Font & Style of Text Extraction from Images

Babita Kaithwas, Priyanka Dhanvijay, Shivani Dhone, Ashwin Shinde
CSE & RTMNU, Nagpur, Maharashtra,
India

Abstract— In this project a new and easy approach for extracting the text from image is presented. The main aim of the project is to convert the old printed text into the digitized form so as to increase the durability of the text and reduce the maintenance cost. There are many different approaches to convert the printed text to digitized text but many of these methods are tedious for the user to use. So in this paper we are trying to give easier approach using matrix matching concept, by implementing the project we are able to extract the text in user friendly and efficient manner.

Keywords— OCR, Pre-processing, Recognition , Segmentation.

I. INTRODUCTION

To increase the durability and reduce the maintenance cost of the printed material, it is necessary to convert the text from image into a digitized format. There are many different approaches to convert the printed text to digitized text but many of these methods are tedious for the user to use. So in this project we are trying to give easier approach for text extraction using matrix matching concept.
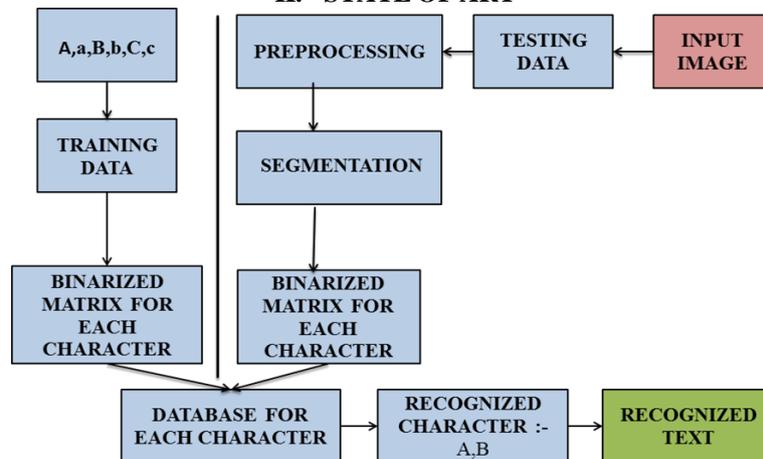
II. STATE OF ART

Fig: Work flow of extraction of text from image

III. TRAINING DATA

The template for the various characters and numbers has been created   i.e dataset in the system so as to train it. This system consist of 4 steps. They are image acquisition, image pre-processing, segmentation of character, feature extraction of characters, character recognition .

- Image Acquisition – The input image is acquired by digital professional camera or scanned copy of an image.
- Image Pre-processing – It is procedure to convert coloured or gray scale image into black and white. Pre-processing the image is very important to make it more accurate for further steps. After the completion of pre-processing a separate database for each character will be created.

IV. TESTING DATA

In this phase image is given as input & again pre-processing will be done. The output of pre-processing phase is provided as an input to segmentation phase.

- Segmentation – Segmentation is the process of decomposition of different objects by extracting their respective boundaries and the text component is isolated from the background.
  In this step the pre-processed image consisting of sequence of character is there by decomposed into sub-image. Firstly the image is segmented line by line, then the line is segmented word by word and further the word is segmented into characters. The segmented character is provided as an input for the next step.

- Recognition – For each input segmented character we create a matrix and then this matrix is matched with the already stored matrix of characters in the dataset. In the recognition step the matrix formed for segmented character are matched with stored matrix of character in dataset. If the matrix is matched then the character is recognized and printed in word file.

V. RESULTS

- Data set-

The data set is created for the matching purpose which is consisting of English alphabets both in lower case and upper case (a-z, A-Z) as well as the numeric (0-9).

Fig.: Data set

Matrix form of each character of data set

For each character in the data set matrix is created, which is further use in recognition step.

VI.   CONCLUSION

The project has successfully able to extract the the from the scanned image. It is necessary or the application that the image must be scanned so as to reduced the noise in the capturing process.

REFERENCE

[1]   Pratibha A. Desai, Sumangala N Bhavikatti, Rajashekar Patil, "Neural Networks Based Offline Handwritten Character Recognition System", SARC-IRAJ International Conference, 16th june 2013

[2]   D.kavitha, P.Shamini, "Handwritten Document into Digitized Text Using Segmentation Algorithm", Special Issue, 4th National Conference on Advanced Computing, Applications & Technologies, May 2014.

[3]   Shilpy Bansal, Mamta Garg, Munish Kumar, "A Technique for Offline Handwritten Character Recognition", IJCAT International Journal of Computing and Technology, Volume 1, Issue 2, March 2014.

[4]   Aparna.A , Prof.I.Muthumani," Optical Character Recognition for Handwritten Cursive English characters", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5, 2014.

[5]   Varun Sanduja , Rajeev Patial," Sobel Edge Detection using Parallel Architecture based on FPGA" , (IJAIS) International Journal of Applied Information Systems , July 2012 .

[6]   Muhammad Naeem Ayyaz, Imran Javed and Waqar Mahmood 1 2: Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction, Pak. J. Engg. & Appl. Sci. Vol. 10, Jan., 2012 (p. 57-67).

[7]   Rohini B. Kharate *1, Dr.S.M.Jagade 2, Sushilkumar N. Holambe, "**A** Brief Review and Survey of Segmentation for Character Recognition", International   Journal of Engineering Sciences, 2(1) January 2013.

[8]   Shilpy Bansal, Mamta Garg, Munish Kumar, "A Technique for Offline Handwritten Character Recognition", IJCAT International Journal of Computing and Technology, Volume 1, Issue 2, March 2014.