MULTIPLE FILE OPTICAL CHARACTER IDENTIFICATION AND RECOGNISATION USING MATLAB

BHARTI SHARMA, ASHUTOSH KUMAR RAO
M.Tech., CSE, Sunderdeep Engineering College, Ghaziabad, Uttar Pradesh
Assistant Professor, CSE, Sunderdeep Engineering College, Ghaziabad, Uttar Pradesh

Abstract **-** Optical character recognition (OCR) is becoming a powerful tool in the field of Character Recognition, now a days. In the existing globalized environment, OCR can play a vital role in different application fields. Basically, OCR technique converts images into editable format. This technique converts images in the form of documents such as we can edit, modify and store data more safely for longtime. This paper presents basic of OCR technique with its components such as pre-processing, Feature Extraction, Classification, post-processing etc. There are various techniques have been implemented for the recognition of character. This Review also discusses different ideas implemented earlier for recognition of a character. This paper may act as a supportive material for those who wish to know about OCR.

Keywords- *OCR, Boundary Detection & noise removal , Feature Extraction*

I. INTRODUCTION

Now a days, globalization is reaching to a great level. In this globalized environment, character recognition techniques also getting a valuable demand in number of application areas. OCR is an effective technique which converts image into suitable format such that data can be edit, modify and stored. This technique performs several operations such as, scans the input image, processes over the scanned image thereby image gets converted into portable formats .For instance, the hard copy of old historical books, novels, etc. .cannot be stored safely for a long time. Rather, its safety has limitations. If we apply OCR technique for such cases, the different historical documents can be stored, modified for a longtime. OCR also having variety of applications in almost all fields, including security. OCR implementation helps us to edit, store and process over the scanned data more effectively. User can handle the stored data whenever he wants with the internet support. So we use Optical character recognition is most successful application used in pattern recognition. purposed OCR system consists of the following basic components:

1. Input Image
2. Pre-processing
3. Feature Extraction
4. segmentation
5. word extraction

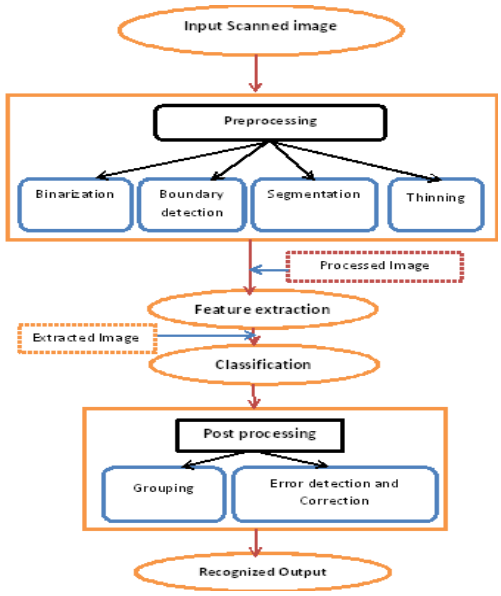

Fig- 1: Processing Stages of OCR Technique

1.1 Input Image

Firstly, image of input data is optically scanned. The scanned image can be any document of different dimensions. This scanned input image is fed to pre-processing section so as to process over that scanned image.

1.2. Pre-Processing

Pre-processing includes several operations over the scanned image, so that input image becomes suitable and comfortable for applying to further sub sections. Basically the objective of pre-processing is to improve the quality of scanned input image. Noise removal, mathematical operations can also be processed in this Pre-processing section. It includes binarization, boundary detection, segmentation, thinning. It performs the several operations over the scanned input data.

1.3 Binarization

Binarization plays an important role in pre-processing. It is necessary to convert a color image into black and white format. So we can process over that black and white image. Basically separation of background and actual image area referred as foreground of a scanned image is called binarization.

1.4 Boundary Detection & noise removal

The binarized image is now applicable for boundary detection noise removal. In this operation the boundaries of scanned image is detected. It detects all the boundaries of image. It is necessary to detect the boundaries so as to select an individual character.

1.5 Segmentation

This is important operation of OCR as rate of recognition is directly proportional to segmentation. In this process, every individual character is separated. This isolates the different sub-parts of an image. It is used to separate pixels of an image as per the contents in data like words, paragraph etc.

1.6 Feature Extraction

For the accuracy of OCR system, the appropriate Feature Extraction method should be selected. While processing over the image some features should be separated. The typical features are Edges, Corners, Ridges, etc. This method of separation is called as Feature Extraction. The accuracy of an OCR technique depends on selection of proper feature extraction method.

1.7 Classification

The feature extracted data must have gone through the process of Classification. This process classifies the extracted individual character in proper way.

1.8 Post-Processing

This is the last and an important phase of OCR technique. It includes different operations like Grouping, Error detection and correction. Whatever the data being operated through different operations such as, binarization, segmentation, Feature
extraction, Classification etc. is fed to post-processing. That means different features of input scanned image are extracted. That feature extracted data is an individual character. It is unable to get detailed information from that individual character. So, it is necessary to collect individual character in appropriate and sequential manner. The process of collecting individual characters of the same contents to form a string is termed as Grouping. By using error detecting and correcting algorithms, errors can also be eliminated.
Finally, we get the recognized output character.

II. LITERATURE REVIEW ON OPTICAL CHARACTER RECOGNITION

Jagruti Chandarana, Mayank Kapadia(2014) , this paper explains comparative analysis between Random Transform and Hough Transform, which are applied for error detection and correction. This paper explains implementation of OCR in Matlab, compared with current working method of OCR. This system achieved recognition rate near about 92%.

Lipi Shah, Ripal Patel. (2014) OCR technique for both handwritten and printed Guajarati script. For this implementation, linear recognition technique has been used. This paper explains how linear recognition technique is efficient in OCR for error detection and correction. This system achieved credit rate near about 89%.

Youssef EsSaadyet. al. 2014, proposed, the method to detect Amazigh handwriting recognition method. The method is based on horizontal and vertical centerline of the character. The characters are segmented into horizontal and vertical lines and position of the character is obtained according to those lines. The features are calculated using sliding window techniques. The characters are classified using MLP. The correction rate obtained was 99.28 % for the 19437 Amazigh printed characters and 96.32% for the 20150 Amazigh handwritten characters.

Dileep Kumar Patel et. al. (2014) proposed the DWT based handwritten character recognition system along with Euclidean distance metrics. The classification of the testing vector is depend upon the Euclidean distance. Minimum disatace decide the class of the test vector. This system gives good accuracy of 90%

Prasad P. Chaudhariet. al. (2014), proposed a grid approach for recognition of an offline handwritten character using grid approach is proposed. Extracted features are train by neural network as classifier of the character in classification stage. The recognition system of experimental results shows that this technique is effective and reliable. The overall procedure results in recognition rate are 96.9%..

ReetikaVermaet. al. (2014), proposed a neural network alog with surf feature approach to solve complex character recognition issue. This approach has been evaluated using noise parameter. The evalution is performed by PSNR and MSE. The classification is done by using back propogation neural network. The success rate of this system is98.77%

Apash Roy et. al.(2014) , proposed a system with feed forward neural network having ability of a machine to interpret handwritten characters from sources like paper document, photograph etc. to digital computerized form is the aim of HCR systems. The neurons of output layer have a feedback connection from their output line. Experimental result shows that an efficient recognition. However, the proposed system is not a complete one. Some other techniques may be combined with this approach to increase the efficiency

of the system. The work can be extended to recognize characters or numerals of some other languages also.

S K R Naganjaneyuluet. al. (2015) proposed a new algorithm to recognize the characters and alphabets from low resolution images. It uses low pass filter in L2 space which slightly improve the performance. The results show that the performance of OCR on low pass filtered images in WSS is far better than the other two cases as images are lower solution. The result shows that OCR giving better performance on the images which are low pass filtered in Weighted Sobolov space.

Kamaljit Kaur et al.(2015) have discussed in their paper that the vehicles are not just identified by the number plate. There are other features that help to identify the vehicle from the remote location as the numberplate can be recognized only when the camera focuses on the number plate. The drawback of this system is that the structure of each and every vehicle is difficult to store in the database.

III. RESULT OUTPUT

Optical character acknowledgment has turned out to be a standout amongst the best uses of innovation in the field of example acknowledgment and manmade brainpower. Numerous business frameworks for performing OCR exist for an assortment of utilizations, in spite of the fact that the machines are as yet not ready to contend with human perusing capacities. Optical Character Recognition manages the issue of perceiving optically prepared characters. Optical acknowledgment is performed disconnected after the composition or printing has been finished, rather than on-line acknowledgment where the PC perceives the characters as they are drawn. Both hand printed and printed characters might be perceived, yet the execution is specifically needy upon the nature of the information records. In the most recent decade the acknowledgment rates of frame perusers close by printed digits and obliged alphanumeric fields have raised altogether (shape perusers for the most part keep running at a high reject/blunder proportion). Numerous scientists now see disconnected and on-line cursive composition as the following test or swing to multi-lingual acknowledgment in an assortment of scripts. Character characterization is likewise a most loved proving ground for new thoughts in example acknowledgment, yet since the greater part of the subsequent analyses are directed on separated characters, the outcomes are not really instantly applicable to OCR.

Finally, we get the recognized output character.

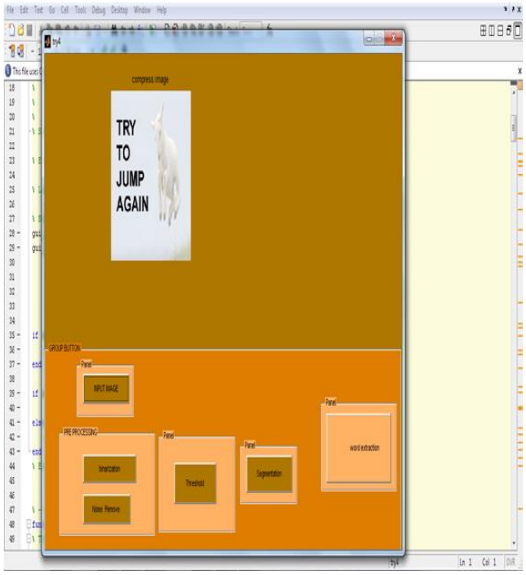

Figure-2: BrowseImage In Gui (Graphical User Interface)

GUI for part of an image processing. To this point I have created a push button which allows me to browse through my working directory and select either a 'jpg' or 'bmp' image.

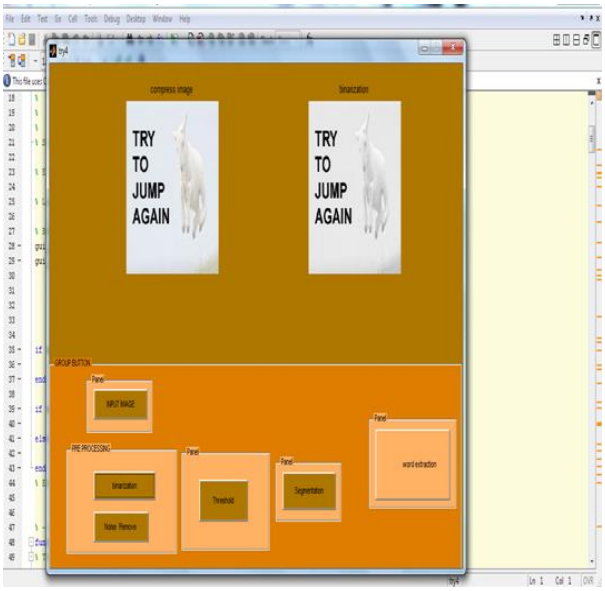

Figure-3: Binarization & noise Removal images

Method was proposed to remove noise & Binarization assumes an essential part in pre-handling. It is important to change over a shading picture into high contrast arrange. So

we can handle over that highly contrasting picture. Fundamentally partition of foundation and real picture territory alluded as frontal area of an examined picture is called binarization.

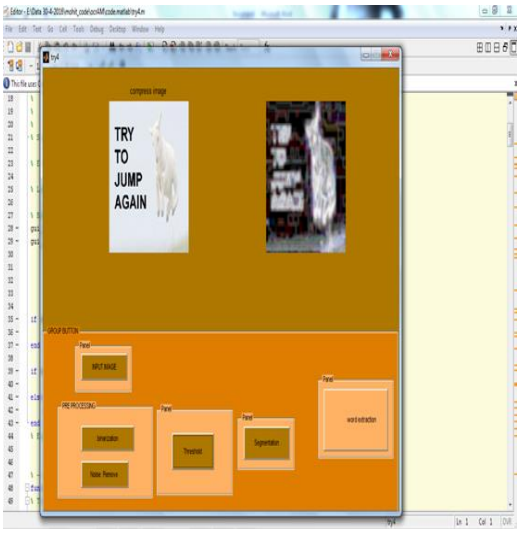

Figure-4: Images Thresholding

The main objective of the Dynamic Threshold Algorithm is to set a threshold for the binary (1 for black, 0 for white) decision about a given pixel. must adapt quickly after leaving a very dark character so that a following lighter character will not be eliminated.

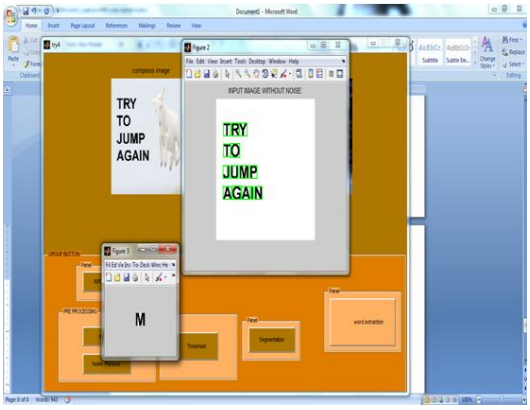

Figure-5: Image Segmentation

The purpose of our character segmentation utility is to produce rotation and scale invariant images of the characters in a specific word from the Rochester flag. The images are binary, and 20x20 pixels in size. The character segmentation techniques used to read "T R Y T O J UMOP AGAIN " from the U. R. flag rely on a number of structural invariants. The indoor environment will allow us to prevent motion of the flag and ensure that it remains flat. A-priori knowledge of the flag's two-dimensional structure allows us to design simple reactive behaviors that can segment the desired characters. We have chosen to read the large collinear letters near the bottom. These characters are written in solid yellow material against the dark blue background, and each occupies nearly 9 square inches. Our text segmentation utility relies on a signal from the camera that can image these characters to an area of at least 20 x 20 pixels. This resolution constraint was required because the character classifier utility expects images of size 20 x 20.

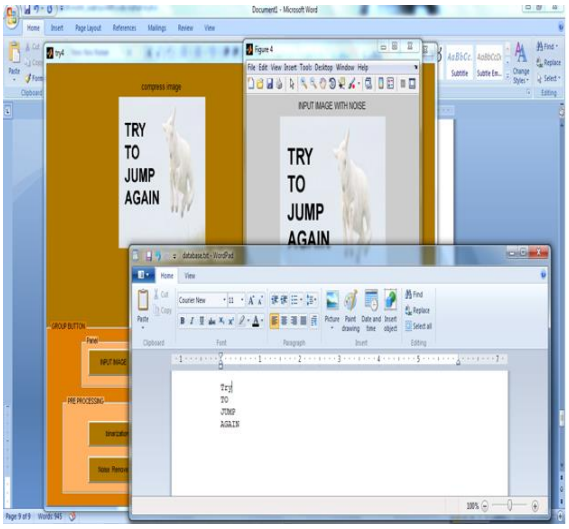

Figure-6: Final Output

This step is not compulsory; it helps to improve the accuracy of recognition. Syntax analysis, semantic analysis kind of higher level concepts might be applied to check the context of recognized character. When image is provided as input to OCR system, its features are extracted and given as an input to the trained classifier. Classifiers compare the input feature with stored pattern and find out the best matching Output.

IV. CONCLUSION

A survey of feature extraction and classification techniques for optical character recognition is studied. A lot of research has been done in this field. Still the work is going on to improve the accuracy of feature extraction and classification techniques. Due to algorithmic simplicity and higher degree of flexibility, template matching and Correlation method is easy to implement with the change of recognition target classes. Its recognition is strongest on monotype and different types of fonts considering the sample input images for example handwritten image and it takes shorter time and does not require sample training but one template is only capable of recognizing characters of the same size. The OCR algorithm which is implemented in MATLAB (R2009.a/64-bit) gives

optimal accuracy on an average as 91.16% and also the Radon transform applied for skew detection and correction gives better results as compared with Hough transform.

V. FUTURE SCOPE

In this research, it is proved that Rough sets theory can be effectively used in reducing feature dimensionality in Telugu OCR. The scope, however, is not only limited to English. The developed methodology can be made as a unified approach for OCR of any language, provided a training set is available. Thus, it can also be used in other language scripts, as well as in handwritten character recognition. Developing commercial OCR systems which can maintain high recognition rates regardless of the irregularities such as the quality of the input documents, and varying font styles is a challenging task for English and other Indian scripts. Compared to European languages, Indian languages have many additional challenges like larger character set due to modifiers, lack of standard test databases, and lack of support from browsers, operating system, and keyboard, etc such as identification of glyph position information, recognizing punctuation marks from the width and height information, handling of confusion pairs of glyphs and touching characters would help improving the performance of Telugu OCRs further. Also, hybridizing the proposed methodology with some of the popular methods such as N-grams may be thought over to enhance the performance to suit commercial OCRs.

VI. REFERENCES

[1] **J R Prasad, U V Kulkarni, R S Prasad,**" Offline handwritten character recognition of Gujarati script using pattern matching", 3Rd InternationalConference on Ant i-counterfeit ing Security and Identificat ion incommunicat ion, **(2009).**

[2] **Y. Sobu , H. Goto , H. Aso,**"Binary tree-based precision-keeping clustering for very fast Japanese character recognition", 25ThInternational Conference on Image and Vision computing, New Zealand,**(2010).**

[3] **M. Kumar, M K Jindal, R K Sharma,**" k-nearest neighbor based offline handwritten Gurumukhi character recognition",InternationalConference on Image Informant ion processing **(2011).**

[4] **Dan ClaudiuCires¸an and Ueli Meier and Luca Maria Gambardella and JurgenSchmidhuber,** "Convolutional Neural Network Committees for Handwritten Character Classification", 2011International Conference on Document Analysisand Recognition, IEEE, **(2011).**

[5] **GeorgiosVamvakas, Basilis Gatos, Stavros J. Perantonis,** "Handwritten character recognition through two-stage foreground sub-sampling" ,*Pattern Recognition*, Volume 43, Issue 8, August **(2010).**

[6] **ShreyDutta, Naveen Sankaran, PramodSankar K., C.V. Jawahar,** "Robust Recognition of Degraded Documents Using Character N-Grams", IEEE,**(2012).**

[7] **Naveen Sankaran and C.V Jawahar,** "Recognition of Printed Devanagari Text Using BLSTM Neural Network", IEEE, **(2012).**

[8] **Yong-Qin Zhang, Yu Ding, Jin-Sheng Xiao, Jiaying Liu and Zongming Guo1,** "Visibility enhancement using an image filtering approach", Zhang et al. EURASIP Journal on Advances inSignal Processing **(2012).**

[9] **W.Badawy,** "Automatic License Plate Recognition (ALPR): A State of the Art Review." **(2012).**

[10] **Ntirogiannis, Konstantinos, Basilis Gatos, and IoannisPratikakis.**"A Performance Evaluation Methodology for Historical Document Image Binarization."(2013).

[11]**SandeepTiwari, Shivangi Mishra, Priyank Bhatia, Praveen Km. Yadav. "**Optical Character Recognition using MATLAB International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)Volume 2, Issue 5, May 2013

[12]. **Majida Ali Abed Hamid Ali** ." Simplifying Handwritten Characters Recognition Using a Particle Swarm Optimization Approach" European Academic Research, Volume I, Issue 5/ August 2013