Extraction of Text & Recognition of Images Using Segmentation & OCR in Matlab

Miss. Rakhshi Fatma Ansari[1], Miss. Namita Mishra[2], Mr. Santosh K. Singh[3]
Post Graduate Student[1, 2], Professor3
Department of IT,
Thakur College of Science & Commerce, Thakur Village, Kandivali (East)
Mumbai-400101, Maharashtra, India

Abstract: Text Extraction plays a major role in finding vital and valuable information. Text extraction involves detection, localization, tracking, binarization, extraction, enhancement and recognition of the text from the given image. These text characters are difficult to be detected and recognized due to their deviation of size, font, style, orientation, alignment, contrast, complex colored, textured background. Due to rapid growth of available multimedia documents and growing requirement for information, identification, indexing and retrieval, many researches have been done on text extraction in images. Several techniques have been developed for extracting the text from an image. Text, which carries high-level semantic information, is a kind of important object that is useful for this task. When a machine generated text is printed against clean backgrounds, it can be converted to a computer readable form (ASCII) using current optical character recognition (OCR) technology. However, text is often printed against shaded or textured backgrounds or is embedded in images. Examples include maps, photographs, advertisements, videos, etc. Current document segmentation and recognition technologies cannot handle these situations well.

Index Terms: **Text Extraction, Image Text, Segmentation, OCR Technique, Document Text Images, Connected Components**

I. INTRODUCTION

In this paper a new system is proposed which extracts text in images. The system takes colored images as input. It detects text on the basis of certain text features: text possesses certain frequency and orientation information; text shows spatial cohesion—characters of the same text string (a word, or words in the same line) are of similar heights, orientation, and spacing. The image is then cleaned up so that the text stands out.

Data and Information these days are widely and vividly available in the form of pictures and videos. The current technology is restricted to extracting text against clean backgrounds. Thus, there is a need for a system to extract text from general backgrounds. There are various applications in which text extraction is useful. These applications include digital libraries, multimedia systems, Information retrieval systems, and Geographical Information systems. The role of text detection is to find the image regions containing only text that can be directly highlighted to the user or fed into an optical character reader module for recognition.

The information from image documents should be converted into text in order to get efficient use and access of it like archiving or reporting that are used in different image based applications such as office works. Image Segmentation and Extraction has been performed for the basics further converting the normal image to gray scale image and binary image to get the desired output i.e. extracting text from the image.

II. LITERATURE REVIEW

Few types of tips and techniques have been used in extracting textual content from the images. Objects containing lesser than the desired pixels are removed in this technique. Properties of image regions and plot bounding regions are measured leading to extracting objects from the image. Text recognition in images is an active research area which attempts to develop a computer application with the ability to automatically read the text from images. Nowadays there is a huge demand of storing the information available on paper documents in to a computer readable form for later use.

One simple way to store information from these paper documents in to computer system is to first scan the documents and then store them as images. However 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 challenges involved are: font characteristics of the characters in paper documents and quality of the images. Due to these challenges, computer is unable to recognize the characters while reading them. Thus, there is a need of character recognition mechanisms to perform document image analysis which transforms documents in paper format to electronic format.

In this paper, we have reviewed and analyzed different methods for text recognition from images. The objective of this review paper is to summarize the well-known methods for better understanding of the reader.

The researchers here have used the OCR technique as well to extract text from images. Optical character recognition or optical character reader (OCR) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast).

Techniques like pre-processing, character recognition, post- processing and application specific optimization are entertained in the OCR Method.

III. METHODOLOGY

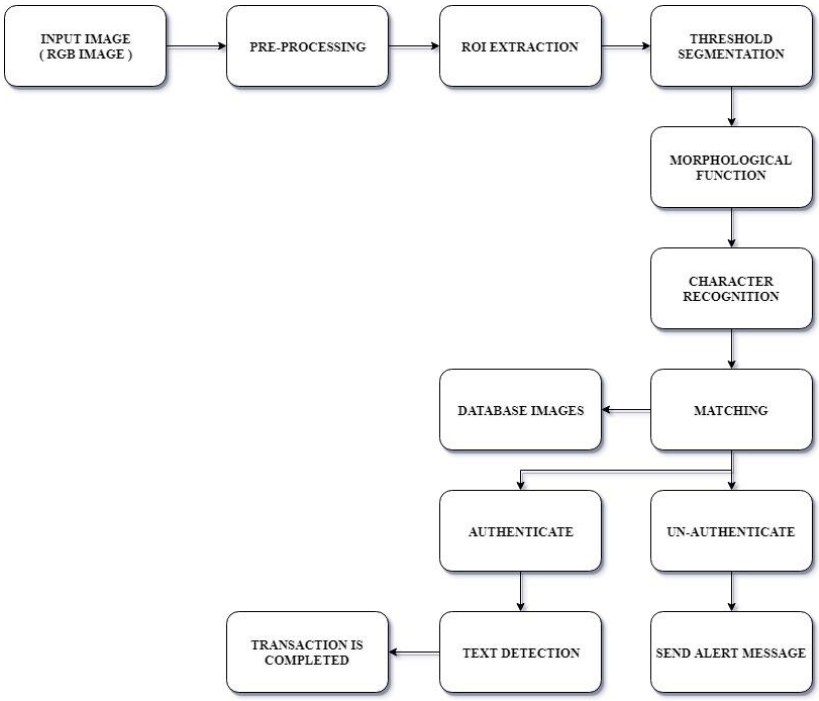


Fig: The stages of extraction of characters from the input.

A. Converting colored image to grayscale: A digital color image is a color image that includes color information for each pixel. There are various color models which are used to represent a color image. These are RGB color model, in which red, green and blue light is added together in various ways to reproduce a broad array of colors. Grayscale images have range of shades of gray without apparent color. These are used as less information needs to be provided for each pixel. In an 8 bit image, 256 shades are possible. The darkest possible shade black is represented as 00000000 and lightest possible shade white is represented as 11111111.

B. Binarization: A Binary image is a digital image that can have only two possible values for each pixel. Each pixel is stored as single bit 0 or 1. The name black and white is often used for this concept. To form a binary image we select a threshold intensity value. All the pixels having intensity greater than the threshold value are changes to 0 (black) and the pixels with intensity value less than the threshold intensity value are changed to 1 (white). Thus the image is changed to a binary image.

C. Connection between Components: For two pixels to be connected they must be neighbors and their gray levels must specify a certain criterion of similarity. For example, in a binary image with values 0 and 1, two pixels may be neighbors but they are said to be connected only if they have same values. A pixel p with coordinates (x, y) has four horizontal and vertical neighbors known as 4-neighbors of p, given as: (x, y+1), (x, y-1), (x+1, y), (x-1, y) and four diagonal neighbors given as: (x+1, y-1), (x-1, y-1), (x+1, y+1), (x-1, y+1). Together these are known as 8-neighbors of p. If S represents subset of pixels in an image, two pixels p and q are said to be connected if there exists a path between them consisting entirely of pixels in S. For any pixel p in S, the set of pixels that are connected to it in S is called a connected component of S.

D. Projections: The method is performed on binary images. It starts scanning from left side of every line and records a change in case of facing the pixel change from zero to one and again to zero. Counting the change does not depend on number of pixels in this method. The text regions will have larger number of transitions from black to white or vice versa, whereas the background region will have lesser number of transitions. If the allocated amount of changes for each row is between two thresholds (low and high thresholds), the row potentially would be considered as text area and the up and down of this row would be specified.

Next, we search vertically for finding the exact location of the text and ignoring these rows as a text. For finding the exact location of the text, we use some heuristics. These heuristics include height and length of the text and the ratio of height to length and enough number of pixels in this area.

E. Reconstruction: After the extraction of text regions from images, the text regions become a bit distorted and difficult to read, thus we recover these components using the original image. The distorted and original images are compared with each other and the pixels which are erased or disfigured are recovered. In case of the OCR technique the text components can further be seen on an excel spread sheet to make it more compatible and clear in understanding.

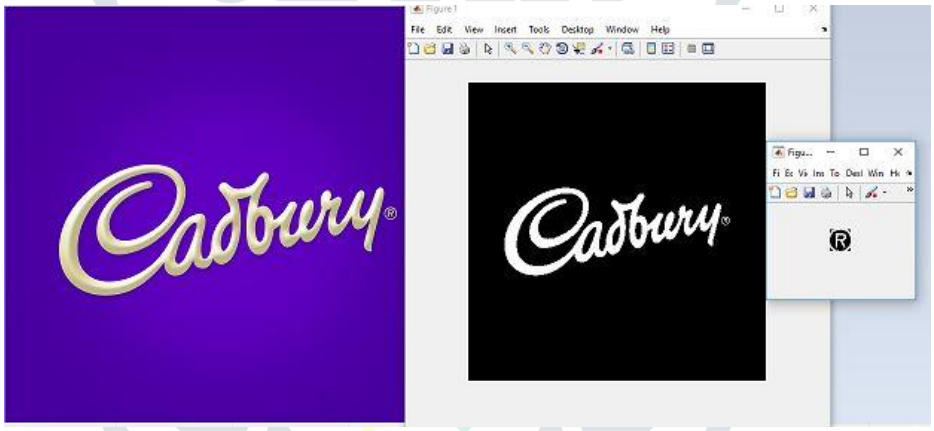
IV. RESULTS

4.1. Method – I : (Image Segmentation and Extraction):

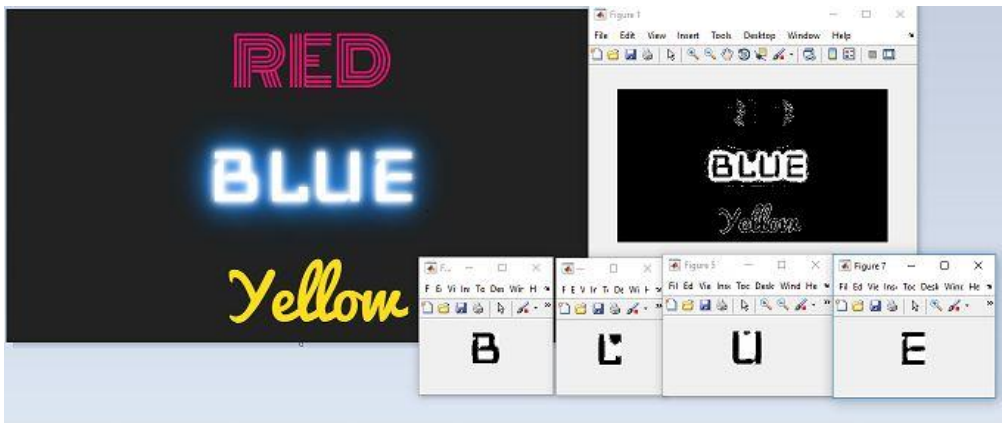
Images of the first method is shown below, which will make it easier to understand on how an image is converted into a gray scale image and binary image and how text is abstracted from it.

i. Before and After:

ii. Before and After:



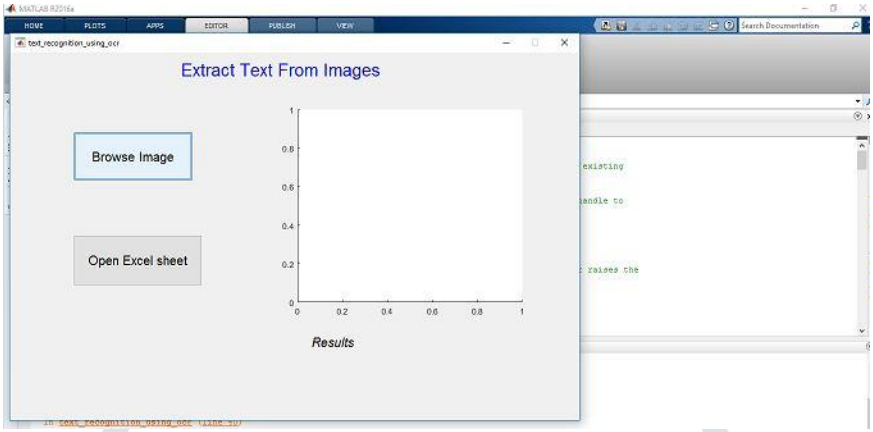
iii. Before and After:



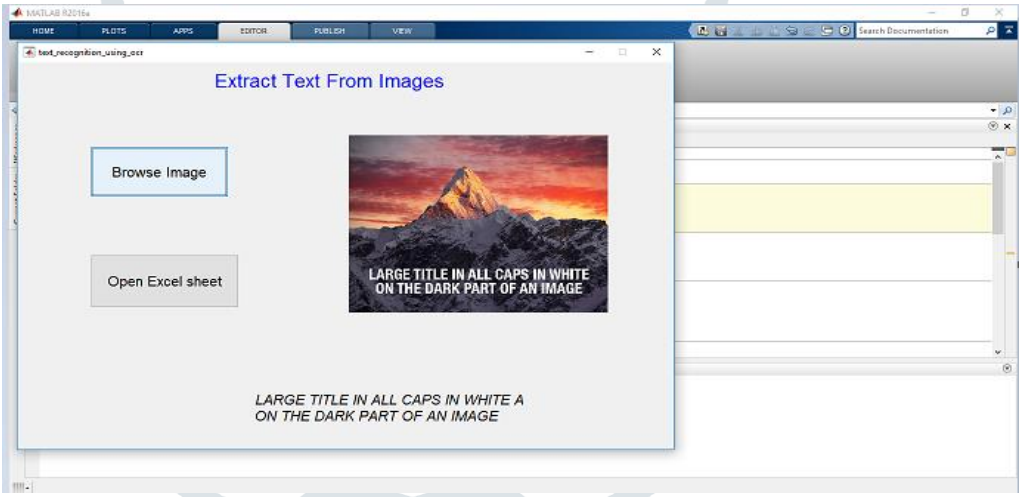
4.2. Method – II : OCR Technique (Optical Character Recognition):

Images of the second method is shown below, which will make it easier to understand on how an image is recognized and the text are extracted from the image and further characterized into clear text expression on an Excel sheet.

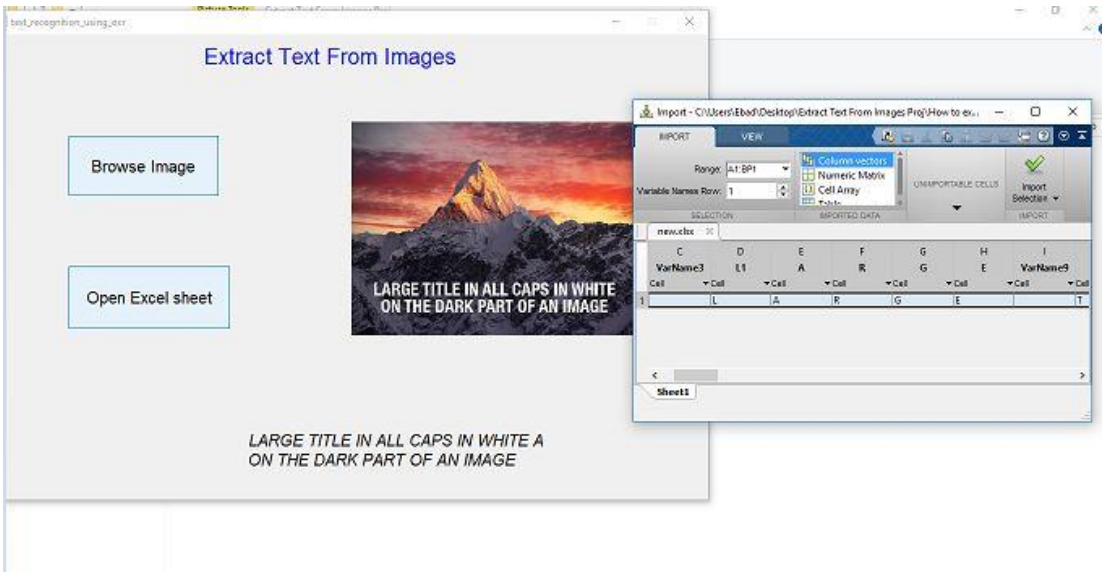
i. **The image selection window appears on the screen, click on "Browse Image", then select any image of your choice, the "result" window shows the result of the image, as shown below:**



ii. **The result after selection of an image, also the text extracted from the images can be seen below:**



iii. **The image appears on the screen, the text extracted from the image appears on the bottom of the window. Then click on "Open Excel Sheet", the text extracted then appears in the Excel sheet as shown in the small window.**



V. CONCLUSION

It was a major task to extract text from images before the invention of these techniques like segmentation and extraction of text from images. This problem has now been solved up to great extent. There are many applications of a text extraction such as Keyword based image search, text based image indexing and retrieval , document analysis, vehicle license detection and recognition, page segmentation ,technical paper analysis, street signs, name plates, document coding, object identification, text based video indexing, video content analysis etc. A number of methods have been proposed in the past for extraction of text in images. These approaches considered the different attributes related to text in an image such as of size, font, style, orientation, alignment, contrast, color, intensity, connected-components, edges etc. The OCR technique is henceforth useful for visually impaired or differently abled users where they can simply now get the text extracted from the digital world and get interacted with it in the real world.

REFERENCES

[1] Ch. Md Mizan, T. Chakraborty* and S. Karmakar, "Text Recognition using Image Processing", International Journal of Advanced Research in Computer Science (IJARCS), 2017, pp. 765-768

[2] Thai V. Hoang , S. Tabbone(2010),"Text Extraction From Graphical Document Images Using Sparse Representation"in Proc. Das, pp 143–150.

[3] Sachin, Grover,Kushal Arora,,Suman K. Mitra(2009),"Text Extraction From Document Images Using Edge Information",IEEE India Council Conference.

[4] Y. Zhan, W. Wang, W. Gao (2006), "A Robust Split-And-Merge Text Segmentation Approach For Images", International Conference On Pattern Recognition,06(2):pp 1002-1005.

[5] Davod Zaravi, Habib Rostami, Alireza Malahzaheh, S.S Mortazavi(2011)," Journals Subheadlines Text Extraction Using Wavelet Thresholding And New Projection Profile", World Academy Of Science, Engineering And Technology .Issue 73.

[6] Arvind, M. Rafi, "Text Extraction from Images Using Connected Component Method" JoAIRA, STM Journal, 2014, 13-18.

[7] A.A. Panchal, Sh. Varde, M.S. Panse, "Character Detection and Recognition System for Visually Impaired People", IEEE, International Conference on Recent Trends in Electronics Information Communication Technology, 2016, pp.1492-1496.

[8] MathWorks, Inc., "Image Processing Toolbox™ User'sGuide", http://in.mathworks.com, September 24, 2017.