Survey paper on Text Recognition Using Image Processing

Mr.Rahul R. Patil
Department of E&TC
ICOER
Wagholi, Pune.

Mr.Audumbar R. Misal
Department of E&TC
ICOER
Wagholi, Pune.

Mr.Ketan R. Nalawade
Department of E&TC
ICOER
Wagholi, Pune.

Abstract

The goal of Text Recognition is to recognize the text from printed hardcopy document to desired format (like .docx). The process of Text Recognition involves several steps including preprocessing, segmentation, feature extraction, classification, post processing. Preprocessing is for done the basic operation on input image like binarization which convert gray Scale image into Binary Image, noise reduction which remove the noisy signal from image. Segmentation stage for segment the given image into line by line and segment each character from segmented line. Future extraction calculates the characteristics of character. A classification contains the database and does the comparison. Nowadays it plays an important role in office, colleges etc

*Keywords**-Preprocessing, segmentation, classification.***

I. INTRODUCTION

Nowadays all over digitization technology is used. Text Recognition usually abbreviated to 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. OCR began as a field of research in artificial intelligence and computational vision. Text Recognition used in official task in which the large data have to type like post offices, banks, colleges etc., in real life applications where we want to collect some information from text written image. People wish to scan in a document and have the text of that document available in a .txt or .docx format.

II. PROBLEM STATEMENT

Aim of this paper is to evaluate the character from given image which is text written image and print on document or word file.

III. LITERATURE REVIEW

Preprocessing is the first step in the processing of scanned image. The scanned image is checked for noise, skew, slant etc. There are possibilities of image getting skewed with either left or right orientation or with noise such as Gaussian. Here the image is first convert into grayscale and then into binary. Hence we get image which is suitable for further processing.

After pre-processing, the noise free image is passed to the segmentation phase, where the image is decomposed into individual characters. Fig.3 shows the image and various steps in segmentation. The binarized image is checked for inter line spaces. If inter line spaces are detected then the image is segmented into sets of paragraphs across the interline gap. The lines in the paragraphs are scanned for horizontal space intersection with respect to the background. Histogram of the image is used to detect the width of the horizontal lines. Then the lines are scanned vertically for vertical space intersection. Here histograms are used to detect the width of the words. Then the words are decomposed into characters using character width computation.

Feature extraction follows the segmentation phase of OCR where the individual image glyph is considered and extracted for features. First a character glyph is defined by the following attributes like height of the character, width of the character.

Classification is done using the features extracted in the previous step, which corresponds to each character glyph. These features are analyzed using the set of rules and labeled as belonging to different classes. This classification is generalized such that it works for single font type. The height of the character and the width of the character, various distance metrics are chosen as the

ISSN: 2278 – 909X

candidate for classification when conflict occurs. Similarly the classification rules are written for other characters. This method is a generic one since it extracts the shape of the characters and need not be trained. When a new glyph is given to this classifier block it extracts the features and compares the features as per the rules and then recognizes the character and labels it.

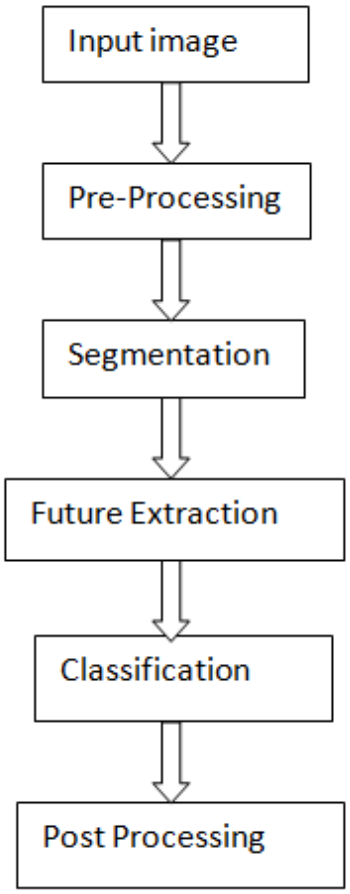

Fig. a) Block Diagram of Text Recognition.

After classification, algorithm will check the classified data with the database where we have already saves the classification of characters, numbers and symbols. Then algorithm will display the result according to the comparison. Here our algorithm will check the pattern of inputted character with database.

IV. ALGORITHMS

1. Start

2. Scan the textual image.

1. Start

2. Scan the textual image.

3. Convert color image into gray image and then binary image.

4. Do preprocessing like noise removal, skew correction etc.

5. Load the DATABASE.

6. Do segmentation by separating lines from textual image.

VI.CONCLUSION

In this paper we proposed algorithm for solving the problem of offline character recognition. We had given the input in the form of images. The algorithm was trained on the training data that was initially present in the database. We have done preprocessing and segmentation and detect the line.

VII.REFERENCES

E-PAPERS:

1] Ramanathan. R. et al., "A Novel Technique for English Font Recognition Using Support Vector Machines", in Advances in Recent Technologies in Communication and Computing, Kottayam, Kerala, 2009, pp. 766-769.

2] Line Eikvil, "Optical Character Recognition", NorskRegnesentral, Oslo, Norway, Rep. 876, 1993.

3] M Usman Raza, et al., "Text Extraction Using Artificial Neural Networks", in Networked Computing and Advanced Information Management (NCM) 7th International Conference, Gyeongju, North Gyeongsang, 2011, pp. 134-137.

4] Araokar, Shashank, 'Visual Character Recognition using Artificial Neural Networks', CoRR, Vol abs/cs/0505016, 2005

ISSN: 2278 – 909X

5] Bertolami, Roman; Zimmermann, Matthias and Bunke, Horst, 'Rejection strategies for offline handwritten text line recognition', ACM Portal, Vol. 27, Issue. 16, December 2006

6] Wikipedia web encyclopedia:

http://en.wikipedia.org/wiki/Main_Page

7] Matt's Mat lab Tutorial Source Code Page:

http://joplin.ucsd.edu/Tutorial/matlab.html

Websites:

1] www.ieee.org.in

2] www.wikipedia.com

ISSN: 2278 – 909X