International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 03 | Mar-2016 www.irjet.net p-ISSN: 2395-0072

Implementation of Font and Style Text Extraction from Image

[1] Ms. Babita Kaithwas, CSE, DBACER, Maharashtra, India
[2] Ms. Priyanka Dhanvijay, CSE, DBACER, Maharashtra, India
[3] Ms. Shivani Dhone, CSE, DBACER, Maharashtra, India
[4] Assistant Prof. Mr. Ashwin Shinde, CSE, DBACER, Maharashtra, India

---

Abstract - *This paper present the easy and new approach to extract the text from the image. The main task of the project is to convert the printed text to digitized form so as to increase the durability of the text and reduce the maintenance cost. In this we are used the matrix matching concept for extracting the text from image. The segmentation is performed on the preprocessed image and then for each segmented letter a matrix is formed for matching purpose in the recognition phase. The aim of project is to present a way for extraction of text from image which is simple as well as user friendly. A generic character recognition system has different stages like noise removal, segmentation, feature extraction and character recognition.*

Key Words: OCR, Pre-processing, Segmentation, Recognition, etc…

1. INTRODUCTION

A lot of work has been done for detecting text in images and a lot has to be done. Optical Character Recognition is a most important gift given by computer science to the mankind. It has made a lot of tedious work easy and speedy. OCR means a technique of recognition of machine printed or handwritten text by computer and then its conversion to an editable from as per the requirement. This is a technique used to convert any raster image of a document into a computer process able format. . Input is digitized image containing any text, which is preprocessed to segment it into normalized individual words. Further feature extraction is used for extracting the features of character. The feature extraction will be done on the segmented characters, and then extracted features of segmented character will be match with the database already created in training phase.

2. LITERATURE REVIEW:

Text extraction is an image is a challenging task in a computer vision. Text extraction plays an important role in providing useful and valuable information. Text in a documents depend upon various factors such as language, styles, font, sizes, color, background, orientation, fluctuating, text line, crossing and touching text lines. A document image contains various information such as texts, pictures, and graphics.[1]

Optical character Recognition (OCR) system foe camera captured image/graphics embedded textual documents for handheld devices. Characters are passed into the recognition module. Experimenting with a set of hundred business card images, captured by cell phone camera, we have achieved a maximum recognition accuracy of 92.74%.[2]

Offline handwritten character recognition has been one of the most engrossing and challenging research areas in the field of pattern recognition in the recent years. Offline handwritten character recognition is very problematic research area because writing styles may vary from user to another. [3]

Character segmentation has long been a critical area of the OCR process. The higher recognition rates for isolated character vs. those obtained for those obtained for words and connecting character strings well illustrate this fact. A good part of recent progress in reading unconstrained printed and return text may be ascribed to more insightful handling of segmentation.[4]

The techniques like neural networks, structural and statistical pattern are available for recognition of text. But the major drawback of neural network is large training data which takes much more time to build this makes the neural network more complicated for an naïve user to understand leading it to less user friendly.[5]

3. RELATED WORK

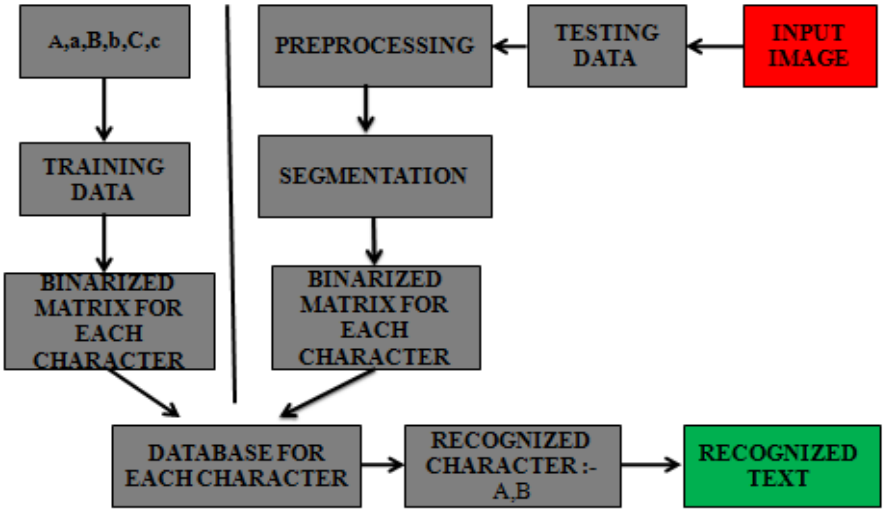

Fig.1: Work flow of text extraction from image.

3.1 Training data:

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

- Image Acquisition

We can acquire a input image by scanned copy of an image or by digital professional camera.

- Image Pre-processing

The input is not suitable for recognition process. Hence, the input image needs to undergo a pre-processing step to convert it into usable for further steps. Binarization technique is used for pre-processing. It is procedure to convert colored or gray scale image into black and white. Pre-processing the input image is very important to make it more accurate for further steps. After the completion of preprocessing a separate database for each character will be created.

3.2 Testing data:

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

- Segmentation

Segmentation step is one of the most important and difficult task in text extraction and recognition. 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 input pre-processed image consisting of sequence of character is there by decomposed into sub-images. 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 data set. In the recognition step the matrix formed for segmented character are matched with stored matrix of character in data set. If the matrix is matched then the character is recognized and printed in word file.

4. IMPLEMENTATION

While executing the system user wants to select the particular image

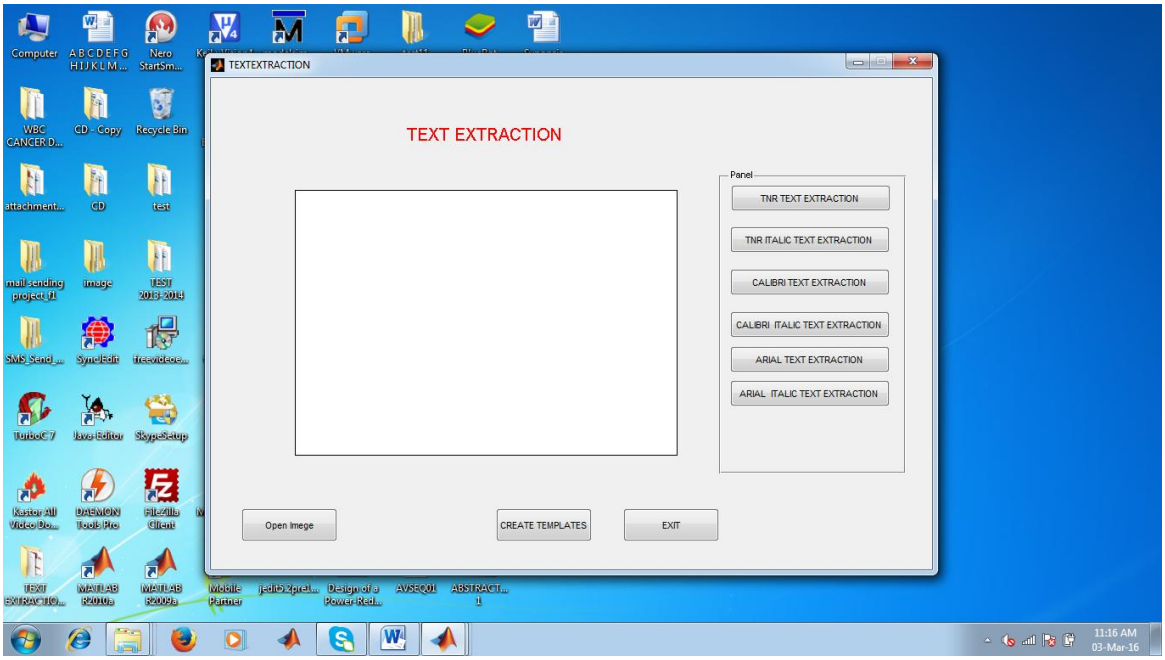

Fig. 2: GUI of the system

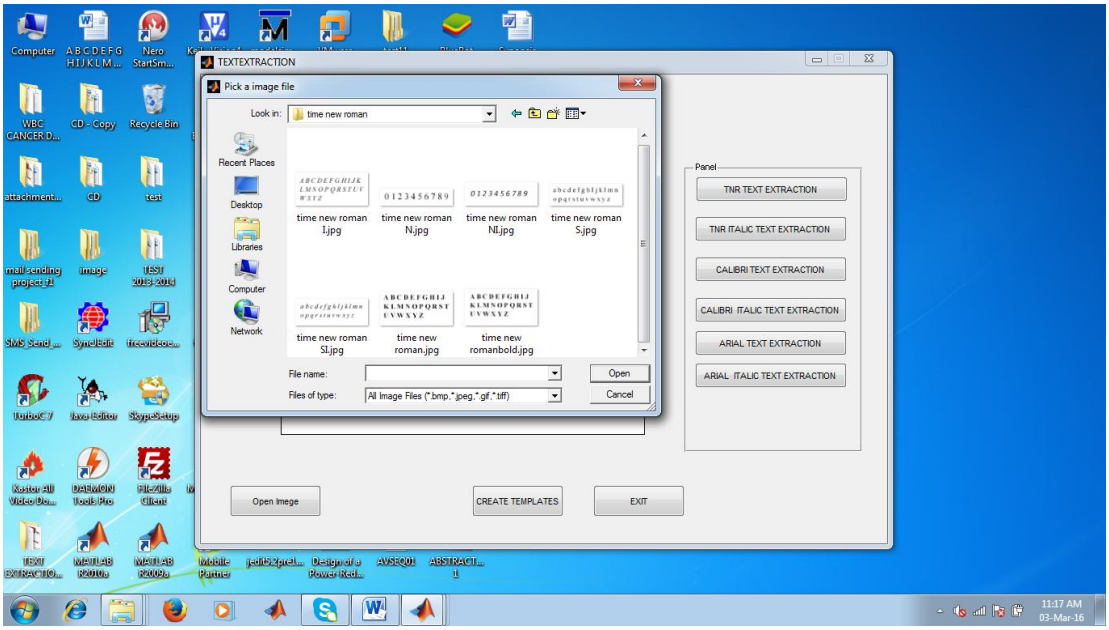

Fig.3: Pop up window after pressing the open image button

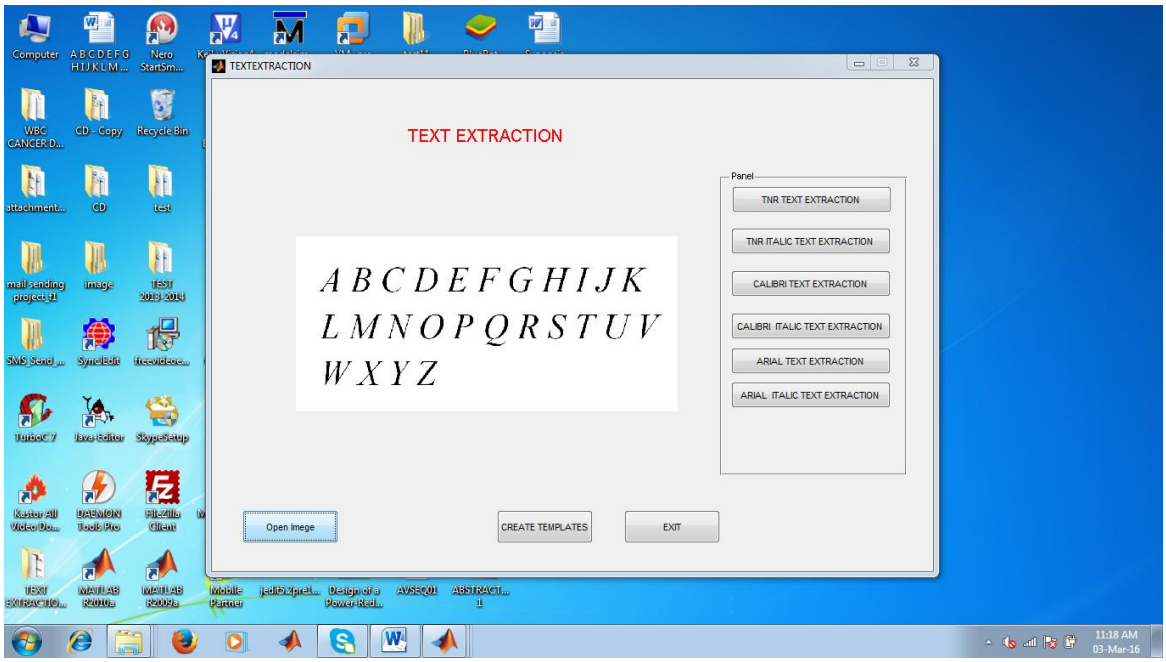

Fig.4: After selecting image

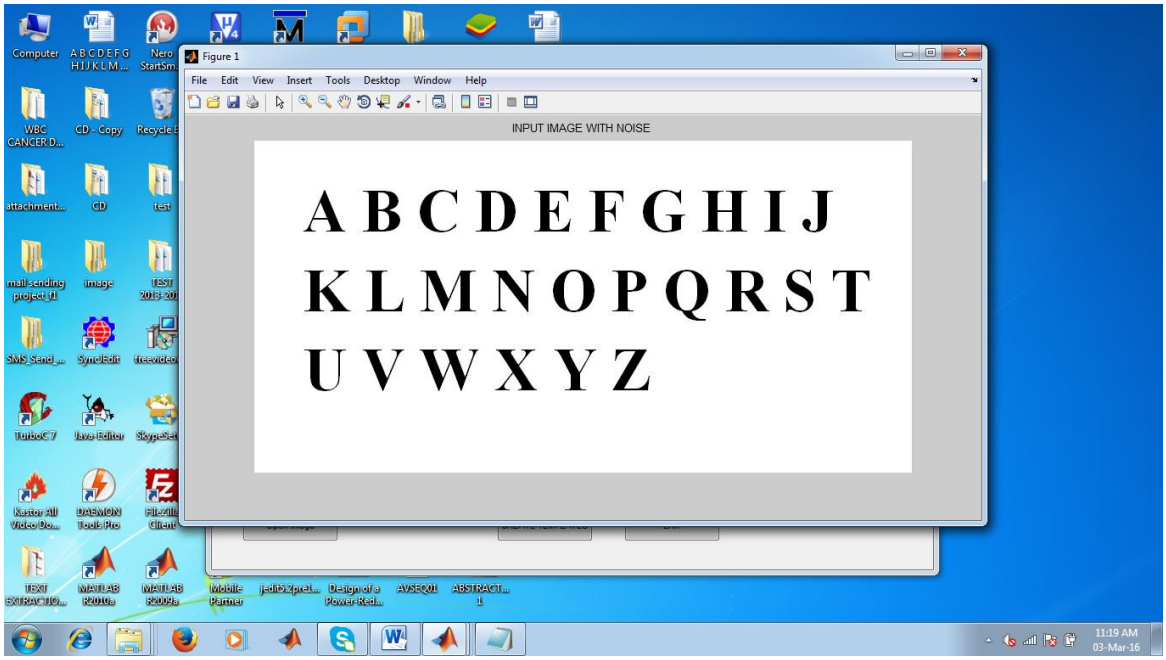

Fig.5: Selected image with noise

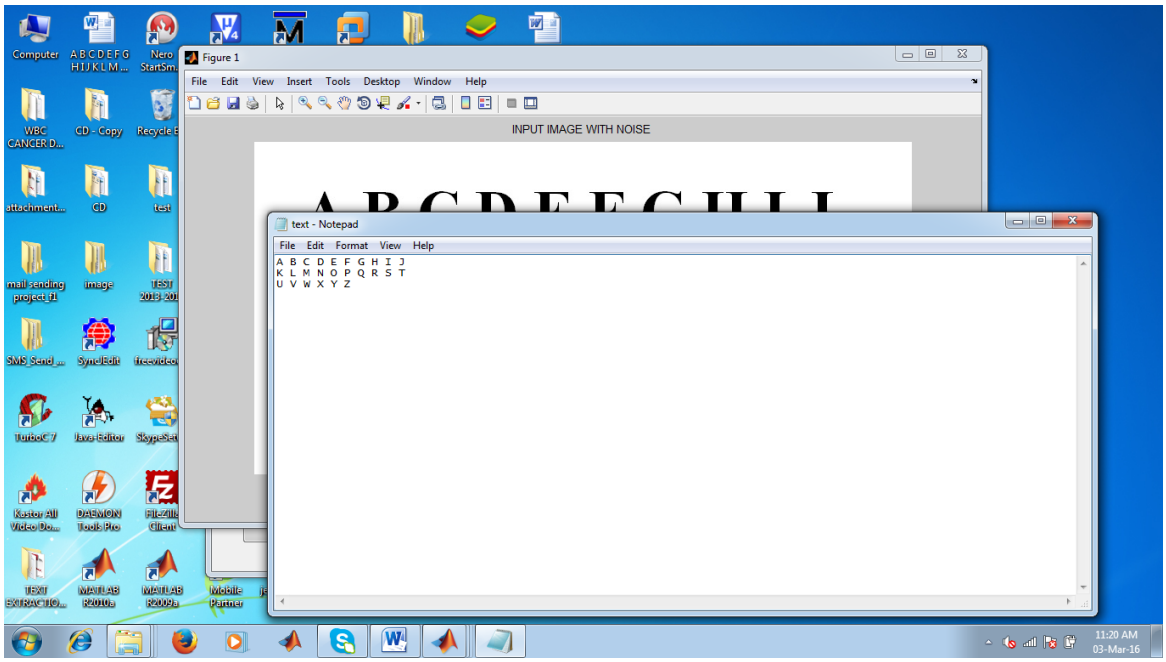

Fig.6: output of the system

5. APPLICATION

There are numerous applications of a text information extraction system, including document analysis, vehicle license plate extraction, technical paper analysis, and object oriented data compression. In the following we have some of these applications.

1. License/container plate recognition
2. Content-based video coding or document coding.
3. Text based image indexing.
4. Industrial automation.
5. Bank application: Extraction image from cheques, reading bank deposit slips.

6. CONCLUSION

In this paper we have successfully able to extract the text from the scanned image. This application is efficiently working on the non-connected letters. A method of text extraction from images is proposed using OCR technique.

ACKNOWLEDMENT

We have great pleasure to express our most sincere regards and deep sense of gratitude to our project guide Mr. A. Shinde, Assistant Professor, Department of Computer Science & Engineering, DBACER, Nagpur for his valuable guidance for completing final year project “Font and Style Text Extraction from Image”.

7. REFERENCES

[1] Deepika Ghai, Neelu Jain “Text Extraction from Document Images”, International Journal of Computer Applications, December 2013.
[2] Raveena Mithe, Supriya Indalkar, Nilam Divekar, “Optical Character Recognition”, International Journal of Recent Technology and Engineering (IJRTE) , March 2013.
[3] ShilpyBansal, MamtaGarg, Munish Kumar, “A Technique for Offline Handwritten Character Recognition”, IJCAT International Journal of Computing and Technology, Volume 1, Issue 2, March 2014.
[4] D.kavitha, P.Shamini, “Handwritten Document into Digitized Text Using Segmentation Algorithm", Special Issue, 4th National Conference on Advanced Computing, Applications & Technologies, May 2014.

[5] Smruti Khati, Rucha Patil, Tulsi Thakur, Harshita Katragadda, Neha Ambulkar, Ketki Bhakare "Text Extraction and Categorization from Image", International Journal of Emerging Technology and Advanced Engineering, Volume 5, Issue 2, February 2015.