GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

OPTICAL CHARACTER RECOGNITION

A Project Report of Capstone Project 2

Submitted by

RAGHAV CHAUHAN
(1613105078/16SCSE105041)

in partial fulfillment for the award of the degree of

Bachelor of Technology

IN

Computer Science and Engineering With Specialization of Cloud Computing and Virtualization

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of
Dr. TAPAS KUMAR
Associate Professor

APR/MAY- 2020

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report " **OPTICAL CHARACTER RECOGNITION**" is the bonafide work of "***RAGHAV CHAUHAN***" who carried out the project work under my supervision.

Signature of the Head

Dr. Munish Shabarwal
PhD(Management),Phd(CS)
Professor & Dean
School of Computing Science
& Engineering

Signature of Supervisor

Dr. Tapas Kumar
Associate Professor
School of Computing Science
& Engineering

ABSTRACT

There are various fields where there is a huge demand for storing information into computer storage disk from data available in printed or handwritten documents or images. This is done to re-utilize this information through computers. One of the way to store information to a system from these documents is to scan the documents and then keep them as image files. For re-utilizing the information, it is difficult to read or query text or other information from the image files. One of the technique for automatically retrieving and storing information, in particular format, from image files is required. One active research area that develop a computer system with the ability to extract and process text from images automatically is Optical character recognition. OCR helps in achieving modification or conversion of any form of text or text-containing documents like handwritten text, printed or scanned text images, into an editable digital format for deeper and processing. Therefore, OCR helps a machine to automatically identify text in such documents.

TABLE OF CONTENTS

INTRODUTION

Optical character recognition is a research area that try to develop a system for extracting and processing text from images automatically. There is a huge demand to store information to a computer storage disk from the data present in printed or handwritten documents for re-utilizing the information through computers. One way to store information to a system from these paper documents is to first scan the documents for storing them as image files. But is very difficult to read or query text or other information from these image files to reutilize them. Therefore a technique to automatically retrieve and store information, in particular text, from image files is needed. There are challenges that needs to be handled for achieving a successful automation. The font characteristics of the characters in paper documents and quality of images are only some of the recent challenges. These challenges leads to incorrect recognition of characters by computer system. So a mechanisms is required for character recognition to perform Document Image Analysis (DIA) that solves these challenges and produces electronic format from the transformed documents in paper format.

OCR is the process of modification of any form of text or text-containing documents such as handwritten text, printed or scanned text images, into an editable digital format for deeper and further processing. OCR technology enables a machine to automatically recognize text in such documents.

1.1 WORKING

Different fonts and ways to write a single character make this issue a challenge to solve. Before selecting an OCR algorithm, the image must be preprocessed for the image to be ready to be “read”.

Pre-processing

OCR software often “pre-process” images to boost the chances of recognition. Techniques include:

1. De-skew: If the document was not correctly aligned when scanned, it may need to be tilted a few degrees clockwise or counterclockwise to create text lines completely horizontal or vertical.

2. Despeckle: Remove positive and negative spots, smoothing edges

3. Binarization: Convert an image to black-and-white (called a “binary image” because there are two colors). The binarization task is conducted as an easy and accurate way to distinguish text (or any other required image element) from the background.

4. Line removal: Cleans up non-glyph boxes and lines.

5. Layout analysis or “zoning”: Identifies columns, paragraphs, captions, etc., as blocks. Particularly useful in multi-column layouts and tables.

6. Line and word detection: Establish word and character shapes baseline, divides words when required.

7. Script recognition: In multiple language documents, the script may transform at the word level and therefore script identification is vital before the relevant OCR can be utilized to manage the particular script.

8. Character isolation or “segmentation”: For OCR characters, various characters linked by image artifacts should be divided, single characters broken into several artifact-based pieces should be linked.

9. Normalization; Normalize aspect ratio and scale.

Feature Extraction

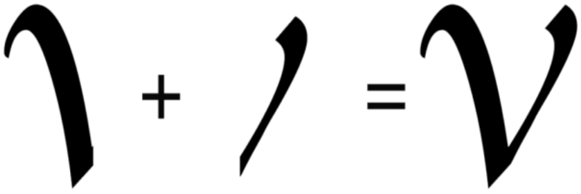
There are two main methods for extracting features in OCR:

1. In the first method, the algorithm for feature detection defines a character by evaluating its lines and strokes.

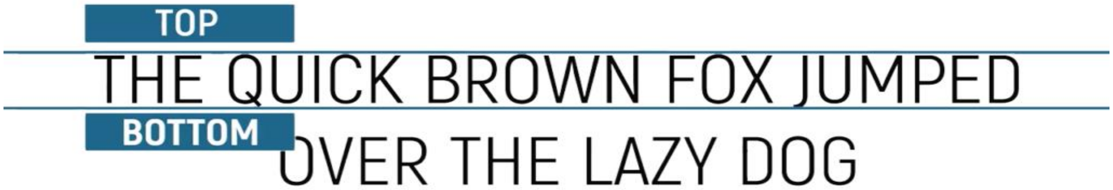
2. In the second method, pattern recognition works by identifying the entire character.

We can recognize a line of text by searching for white pixel rows that have black pixels in between. Similarly, we can recognize where a character starts and finishes.

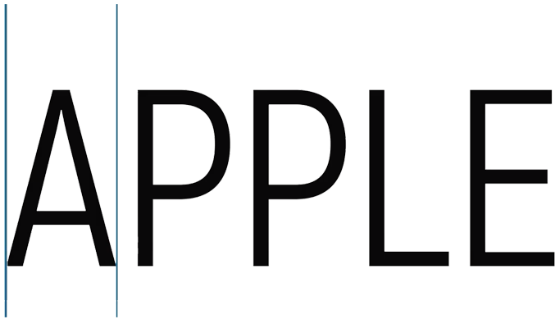
The following pictures show the visualization of these methods respectively:



Method 1 – Feature detection.

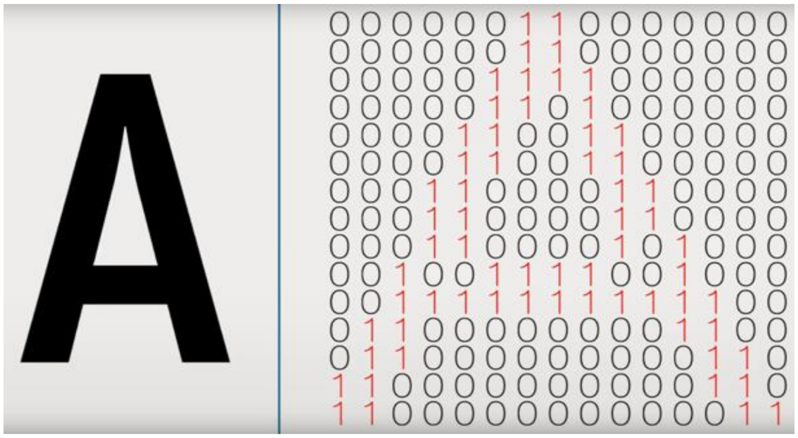


Method 2 – Pattern recognition on a row of text.



Method 2 – Pattern recognition on a single character.

Next, we convert the image of the character into a binary matrix where white pixels are 0s and black pixels are 1s as shown in the following image:



```
0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 0 0 1 1 1 1 0 0 0 0 0 0
0 0 0 0 0 1 1 0 1 0 0 0 0 0 0
0 0 0 0 1 1 0 0 1 1 0 0 0 0 0
0 0 0 0 1 1 0 0 1 1 0 0 0 0 0
0 0 0 1 1 0 0 0 0 1 1 0 0 0 0
0 0 0 1 1 0 0 0 0 1 1 0 0 0 0
0 0 0 1 1 0 0 0 0 1 0 1 0 0 0
0 0 1 0 0 1 1 1 1 0 0 1 0 0 0
0 0 1 1 1 1 1 1 1 1 1 1 1 0 0
0 1 1 0 0 0 0 0 0 0 0 1 1 0 0
0 1 1 0 0 0 0 0 0 0 0 0 1 1 0
1 1 0 0 0 0 0 0 0 0 0 0 1 1 0
1 1 0 0 0 0 0 0 0 0 0 0 0 1 1
```

Sample of binary matrix.

Then, by using the distance formula, we can find the distance from the center of the matrix to the farthest 1.

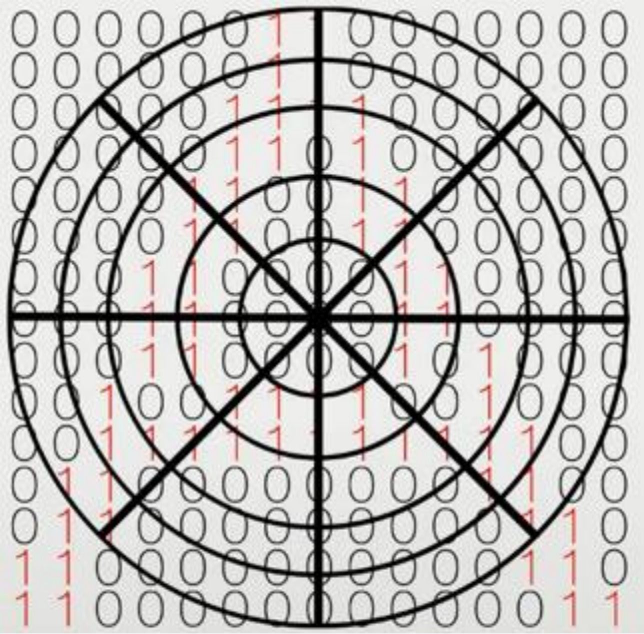
$$d = \sqrt{\left(y_2 - y_1\right)^2 + \left(x_2 - x_1\right)^2}$$

The distance formula.

We then create a circle of that radius and split it up into more granular sections.

At this stage, the algorithm compares each subsection to a database of matrices representing characters with different fonts to identify the character it has most in common statistically.

It makes it easy to bring printed media into the digital world by doing this for every line and character.



Compare each subsection against the matrix database.

Post-processing

OCR accuracy can be improved if the output is limited by a lexicon (a list of words permitted in a document). For instance, this could be all the words in English, or a more technical lexicon for a particular field.

This method can be less efficient if the document contains words that are not in the lexicon, like proper nouns.

Fortunately, to improve accuracy, there are OCR libraries available online for free. The Tesseract library is using its dictionary to control the segmentation of characters.

The output stream can be a single string or a character file, but more advanced OCR systems retain the original page structure and, for example, create a PDF containing both the original image pages and a searchable textual image.

1.2 ADVANTAGES

Optical character recognition has been performed in a numerous of applications. We discussed some of these application areas in this section.

Handwriting Recognition:

The capacity of a PC to get and translate intelligible handwritten data from sources, for example, paper records, photos, touch-screens and different gadgets ,is termed as Handwriting Recognition. The picture of the written content can be detected "off line" through a bit of paper by optical scanning (OCR) or clever word recognition. On contrary, the developments of the pen tip may be detected "on line", for instance by a pen-based PC screen surface.

Receipt Imaging:

Receipt imaging is broadly utilized as a part of numerous organizations applications to monitor financial records and keep accumulation of payments from heaping up. OCR simplifies information gathering and analysis, among different procedures in government offices and autonomous organization.

Legal Industry:

Legal industry is likewise one of the recipients of the OCR innovation. OCR is utilized to digitize documents, and to specifically enter into PC database. Legitimate experts can further search documents required from tremendous databases by basically writing a few keywords.

Banking;

Another important use of OCR is in banking , where OCR is utilized for processing cheques without human intervention. A cheque can be given to a machine where the framework filters the sum to be issued and the correct measure of cash is exchanged. Innovation has been idealized for printed cheque, and is genuinely precise for handwritten checks diminishing the hold-up time in banks.

Healthcare:

To process printed material, medicinal services have likewise seen an expansion in the utilization of OCR innovation. Medicinal service experts continuously requires to manage extensive volumes of documents for every patient, that also includes protection frames and in addition general health forms. It is valuable to input relevant information into an electronic database ,to stay aware of every one of this data. With OCR processing tools, we can extract data from structures and put it into databases, so that each patient's information is quickly recorded and retrieved when needed in future.

Captcha:

A CAPTCHA is a system that can create and grade tests that human can pass yet current software technology can't. Malicious programmer can make software to misuse personal information on websites. In CAPTCHA, a picture containing an arrangement of letters and numbers is produced with various size and textual styles, distracting backgrounds, arbitrary portions, highlights and noise to avoid reading text via OCR. Current OCR frameworks can be used to eliminate the noise and portion the picture to make the picture tractable by such malicious users.

Automatic Number Plate Recognition:

Automatic number plate recognition is utilized as a mass observation method making utilization of optical character recognition on pictures to recognize vehicle registration plates. ANPR has been made to store the pictures caught through the cameras including the numbers caught from license plate. ANPR innovation own to plate variety from one place to another as it is an area particular innovation. They are utilized by different police forces and as a technique for electronic toll accumulation on pay- per-use streets.

2.1CHALLENGES

For achieving good quality and high accuracy character recognition, OCR techniques requires high quality and high resolution images with basic structural properties like high differentiating text and background. One of the important and determining factor in the accuracy and success of OCR is the way images are generated, this affects the quality of images. OCR with images produced by scanners has high accuracy and good performance. In comparison to images produced by cameras usually are not as good of an input as scanned images to be used for OCR because of the environmental or camera related factors. There are numerous errors that can be generated , these are discussed below.

Scene Complexity:

There are large numbers of man-made objects which are included in camera taken images such as paintings, buildings, and symbols, in natural environment. These objects consists of comparative structures and appearances to text that makes text recognition very difficult in the processed image. Text itself is regularly laid out to encourage decipherability. There is a challenge with scene intricacy is that the surrounding scene makes it hard to differentiate text from non-text.

Conditions of Uneven Lighting:

While taking images in natural environments gives an uneven lighting and shadows. This is huge challenge for OCR as it lowers down the required characteristics of the image and causes less accurate detection, segmentation and recognition results. Such condition with uneven lighting distinguishes a scanned image from that produced with a camera. The lack of such disparities in lighting and shadows makes scanned images preferred over camera images for their better characteristics and quality. Using an on-camera flash may discard such problems with uneven lighting, it introduces new challenges.

Skewness (Rotation):

For OCR systems, the point of view for the input image which is taken through camera of hand-held device and other gadgets which is used for taking image is not fixed like a scanner input, which skewing of text lines from their unique orientation might be noticed. Poor results with great degree will be taken in observation when such a skewed image is fed to the OCR classifier. There are various techniques available for the purpose of deskew the image documents, like Projection Profile, RAST algorithm, Hough transform, methods of Fourier transformation, etc.

Blurring and Degradation:

As working over a variety of distances are intended to various digital cameras, a critical factor is the digital camera’s focusing. To get the best accuracy of character recognition and character segmentation, character sharpness is needed. With large apertures and short distances, uneven focus can be seen when there is a change in small point of view. For the most part connected with photography, there consists two kinds of obscure which is: out of focus obscure and movement obscure. When catching a moving item, when the shade rate of the camera is not sufficiently high, the sensor observes a continually changing scene. Accordingly, blurring will be observed in parts in motion.

Multilingual Environments:

Albeit a large portion of the languages of Latin consist of many characters, languages for example, Japanese, Chinese and Korean, includes a large number of character classes. Connected characters exists with the Arabic languages, that according to context causes in the changing of writing shape. In Hindi syllables represent by combining alphabetic letters into thousands of shapes. In multilingual situations, OCR in scanned documents stays as a primary research issue, since OCR in complex symbolism is more troublesome.

2.2EXISTING SYSTEM

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.

DRAWBACKS OF EXISTING SYSTEM:

The drawback in the early OCR systems is that they only have the capability to convert and recognize only the documents of English or a specific language only. That is, the older OCR system is uni-lingual.

2.3PROPOSED SYSTEM

Our proposed system is OCR on a grid infrastructure which is a character recognition system that supports recognition of the characters of multiple languages. This feature is what we call grid infrastructure which eliminates the problem of heterogeneous character recognition and supports multiple functionalities to be performed on the document. The multiple functionalities include editing and searching too where as the existing system supports only editing of the document. In this context, Grid infrastructure means the infrastructure that supports group of specific set of languages. Thus OCR on a grid infrastructure is multi-lingual.

BENEFITS 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.

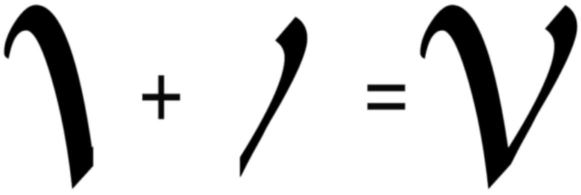
2.4CONCLUSION

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 85 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.

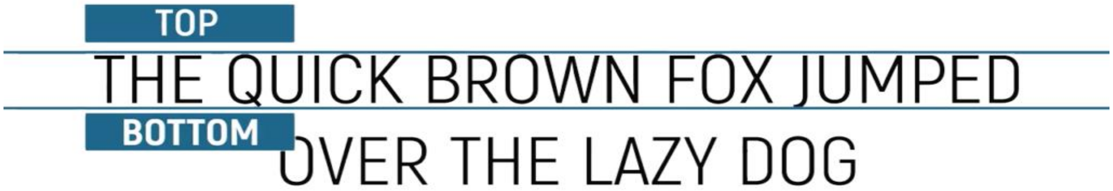
LIST OF FIGURES

1.



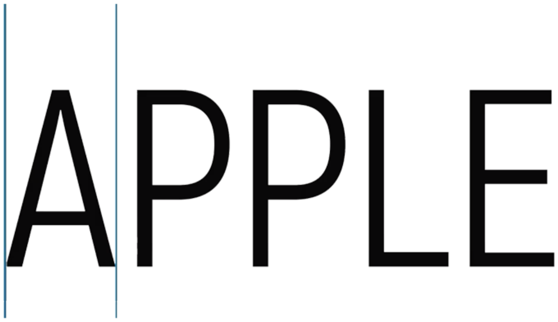
Method 1 – Feature detection.

2.



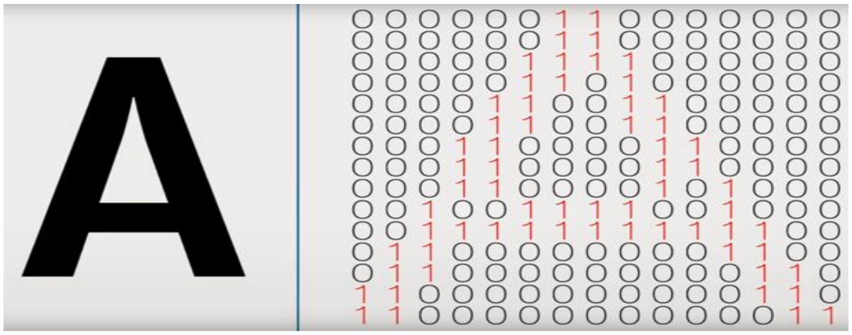
Method 2 – Pattern recognition on a row of text.

3.



Method 2 – Pattern recognition on a single character.

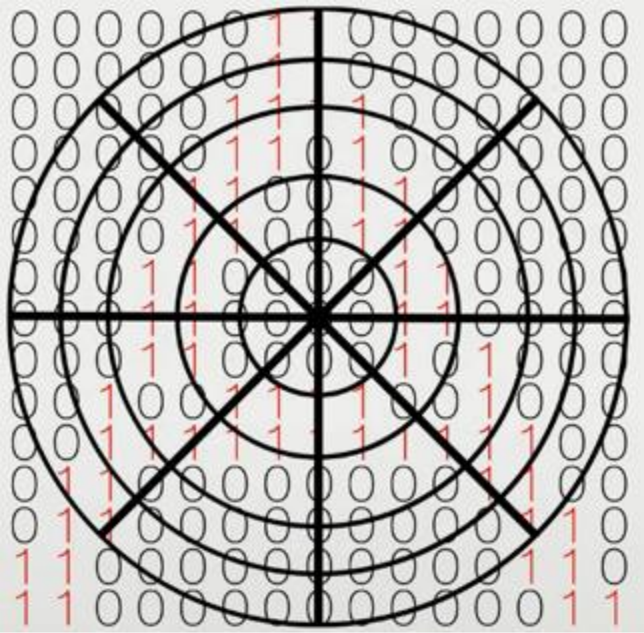
4.



```
0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
0 0 0 0 0 1 1 1 1 0 0 0 0 0 0
0 0 0 0 0 1 1 0 1 0 0 0 0 0 0
0 0 0 0 1 1 0 0 1 1 0 0 0 0 0
0 0 0 0 1 1 0 0 1 1 0 0 0 0 0
0 0 0 1 1 0 0 0 0 1 1 0 0 0 0
0 0 0 1 1 0 0 0 0 1 1 0 0 0 0
0 0 0 1 1 0 0 0 0 1 0 1 0 0 0
0 0 1 0 0 1 1 1 1 0 0 1 0 0 0
0 0 1 1 1 1 1 1 1 1 1 1 1 0 0
0 1 1 0 0 0 0 0 0 0 0 1 1 0 0
0 1 1 0 0 0 0 0 0 0 0 0 1 1 0
1 1 0 0 0 0 0 0 0 0 0 0 1 1 0
1 1 0 0 0 0 0 0 0 0 0 0 0 1 1
```

Sample of binary matrix.

5.



Compare each subsection against the matrix database.