Impact Factor (SJIF): 4.542

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 4, Issue 4, April-2017

A Simple Effective Algorithm on Text Recognition System

Ronakkumar Pravinbhai Patel[1]

[1]*Computer Department, B & B institute of Technology, VallabhVidyanagar, Anand, Gujarat*

Abstract — Text recognition system is read text from images and store text for future use. It is oldest research area in image processing. Now a day there is a huge demand in storing the information available in paper documents format in to a computer storage disk and then later reusing this information by searching process. When we want store information as image in our computer system, first scan the documents and then store them as images. But to reuse this information it is very difficult. We will read the individual contents and searching the contents form these documents line-by-line and word-by-word. Thus there is a need of Text recognition system to perform Document Image Analysis (DIA) which transforms documents in paper format to electronic text format. In this paper we have applied simple and effective technique and getting better result than available recognition method. The challenges involved in this the font characteristics of the texts in paper documents and quality of images.

Keywords- Document Image Analysis(DIA), electronic text format, Text recognition, storage disk, font characteristics.

I. INTRODUCTION

There is growing demand for the software to recognize texts in computer system when information is scanned through paper documents, either modern or historical. These days there is a huge demand in " storing the information available in paper documents format in to a computer storage disk and then later reusing this information by searching process.". During the last decades a lot of research has been done in the field of Optical Character Recognition (OCR). Numerous commercial software have been released that convert digitized documents into text files, usually in ASCII format. Although these software process machine printed documents successfully, when it comes to handwritten documents the results are not satisfactory enough.

There are two types of Recognition method available. Simply it's say online and offline recognition. Online is use of huge database and recognition accuracy is high, but offline method is very difficult. We applied offline method using matlab software. Our method is working on a single background image and it is very effective.

Moreover, such products are unable to process historical documents due to their low quality, lack of standard alphabets and presence of unknown fonts. To this end, recognition of historical documents is one of the most challenging tasks in OCR.

In the literature, historical document processing is mainly focused on document retrieval. Word-spotting techniques for searching and indexing historical documents have been introduced. In [1], word images are grouped into clusters of similar words by using image matching to find similarity. Then, by annotating "interesting" clusters, an index that links words to the locations where they occur can be built automatically. Yang et al.[2] has proposed a novel adaptive binarization method based on wavelet filter is proposed in this paper, which shows comparable performance to other similar methods and processes faster, so that it is more suitable for real-time processing and applicable for mobile devices. The proposed method is evaluated on complex scene images of ICDAR 2005 Robust Reading Competition, and experimental results provide a support for our work.

In [3] and [4] holistic word recognition approaches for historical documents are presented based on scalar and profile-based features and on matching word contours respectively. Their goal is to produce reasonable recognition accuracies which enable performing retrieval of handwritten pages from a user-supplied ASCII query. In [4], a word spotting technique based on combing synthetic data and user feed-back for keyword searching in historical printed documents is described. Sankaran et al. [5] has presented present a novel recognition approach that results in a 15% decrease in word error rate on heavily degraded Indian language document images. OCRs have considerably very good performance on good quality documents, but fail easily in presence of degradations.

Gur et al. [6] has discussed that text recognition and retrieval is a well-known problem. Automated optical character recognition (OCR) tools do not supply a complete solution and in most cases human inspection is required. In this paper the authors suggest a novel text recognition algorithm based on usage of fuzzy logic rules relying on statistical data of the analysed 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 otherwise. The authors focused on Rashi fonts associated with commentaries of the Bible that are actually handwritten calligraphy. Rhead et al. [7] has considered real world UK number plates and relates these to ANPR. It considers aspects of the relevant legislation and standards when applying them to real world number plates. The varied manufacturing techniques and varied specifications of component parts are also noted. The varied fixing methodologies and fixing locations are discussed as well as the impact on image capture. Badawy, W. et al. [8] has discussed the Automatic license plate recognition (ALPR) is the extraction of vehicle license plate information from an image or a sequence of images. The extracted information can be used with or without a database in many applications, such as electronic payment systems (toll payment, parking fee payment), and freeway and arterial monitoring systems for traffic surveillance. The ALPR uses either a color, black and white, or infrared camera to take images.

In this paper, an off-line recognition system for either machine printed. It consists of a pre-processing stage where documents are converted into binary images, a top – down segmentation technique that extracts the texts, the creation of a database by the extracted texts and a recognition stage where the database is used for converting any document into text file. The paper is organized as follows: in Sect. 2, Construction of Text recognition system which contains flow chart of algorithm and output of each phase Finally, Sect. 3, a Conclusion is given.

II. CONSTRUCTION OF TEXT RECOGNITION SYSTEM

In this section we discuss about the overall architecture of Text recognition system. A Text recognition system receives an input in the form of image which contains some text information. The output of this system is in electronic format i.e. text information in image are stored in computer readable form. The Text recognition system can be divided in following module: (A) Pre-processing Module, (B) Character Extraction, (C) Text Recognition Module and (D) Post-processing Module. The overall architecture is shown in figure 1.

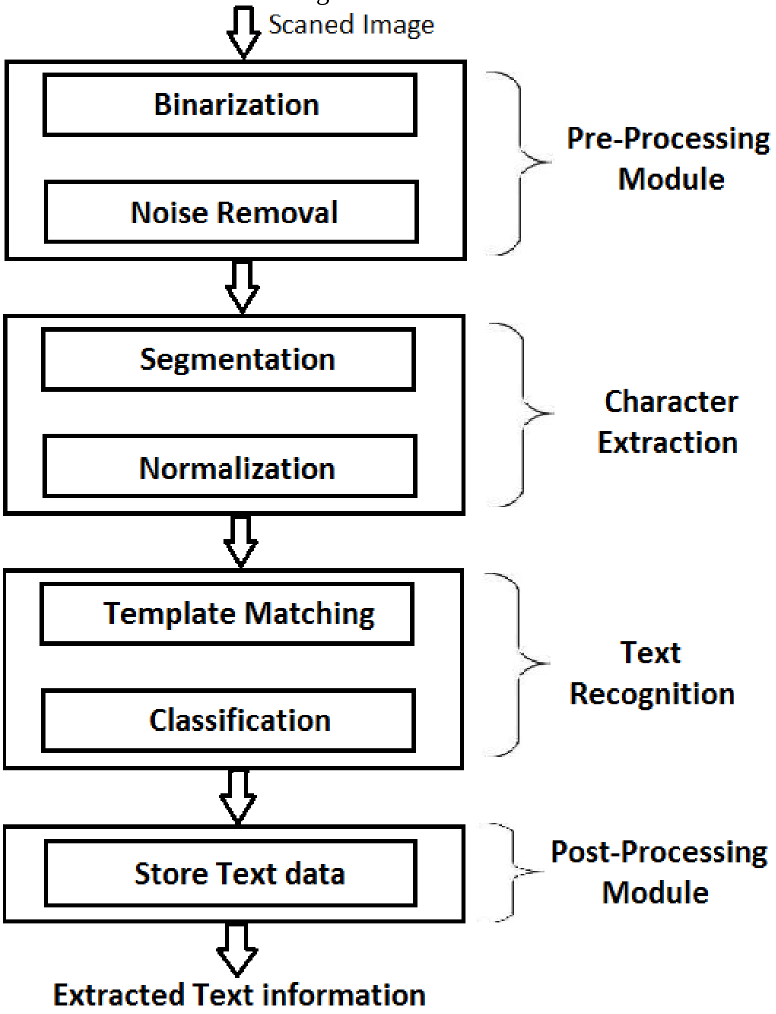


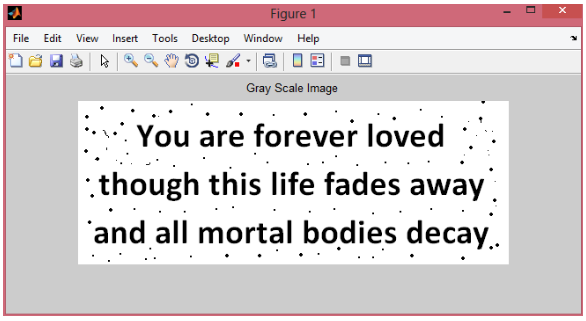
Figure 1 Architecture of proposed algorithm

A) Pre-processing Module

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 analysed so that's the important information can be retrieved. So to improve quality of the input image, we perform some operation on it such as

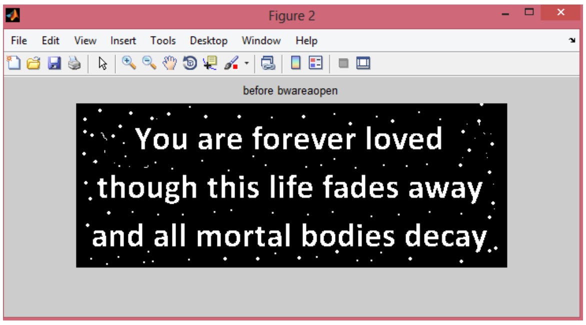
1. Binarization

The image is taken and is converted to gray scale image. The gray scale image is converted to binary image. This process is called Digitization of image (Binarization)[8]. Figure 2 Shows original image contain text. Figure 3 Show the binary image of figure 2.



You are forever loved
though this life fades away
and all mortal bodies decay

Figure 2 Input Original Image

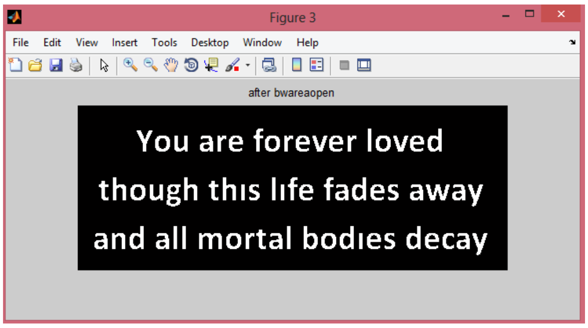


You are forever loved
though this life fades away
and all mortal bodies decay

Figure 3 Binary Image

2. Noise removal

Noise removal is one of the most important processes. Practically any scanner is not perfect; the scanned image may have some noise. This noise may be due to some unnecessary details present in the image. By applying suitable methods the denoised image is produced. Due to this quality of the image will increase and it will affect recognition process for better text recognition in images. And it results in generation of more accurate output at the end of text recognition processing. Figure 4 Shows noise is removed from figure 3.



You are forever loved
though this lıfe fades away
and all mortal bodıes decay

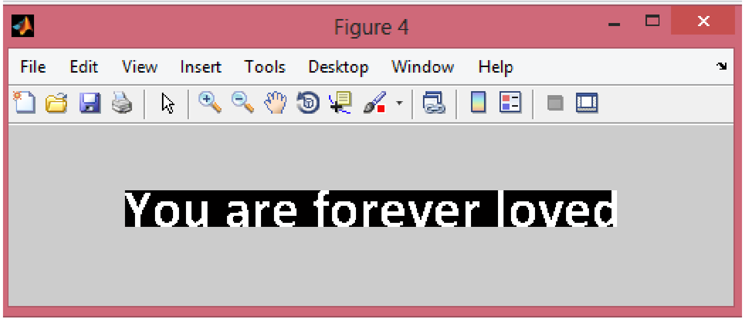
Figure 4 Denoised Image

B) Character Extraction

These are done using two parts Segmentation and Normalization.

1. Segmentation

After pre-processed image find location of text on image. Separate each line and then each character using space from image it is segmentation [10]. Segmentation is done to make the separation between the individual characters of an image. Figure 5 Shows Separate line from figure 4. Figure 6 shows separate character from lines.



You are forever loved

Figure 5 Separated line

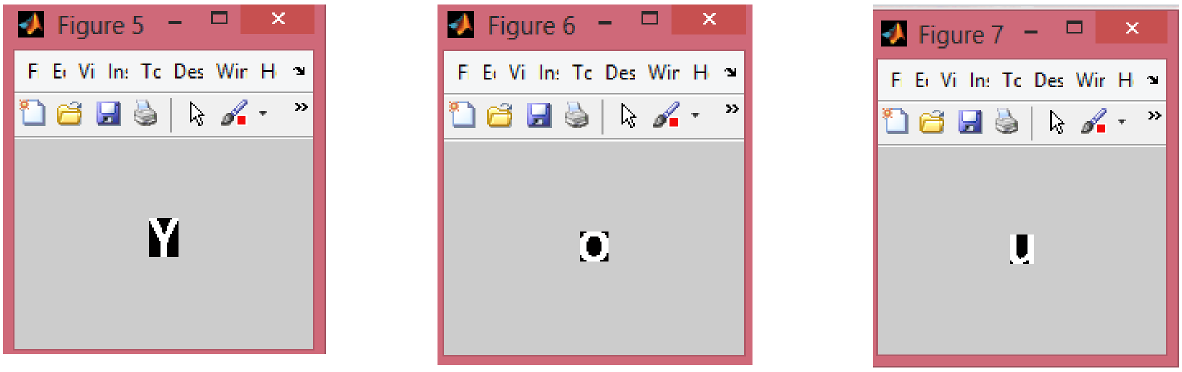


Figure 6 Separated Character from line

2. Normalization

Normalization is one of the important processing operations for text recognition. The normalization is applied to obtain characters of uniform size, slant and rotation. We convert all Separated character into 24x42 image same as template character size. This is important for perfect matching.

C) Text Recognition Module

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

1. Template Matching

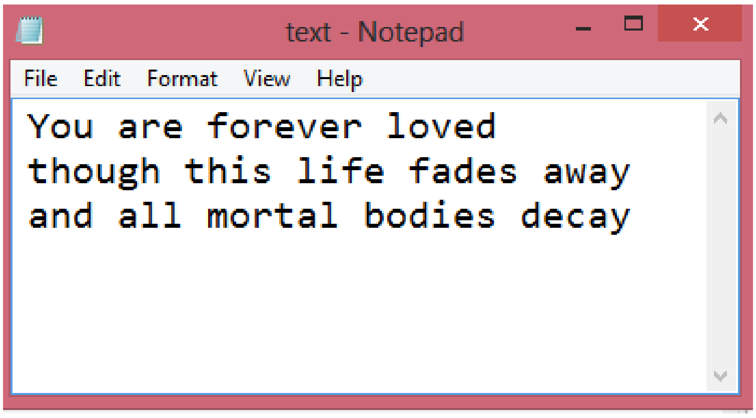
The image from the extraction stage is correlated with all the templates which are preloaded into the system. Once the correlation is completed, the template with the maximum correlated value is declared as the character present in the image [1].

2. Feature Extraction

Another Technique is 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 technique 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.

D) Post-processing Module

The output of Text 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. Figure 7 Shows Output texts.



```
You are forever loved
though this life fades away
and all mortal bodies decay
```

Figure 7 Output Text file

III. CONCLUSION

This paper tells about OCR system for offline Text recognition. The systems have the ability to yield excellent results. Pre-processing techniques used in document images as an initial step in text recognition systems were presented. The feature extraction step of Text recognition is the most important. It can be used with existing OCR methods, especially for English text. This system offers an upper edge by having an advantage i.e. its scalability, i.e. although it is configured to read a predefined set of document formats, currently English documents, it can be configured to recognize new types. Future research aims at new applications such as online Text recognition used in mobile devices, extraction of text from video images, extraction of information from security documents and processing of historical documents.

REFERENCES

[1] T.M.Rath and R. Manmatha, "Word spotting for historical documents", International Journal on Document Analysis and Recognition (IJDAR), Vol.9, No 2 – 4, pp. 139– 152 , 2006.

[2] Yang, Jufeng, Kai Wang, Jiaofeng Li, Jiao Jiao, and Jing Xu, "A fast adaptive binarization method for complex scene images", 19th IEEE International Conference on Image Processing (ICIP), 2012.

[3] V. Lavrenko, T. M. Rath, R. Manmatha: "Holistic Word Recognition for Handwritten Historical Documents", Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04),pp 278-287, 2004.

[4] T. Adamek, N. E. O'Connor, A. F. Smeaton, "Word Matching Using Single-Closed Contours for Indexing Handwritten Historical Documents", International Journal on Document Analysis and Recognition (IJDAR), special Issue on Analysis of Historical Documents, 2006.

[5] Shrey Dutta, Naveen Sankaran, PramodSankar K., C.V. Jawahar, "Robust Recognition of Degraded Documents Using Character N-Grams", IEEE, 2012.

[6] Gur, Eran, and ZeevZelavsky, "Retrieval of Rashi Semi-Cursive Handwriting via Fuzzy Logic", Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. IEEE, 2012.

[7] Rhead, Mke, "Accuracy of automatic number plate recognition (ANPR) and real world UK number plate problems." IEEE International Carnahan Conference on Security Technology (ICCST), 2012.

[8] Badawy, W. "Automatic License Plate Recognition (ALPR): A State of the Art Review." IEEE International Conference on Document Analysis and Recognition, 2012.

[9] "Combination of Document Image Binarization Techniques", 2011 International Conference on Document Analysis and Recognition.

[10] "α-Soft: An English Language OCR", 2010 Second International Conference on Computer Engineering and Applications. Junaid Tariq, Umar Nauman Muhammad Umair Naru.

[11] "A Review on the Various Techniques used for Optical Character Recognition", Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi/International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 1,Jan-Feb 2012.