisions are made in the choice of the binarisation employed for a given input image type; since the quality of the binarisation method employed to obtain the binary result depends on the type of the input image (scanned document, scene text image, historical degraded document etc.).
- Line removal – Cleans up non-glyph boxes and lines
- Layout analysis or "zoning" – Identifies columns, paragraphs, captions, etc. as distinct

blocks. Especially important in multi-column layouts and tables.

- Line and word detection – Establishes baseline for word and character shapes, separates words if necessary.
- Script recognition – In multilingual documents, the script may change at the level of the words and hence, identification of the script is necessary, before the right OCR can be invoked to handle the specific script.
- Character isolation or "segmentation" – For per-character OCR, multiple characters that are connected due to image artifacts must be separated; single characters that are broken into multiple pieces due to artifacts must be connected.

3. FEATURE EXTRACTION

Feature extraction decomposes glyphs into "features" like lines, closed loops, line direction, and line intersections. These are compared with an abstract vector-like representation of a character, which might reduce to one or more glyph prototypes. General techniques of feature detection in computer vision are applicable to this type of OCR, which is commonly seen in "intelligent" handwriting recognition and indeed most modern OCR . Nearest neighbour classifiers suchas the k-nearest neighbors algorithm are used to compare image features with stored glyph fea- tures and choose the nearest match.

4. CHARACTER RECOGNITION

There are two basic types of core OCR algorithm, which may produce a ranked list of candidate characters.

Matrix matching involves comparing an image to a stored glyph on a pixel-by-pixel basis; it is also known as "pattern matching", "pattern recognition", or "image correlation". This relies onthe input glyph being correctly isolated from the rest of the image, and on the stored glyph being in a similar font and at the same scale. This technique works best with typewritten text and does not work well when new fonts are encountered.

The second pass is known as "adaptive recognition" and uses the letter shapes recognised with high confidence on the first pass to recognise better the remaining letters on the second pass. This is advantageous for unusual fonts or low-quality scans where the font is distorted (e.g. blurred or faded).

5. POST-PROCESSING

OCR accuracy can be increased if the output is constrained by a lexicon – a list of words that are allowed to occur in a document. This might be, for example, all the words in the English language, or a more technical lexicon for a specific field. This technique can be problematic if the document contains words not in the lexicon, like proper nouns. It uses its dictionary to influence the character segmentation step, for improved accuracy.

The output stream may be a plain text stream or file of characters, but more sophisticated OCR systems can preserve the original layout of the page and produce, for example, an annotat- ed PDF that includes both the original image of the page and a searchable textual representation.

"Near-neighbor analysis" can make use of co-occurrence frequencies to correct errors, by noting that certain words are often seen together. For example, "Washington, D.C." is generally far more common in English than "Washington DOC".

Knowledge of the grammar of the language being scanned can also help determine if a word is likely to be a verb or a noun, for example, allowing greater accuracy.

The Levenshtein Distance algorithm has also been used in OCR post-processing to further optimize results from an OCR API.

CHAPTER 3

1. ASSUMPTIONS MADE

1) The input scanned document is assumed to be only in jpg, gif or jpeg format.
2) The input scanned document only consists of text in black written on a white background, itcontains no graphical images.
3) After loading the image, first line segmentation is performed and then only word segmentationcan be performed .
4) Lines can be dropped , dragged , added or deleted only after default line segmentation hasbeen performed.
5) For loading another image, the clear button is pressed and then the image is loaded.

2. APPLICATIONS

OCR engines have been developed into many kinds of domain-specific OCR applications, such as receipt OCR, invoice OCR, check OCR, legal billing document OCR.

They can be used for:

- Data entry for business documents, e.g. check, passport, invoice, bank statement and receipt
- Automatic number plate recognition
- Automatic insurance documents key information extraction
- Extracting business card information into a contact list
- More quickly make textual versions of printed documents, e.g. book scanning for Project Gutenberg
- Make electronic images of printed documents searchable, e.g. Google Books
- Converting handwriting in real time to control a computer (pen computing)
- Defeating CAPTCHA anti-bot systems, though these are specifically designed to prevent OCR
- Assistive technology for blind and visually impaired users

3. CONCLUSION

In this chapter, preprocessing techniques used in document images as an initial step in character recognition systems were presented. Future research aims at new applications such as online character recognition used in mobile devices, extraction of text from video images, extraction of information from security documents and processing of historical documents. The objective of such research is to guarantee the accuracy and security of information extraction in real time applications. Even though many methods and techniqueshave been developed for preprocessing there are still problems that are not solved completely and more investigations need to be carried out in order to provide solutions.

Most of preprocessing techniques are application-specific and not all preprocessing techniques have to be applied to all applications. Each application may require different preprocessing techniques depending on the different factors that may affect the quality of its images, such as those introduced during the image acquisition stage.

Image manipulation/enhancement techniques do not need to be performed on an entire imagesince not all parts of an image is affected by noise or contrast variations; therefore, enhancement of a portion of the original image maybe more useful in many situations. This is obvious when an image contains different objects which may differ in their brightness or darkness from the other parts of the image; thereby, when portions of an image can be selected, either manually or automatically according to their brightness such processing can be used to bring out local detail.

4. REFERENCES

1) Davies, E. (2005). *Machine Vision – Theory Algorithms Practicalities,* Third Edition*,* Morgan Kaufmann Publishers, ISBN 13: 978-0-12-206093-9, ISBN-10: 0-12-206093-8.

2) Fischer, S., (2000). *Digital Image Processing: Skewing and Thresholding,* Master ofScience thesis, University of New South Wales, Sydney, Australia.

3) Gonzalez, R.; Woods, R. & Eddins, S. (2004). *Digital Image Processing using MATLAB*, Pearson Education Inc., ISBN 0-13-008519-7.