International Journal of Engineering and Management Research
e-ISSN: 2250-0758 | p-ISSN: 2394-6962
https://ijemr.vandanapublications.com

Peer Reviewed & Refereed Journal
Volume-14, Issue-2 (April 2024)
https://doi.org/10.5281/zenodo.11171827

Optical Character Recognition from Images

Angel[1], Jean Jisha .M[2] and Vijayalakshmi Shivkhumar[3]
[1]St. Joseph's University, Bengaluru, INDIA
[2]St. Joseph's University, Bengaluru, INDIA
[3]St. Joseph's University, Bengaluru, INDIA

[2]Corresponding Author: jeanjisha31@gmail.com

Received: 23-03-2024 Revised: 07-04-2024 Accepted: 27-04-2024

ABSTRACT

Analysis of document images for information extraction has become very prominent in recent past. Wide variety of information, which has been conventionally stored on paper, is now being converted into electronic form for better storage and intelligent processing. This needs processing of documents using image analysis, processing methods. This article provides an overview of various methods used for digital image processing using three main components: Pre-processing, Feature extraction and the Classification. Pre-processing feature extraction and classification. Classification is an important step in Office Automation, Digital Libraries, and other document image analysis applications.

***Keywords*—** Optical Character, Images, Digital Libraries

I. INTRODUCTION

The aim of document scanner text recognition is to classify optical patterns (often contained in a digital image) corresponding to alphanumeric or other characters. The process of scanner involves several steps including segmentation, feature extraction, and classification.

a. **Aim**: To scan the image and to retrieve the text from the given image
b. **Objective**: The objective of Document Image analysis is to recognize the text & graphics components in image of documents & to extract intended information from them. Two categories of document image analysis can be defined.

II. TEXT PROCESSING

Deals with the textual components of a document image & its task are;

- Determining the skew (any tilt at which the document may have been scanned in the computer).

- Finding columns, paragraphs, textual lines, words, recognizing the text (Possibly its attributes such as size, font etc.) by scanner

III. PRELIMINARY CONCEPTS

Line detection from image

- Start scanning the image horizontally from the topmost left corner row by row.
- if any black pixel is encountered in a row make the row status as '0'.
- If no black pixel in encountered in a row while tracing it then marks the row status as '1'.
- By counting and following the total numbers of continuous '0' from row status vector number and position of lines can be obtained

Character detection from the line

- Take a single line under consideration.
- Start scanning the image vertically from the topmost left corner column by column.
- If any black pixel is encountered in a column mark the column status to '0'.
- If no black pixel in encountered in a column while tracing it then marks the column status as '1'.
- By counting and following the total numbers of continuous '0' from column status vector number and position of lines can be obtained.

IV. LITERATURE SURVEY

Text recognition from images is still active research in the field of pattern recognition. To address the issues related to text recognition many researchers have proposed different technologies, each approach or technology tries to address the issues in different manner.

Automatic License Plate Recognition (ALPR): Badawy, W. et al. has discussed the Automatic license plate recognition (ALPR) is the extraction of vehicle license plate information from an image or a sequence of images. extracted information can be used with or without a database in many applications, such as electronic payment systems (toll payment, parking fee payment), and free away and arterial monitoring systems for traffic surveillance The ALPR uses either a color, black and white, or infrared camera to take images.

Error detection: Sankaran et al. has proposed a novel recognition approach that result in a 15% decrease in word error rate on heavily degraded Indian language document images.

Improve Optical Character Recognition Using Templates & Correlation: Yang et al. has proposed a novel adaptive binarization method based on wavelet filter is proposed. This approach was processes faster, so that it is more suitable for real-time processing and applicable for mobile devices. They evaluated this adaptive method on complex scene images of ICDAR 2005 database.

Document Image Binarization: Ntirogiannis et al. has studied that the document image binarization is of great importance in the document image analysis and recognition pipeline since it affects further stages of the recognition process. The evaluation of a binarization method aids in studying its algorithmic behavior, as well as verifying its effectiveness, by providing g qualitative and quantitative indication of its performance. They proposed a pixel-based binarization evaluation methodology for historical handwritten/machine-printed document images.

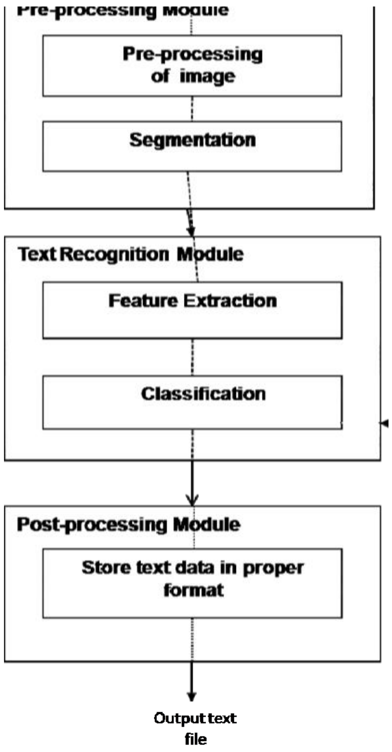
Automated optical character recognition: Gur et al. has discussed some problems in text recognition and retrieval. Automated optical character recognition (OCR) tools do not supply a complete solution and in most cases human inspection is required. They suggest a novel text recognition algorithm based on usage of fuzzy logic rules relying on statistical data of the analyzed font. The new approach combines letter statistics and correlation coefficients in a set of fuzzy based rules, enabling the recognition of distorted letters that may not be retrieved d otherwise. They focused on rashi fonts associated with commentaries of the bible that are actually handwritten calligraphy.

Optical Character Recognition System: Jawahar et al. has proposed a recognition scheme for the Indian script of Devanagari. They used a approach does not require word to character segmentation, which is one of the most common reason for high word error rate. They have been reported a reduction of more than 20% in w word error rate and over 9% reduction in character error rate while comparing with the best available OCR system.

V. PROPOSED ARCHITECTURE

In this section we describe the overall architecture of Text recognition system in the document. A Text recognition system receives an input in the form of document which contains some text information. The output of this system is in electronic format i.e., text information in image is stored in computer readable from our text recognition system divided in following module

- Pre-processing Module
- Text Recognition Module
- Post-processing Module



1. Pre-Processing Module

The Paper document is generally scanned by the optical scanner and is converted in to the form of a picture. A picture is the combinations of picture elements which are also known as pixels. The pixels contain basically two values ON and OFF. The ON value points that's the pixel is visible and the OFF-value points that's the pixel is not visible. At this stage we have the data in the form of image and this image can be further analyzed so that's the important information can be retrieved. So, to improve quality of the input image and make it suitable for further analysis, we perform some operation on it such as Grayscale conversion, Binary image conversion and the most important is segmentation. In this we perform some operation on scan image such as:

Pre-Processing of Images

a. Scanning printed documents and storing the documents as snapshots or images.
b. Processing those image-based documents, Converting these image-based documents into proper format such as Greyscale and Binary format.

Segmentation

The segmentation is the most important process in document scanner. Segmentation is done to make the separation between the individual characters of an image. Segmentation is one of the most important phases in this project. The performance of this project is depending on segmentation. Segmentation subdivides an image into its constituent regions or objects. Basically, in segmentation, we try to extract basic constituent of the script, which are certainly characters. This is needed because our classifier recognizes these characters only. We perform the segmentation of character from image by applying Line detection and Character detection.

2. Text Recognition Module

This module can be used for text recognition in output image of pre-processing model and give output data which are in computer understandable form. Hence in this module following techniques are used.

Feature Extraction

Feature extraction is the process to retrieve the most important data from the raw data. The most important data means that’s on the basis of that’s the characters can be represented accurately. To store the different features of a character, the different classes are made. There are many techniques used for feature extraction like Principle Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Chain Code (CC), zoning, Gradient Based features, Histogram etc.

In this we use matrix feature extraction method. In this method first we convert the image to binary matrix i.e., black and white image convert to matrix form, text image is converted in to the matrix of 0’s and 1’s from this matrix data. We extract text character line by line and word by word by using above segmentation method. After that segmented characters data are normalized and store in fixed dimension as a feature of that character

Classification

The classification is the process of identifying each character and assigning to it the correct character class, so that texts in images are converted in to computer understandable form. This process used extracted feature of text image for classification i.e., input to this stage is output of the feature extraction process. Classifiers compare the input feature with stored pattern

In this we use Artificial Neural Network (ANN) for classification because neural network can get itself trained automatically on the basis of efficient tools for learning large databases and examples. This approach is non algorithmic and trainable. There are the different types of neural networks which can be used for the classification from which we used Kohonen neural network.

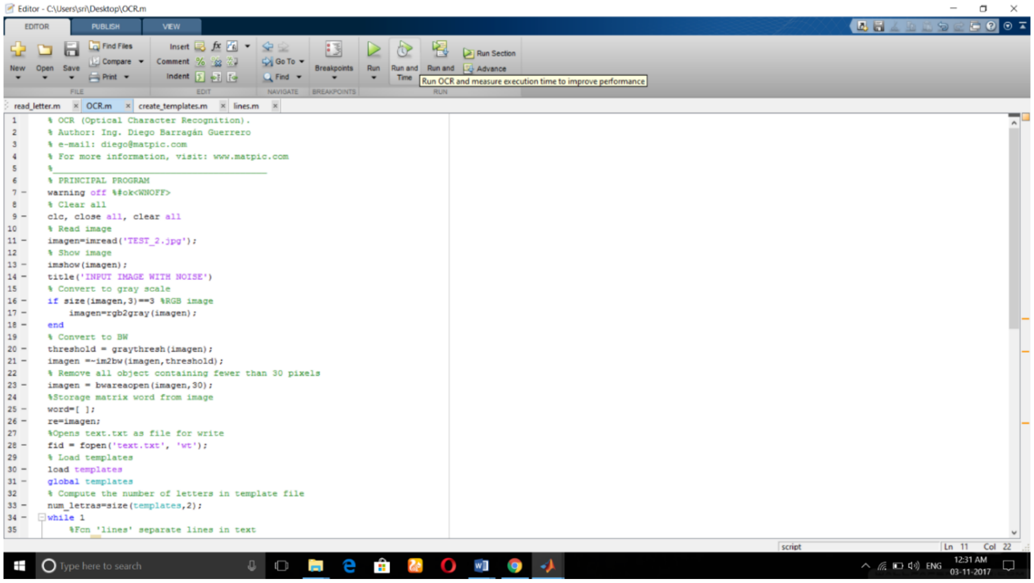
3. Post-Processing Module

The output of Text and find out best matching class for input.

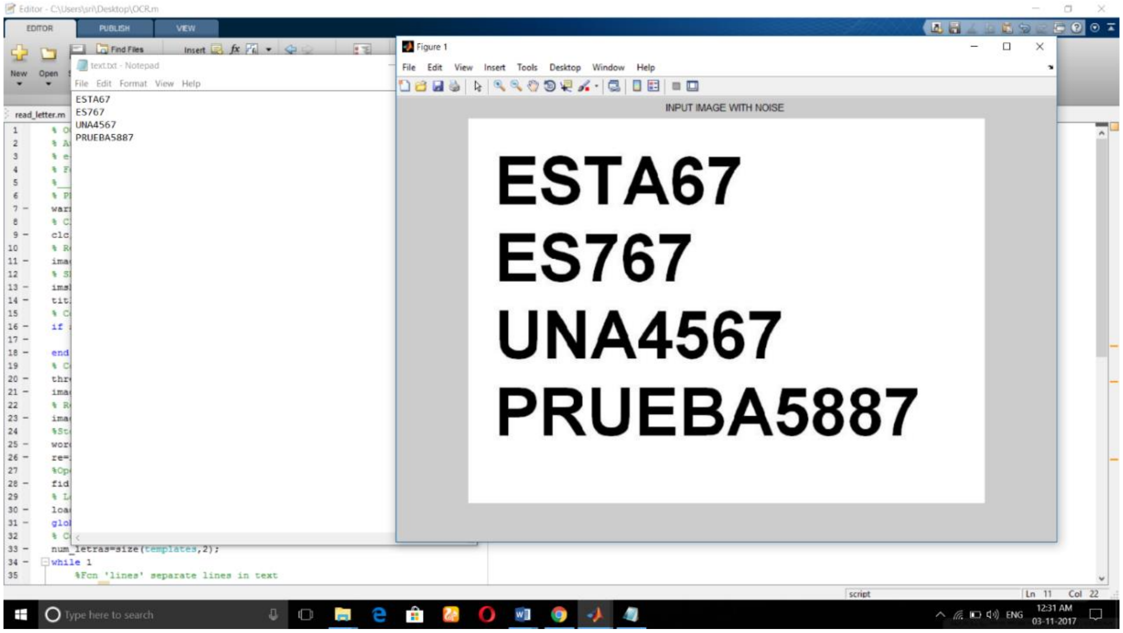
Recognition Module is in the form text data which is understand by computer, so there need to store it in to some proper format (i.e., txt or MS-Word) for farther use such as Editing or Searching in that data.

VI. EXPERIMENTAL RESULT

The Paper document is generally scanned by the optical scanner and is converted in to the form of a picture. A picture is the combinations of picture elements which are also known as pixels. At this stage we have the data in the form of image and this image can be further analyzed so that’s the important information can be retrieved. So, we apply our method of text recognition which disused in this paper and output results are shown in the form of following images.

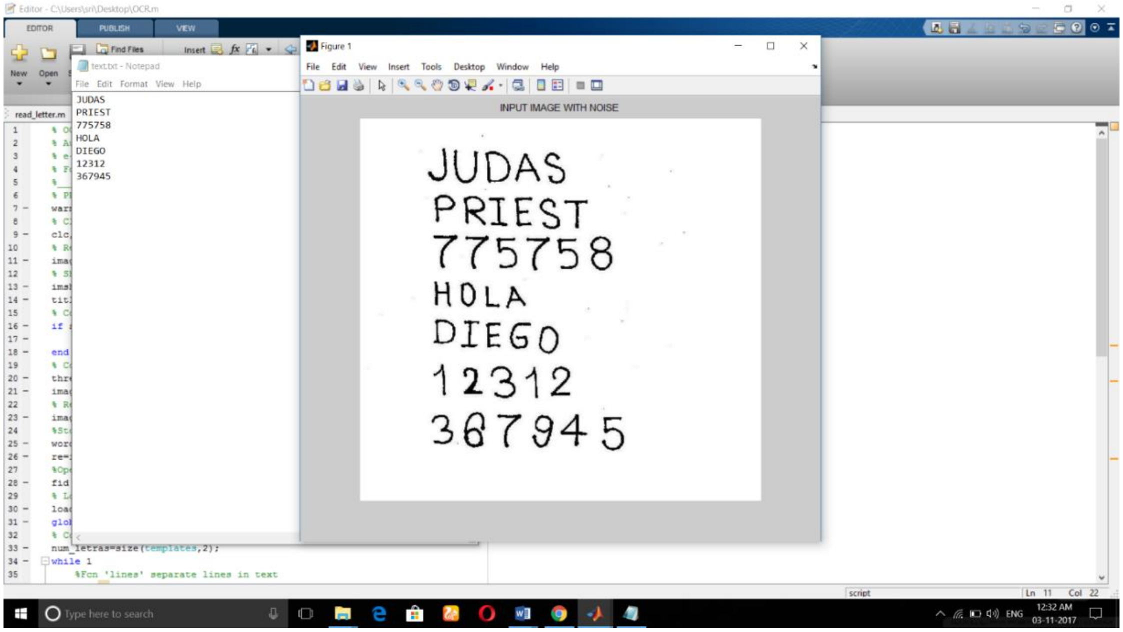


```
% OCR (Optical Character Recognition).
% Author: Ing. Diego Barragán Guerrero
% e-mail: diego@matpic.com
% For more information, visit: www.matpic.com
%________________________________________
% PRINCIPAL PROGRAM
warning off %#ok<WNOFF>
% Clear all
clc, close all, clear all
% Read image
imagen=imread('TEST_2.jpg');
% Show image
imshow(imagen);
title('INPUT IMAGE WITH NOISE')
% Convert to gray scale
if size(imagen,3)==3 %RGB image
    imagen=rgb2gray(imagen);
end
% Convert to BW
threshold = graythresh(imagen);
imagen =~im2bw(imagen,threshold);
% Remove all object containing fewer than 30 pixels
imagen = bwareaopen(imagen,30);
%Storage matrix word from image
word=[ ];
re=imagen;
%Opens text.txt as file for write
fid = fopen('text.txt', 'wt');
% Load templates
load templates
global templates
% Compute the number of letters in template file
num_letras=size(templates,2);
while 1
    %Fcn 'lines' separate lines in text
```



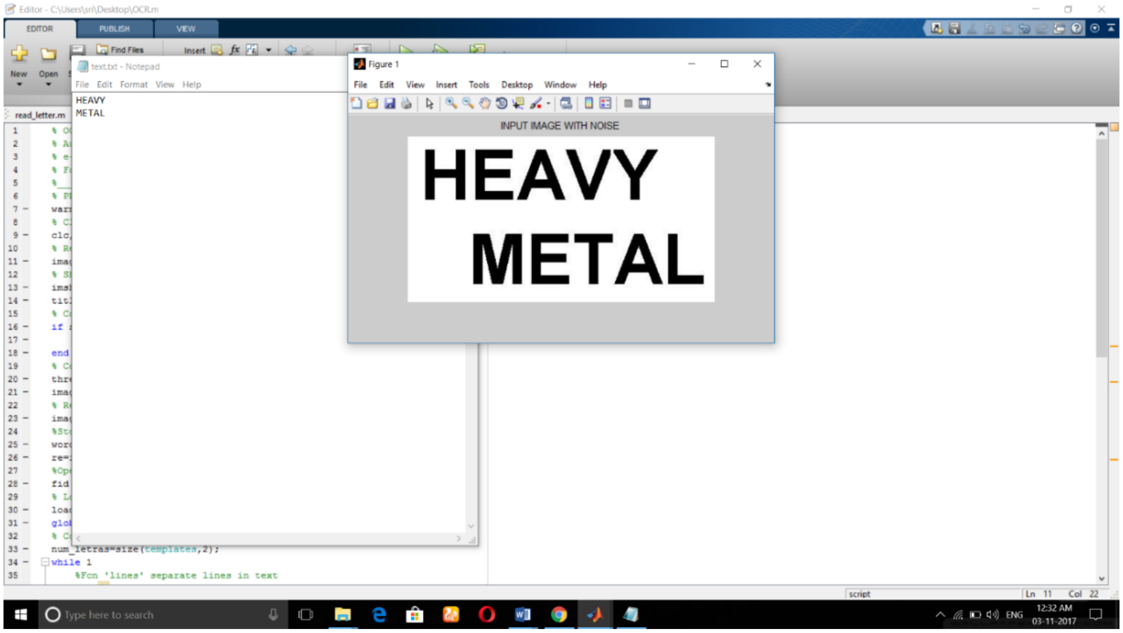
```
ESTA67
ES767
UNA4567
PRUEBA5887
```

```
num_letras=size(templates,2);
while 1
    %Fcn 'lines' separate lines in text
```



```
JUDAS
PRIEST
775758
HOLA
DIEGO
12312
367945
```

```
num_letras=size(templates,2);
while 1
    %Fcn 'lines' separate lines in text
```



```
HEAVY
METAL
```

INPUT IMAGE WITH NOISE

HEAVY
METAL

```
num_letras=size(templates,2);
while 1
    %Fcn 'lines' separate lines in text
```

VII. FUTURE WORK

In this paper, we retrieved a foreground text from complex document. Next, we are going to extract from image document.

VIII. CONCLUSION

We proposed and discussed method document scanner text recognition. wide area for researcher in pattern recognition. A lot of research work has been done and is still being done in document scanner text recognition for various languages. More and more researchers are attracted to this challenging field. Each stage optical character recognition has its own significance and should be designed properly for better results.

REFERENCES

[1] F. Fisher. (2001). Digital camera for document acquisition. In: *Proc. Symposium on Document Image Understanding Technology*, pp. 75–83.

[2] D. Doermann, J. Liang & H. Li. (2003). Progress in camera based document image analysis. In: *Seventh International Conference on Document Analysis and Recognition*, pp. 606–616. IEEE.

[3] G. Strang. (2003). *Introduction to linear algebra*. Cambridge Publication.

[4] C. Harris & M. Stephens. (1988). A combined corner and edge detector. In: *Alvey Vision Conference*, *15*, pp. 50, Manchester, UK.

[5] Distributed Desktop Conferencing System (MERMAID) Based on Group Communication *Architecture Kazutoshi MAENO, Shiro SAKATA and Toyoko. Ohmori Network Research Laboratory C&C Systems Research Laboratories Nec Corporation*.