



Offline OCR for Handwritten English Characters and Numerals using Shape numbers and Chain codes

N. Shobha Rani, Lakshmi Prasanna Kumari.R, Nikhil Bopaya .M.A

Department of Computer Science

Amrita Vishwa Vidyapeetham, Mysore campus, Mysore, Karnataka, INDIA

Abstract: Automation of text document images into editable form is the primary concern for the document image understanding systems. This paper proposes an approach of converting the handwritten English documents into editable form. The documents considered include both English alphabets and numerals. The algorithm is mainly focused on segmentation touching characters in handwritten script using chain code features in the regions of minimum valleys obtained using vertical histograms of the word. The principal component analysis features are used for classification of segmented characters and numerals. The algorithm has given encouraging results with an overall accuracy of 89%.

Keywords: Scanned images, OCR, machine –encoded text, English characters, shape numbers, chain codes.

I. Introduction

Image processing is a method that find its application in cases where in an image is to be converted into its corresponding digital form and there a few operations which are to be performed on the same. Basically this a type of dispensation of signals in which the image in question is taken as an input and the output being the same image or the characteristics that is associated with that particular image. Most frequently than not the Image Processing System includes certain predetermined signal processing methods which are applied upon the input image. Image Processing is a domain that is ever so steadily forming the integral core research area within the engineering and computer science disciplines. There are three different steps which are followed in a general Image Processing scenario. i.e., acquiring the image using an optical scanner or by using a camera, the acquired image is now analyzed and manipulated by using data compression, image enhancement techniques and by spotting some patterns that are not visible to naked eyes and the last stage being the Output will have an enhanced image in terms of quality or a corresponding report based on analysis of that image. The main purpose of Image Processing is to get a clear observation of the objects that is not visible, to create a better image in terms of the sharpness and restoration, to get an image of interest, to measure various objects in an image and to distinguish the objects in an image.

A. Types of Image processing

There are basically two methods that are used for Image Processing, Analog and Digital Image Processing. When the image of interest are hard copies like printouts we use Analog image processing. Analog image processing or visual image processing in combination with different fundamentals of interpretation are used by Image analysts. Another tool that is used in Image Processing through visual techniques is Association, because of which the analysts use a well grafted combination of his/her personal knowledge and the collateral data to image processing. When the interested image is in digital format it is when the Digital Image Processing comes into picture, the manipulation to such images are done by using computer and thus the name Digital Image Processing. In the rapidly growing field of Image Processing, Handwriting recognition is one of the most talked about area. There are a whole lot of discrepancies which are to be met with their corresponding solutions when the characters from a scanned document has to be recognized, this is the reason behind we using the standard hierarchy of Image Processing procedures to find an optimal and efficient way with which we can meet the objective of our project, an OCR than can recognize handwritten characters and numerals with a good speed and accuracy. There are five major image processing techniques:

A.1 Preprocessing- When the input image is at its lowest level of abstraction there will be a lot of features of that image that will require enhancement to be able to proceed with the further steps. The main objective behind this step is to improve the image by avoiding the unwanted discrepancies which might well lead to less accurate results at the later part.

Four categories of image processing methods are

- Pixel brightness transformations.
- Geometric transformations.
- The use of a local neighborhood of the processed pixel by certain Preprocessing techniques.
- Image restoration that requires knowledge about the entire image.

A.2. Segmentation- It is the process in which a digitized image is divided into pixels, An image is a collection of n number of pixels, and the process of assigning a label to each and every pixel in that image is called segmentation. The idea behind labeling all the pixels is that the pixels with the same level share certain characteristics. The goal of image segmentation being bringing the image of interest into a meaningful representation which will make it easier to be analyzed and will aid in the future steps.

A.3 Character segmentation

When we talk about digitizing a handwritten document, there is countless number of advantages that come into our mind. For an instance, considering a sheet of paper where there is very important information but the characters are not clear. The essence of our project being to just get this sheet digitized with a scanner or so and then with a few image processing techniques we process this image and reveal the information the paper had to. OCR (Optical Character Recognition) is a branch of computer science which reads texts from the paper and images into a digitized form (which can be manipulated by the computer). An OCR helps us to load the computer with a file (a written document for example) and then edit the file in any of the word processors the computer offers. Usually the images are read into the computer with the help of a scanner and with the help of some comprehensive software's to analyze the documents taken from the scanner. The amount of advantages we can actually have when we have our data in an electronic or digital format is enormous, with the help of a computer we can make our data reach any part of the world within no time (E-mail). OCR helps in getting these handwritten documents to be in digital format without actually having to type all the data.

If there is something important that will gauge the performance of an OCR then it has to be the way in which that particular OCR segments the individual characters, the characteristics of the individual characters may be very different and the quality of the document are a few major challenges that might hamper the performance of the OCR. The spaces two individual persons will leave between two characters they write will be unique, this is another thing that will add to make segmentation of individual characters in handwritten documents very complicated. There are a few other concerns that will make segmentation very complex, touching characters and overlapping of characters being the others.

In a handwritten document there is a high possibility that two adjacent characters appear to be very closely written, called "touching characters" in terms of Image Processing. To be able to identify these characters, these characters should be separated first. So we have to come up with an efficient way with which we can extract these characters individually.

A.4. Feature Extraction- When there is an input which has very large set of data and expected to be redundant as well, with feature extraction we can transform the data into a set of features which will represent the same data. The transformation from the input data to a set of features is called feature extraction.

A.5. Classification- The classification includes a broad range of decision-theoretic approaches to the identification of images. Image classification analyzes the numerical properties of various image features and organizes data into categories. The intent of the classification process is to categorize all pixels in a digital image into one of several land cover classes or themes. This categorized data are used to produce thematic maps of the land cover present in an image.

For classification, we have used three different classifiers SVM, MLP and Naïve bayes, with 5 fold and 10 fold cross validations are used to recognize various characters and to achieve a better accuracy. SVM is also divided into 4 Kernels namely, Linear, Polynomial, RBF and Sigmoid Kernels.

A.6. Recognition The process of representing a language in a symbolic representation from its spatial form of graphical marks is called recognition. When any digital image is taken into consideration, there a few objects and features in it, identification and detection of such objects and features is done by recognition.

A.7. Post preprocessing- after the classification is done the accuracy has to be checked for the validation and verification of the results, change the detection and monitoring based on two or more dates of classified images; and the classified images are integrated with the other data layers and other conventional survey maps.

II. Literature Survey

There are numerous and very intensive experimentations carried out in the area of printed and handwritten English OCR scripts since 1970's. Some of the details related to the area of English character recognition are presented as follows. The OCR for English which used an algorithm based on the principle of Artificial Neural Network (ANN) showed results that only a few character patterns were recognized with accuracy more than 50%. [1] A new algorithm which used median filters was developed to avoid the discrepancies in the OCR while recognizing the character patterns. The samples used were of high quality (very little noise), the results showed great improvement in speed as well accuracy with which the OCR recognized the character patterns [2]. By using the 40-point feature extraction technique the different attributes and character morphology of all the individual characters were highlighted and stored in a Neural Network created in Matlab, the features extracted from the tested characters are matched with those previously stored in the Neural Network, the maximum percentage to which the match would occur between the ones in the Neural Networks and the tested characters were represented with the aid of a graph [3]. The usage of three different classifiers namely SVM, MLP, Naïve bayes during classification was found to be useful, using the neighborhood foreground pixel density with PCA

the accuracy level was found to be an encouraging 91.95% [4]. A new system that can read a predefined set of document formats was developed [5]. A system whose performance was validated on sample datasets of known and unknown users had its own shortcomings when it could not avoid segmentation errors in some digit samples, but it showed enough promise to be used as an open source for OCR engine for recognition of handwritten numerals of Roman script with a fairly acceptable accuracy [6]. There was an accuracy of an encouraging 94.1% when an hybrid approach was used with neural network as classifier to recognize handwritten numeral characters [7]. The result of structure analysis shows that if the number of hidden nodes increases the number of epochs taken to recognize the handwritten character is also increases [8]. Neural Networks are mostly used for Pattern Recognition task, by using parameters like number of Hidden Layer, size of Hidden Layer and epochs and Multilayer Feed Forward network with Back propagation, different models of Neural Network and applied to the test set on each to find the accuracy of the respective Neural Network [9]. Feature extraction is an integral part of any recognition system. The aim of feature extraction is to describe the pattern by means of minimum number of features that are effective in discriminating pattern classes. The gradient measures the magnitude and direction of the greatest change in intensity in a small neighborhood of each pixel. (Gradient refers to both the gradient magnitude and direction). Gradients are computed by means of the Sobel operator. In this paper an effort is made towards recognition of English Characters and obtained recognition accuracy of 94%. Due to its logical simplicity, ease of use and high recognition rate, Gradient Features should be used for recognition purposes [10]. By projecting the handwritten characters on different sized grids there was experiment to recognize English handwritten characters using Matlab Neural Network toolbox. There was an image which was acquired and was made to undergo noise filtering, smoothing and normalization of scanned image, rendering image suitable for segmentation where image is decomposed into sub images, Feature Extraction improves recognition rate and misclassification. Character extraction and edge detection algorithm are used for training the neural network to classify and recognize the handwritten characters [11]. There was another study that explores the existing ring based method, the new sector based method and the combination of these, termed the Fusion method for the recognition of handwritten English capital letters. The variability associated with the characters is accounted for by way of considering a fixed number of concentric rings in the case of the ring based approach and a fixed number of sectors in the case of the sector approach. Structural features such as end points, junction points and the number of branches are used for the pre-classification of characters, the local features such as normalized vector lengths and angles derived from either ring or sector approaches are used in the training using the reference characters and subsequent recognition of the test characters. The recognition rates obtained were encouraging [12]. The proposed system extracts the geometric features of the character contour. These features are based on the basic line types that form the character skeleton. The system gives a feature vectors as its output. The feature vectors so generated from a training set were then used to train a pattern recognition engine based on Neural Networks so that the system can be benchmarked [13]. Object detection based on image processing is the task of identifying a designated object on a static image or a sequence of video frames, out of all the methods used as of now they were categorized into gradient-based and edge based feature extraction methods, depending on the low level features they use. An image can also be considered as a grid of image patches, it is therefore reasonable to incorporate the concept of granules to gradient for a review [14]. The researches on the English language script are very successful and many approaches has achieved an accuracy of almost 100% in case of printed characters and even isolated and even touching character segmentation of the handwritten scripts in English is currently under progress. The proposed experimentation mainly focuses on recognition of printed and handwritten English scripts together. The paper is organized as follows, section 3 discusses the proposed methodology, section 4 addresses the results of experimentation and section 5 concludes the work.

III. Proposed Methodology

Algorithm:

Step 1: Read an input image

Step 2: Pre-process the input image

- a. *Convert to gray scale image*
- b. *Apply Gaussian filtering to remove the noise*
- c. *Threshold the image to obtain binary image*
- d. *Resize the image*

Step 3: Perform the line segmentation of the document

Segment the document into lines using horizontal projection histograms

1. *Mark all the rows corresponding to a zero valley before a peak is considered as starting of line.*
2. *Mark all the rows corresponding to a zero valley after the peak is considered as ending of line.*
3. *Sort all the marked rows*

4. Extract the portions of image for each two adjacent marked row entries.
5. Save all the extracted portions or lines as separate image variables

Step 4: Perform the word segmentation of the lines

- a. For each line compute the vertical projection profiles.
- b. Mark all the columns corresponding to a zero valley that comes before every peak.
- c. Mark all the columns corresponding to a zero valley that comes after every peak.
- d. Sort all the marked columns
- e. Extract the portions of line that comes between each pair of marked columns.
- f. Save all extracted portions or words as separate image variables

Step 5: Perform the character segmentation of each word.

(As the words obtained in this case are composed of touching characters)

- a. Compute shape numbers of various touching portions between the characters.
- b. Train the shape numbers using if then rules to the database
- c. For each word, starting with medial axis the 8-directional chain codes are computed.
- d. Obtain the first difference or shape number of the chain code.
- e. Compare with trained shape numbers to identify the column where segmentation has to be performed to extract a character.
- f. Mark all the columns in each word where the shape number is matched.
- g. Extract the portions of word that comes between each pair of marked columns.
- h. Save all the extracted portions or characters.

Step 6: Feature Extraction, classification, recognition and post-processing

- a. For each character extracted compare the trained character templates with extracted characters from word on pixel by pixel basis
- b. Template convolution is performed between the test character extracted and trained characters in the database.
- c. The template with max pixel count match is considered as the recognized one.
- d. The recognized template in the database is further sent for post-processing.
- e. The Unicode of corresponding matched template is written to the editable file.

Step 7: Finish.

A. Sample Input images:

Some of the samples considered as input to the proposed system is as presented in figure 1.

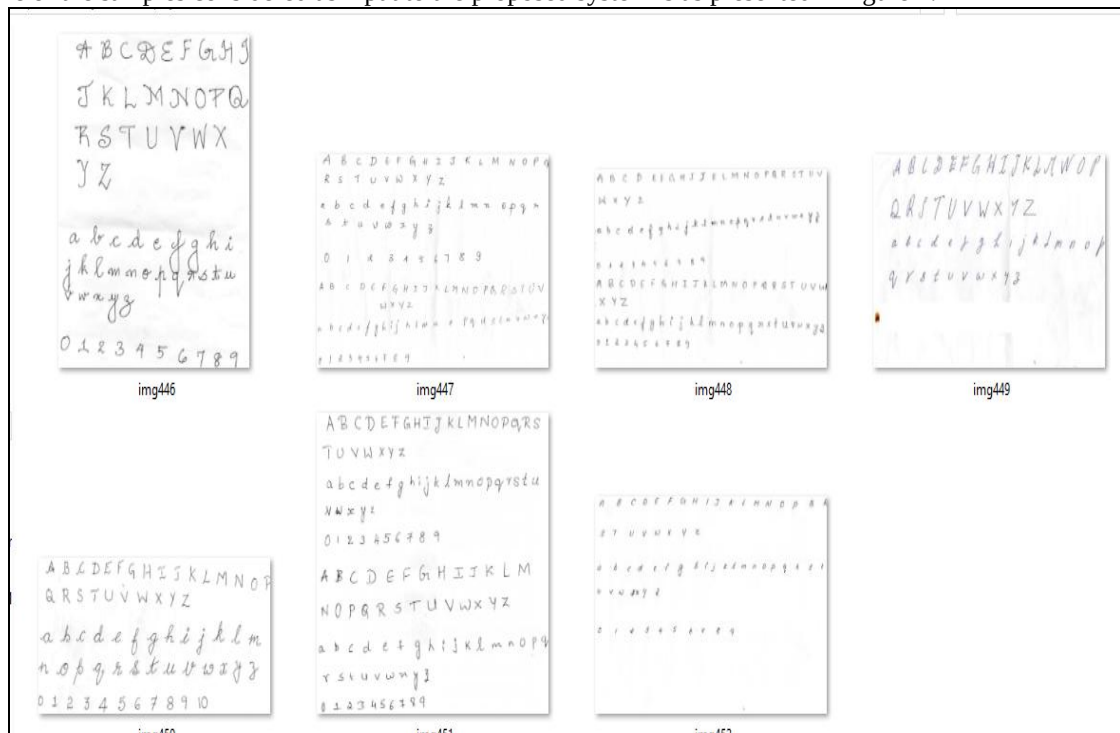


Figure 1: Samples considered as input to the proposed system.

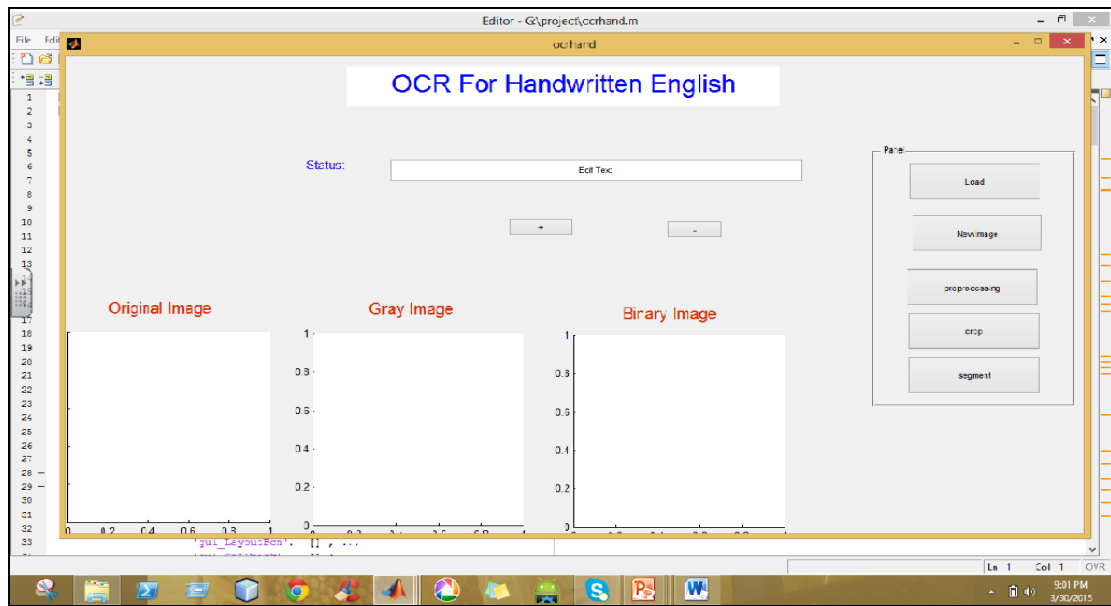


Figure 2: User Interface of Skew detection and correction.

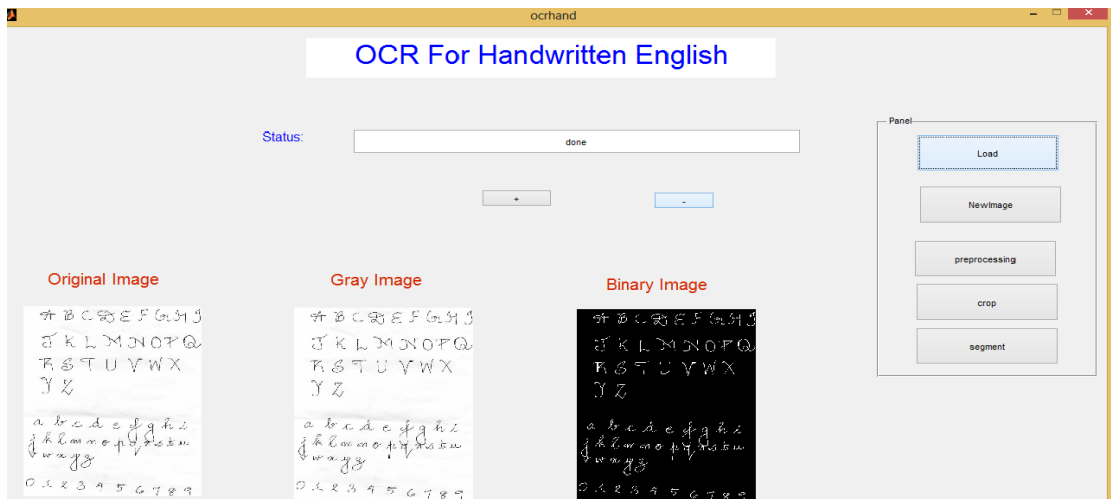


Figure 3: The input image with the gray and binary formats.

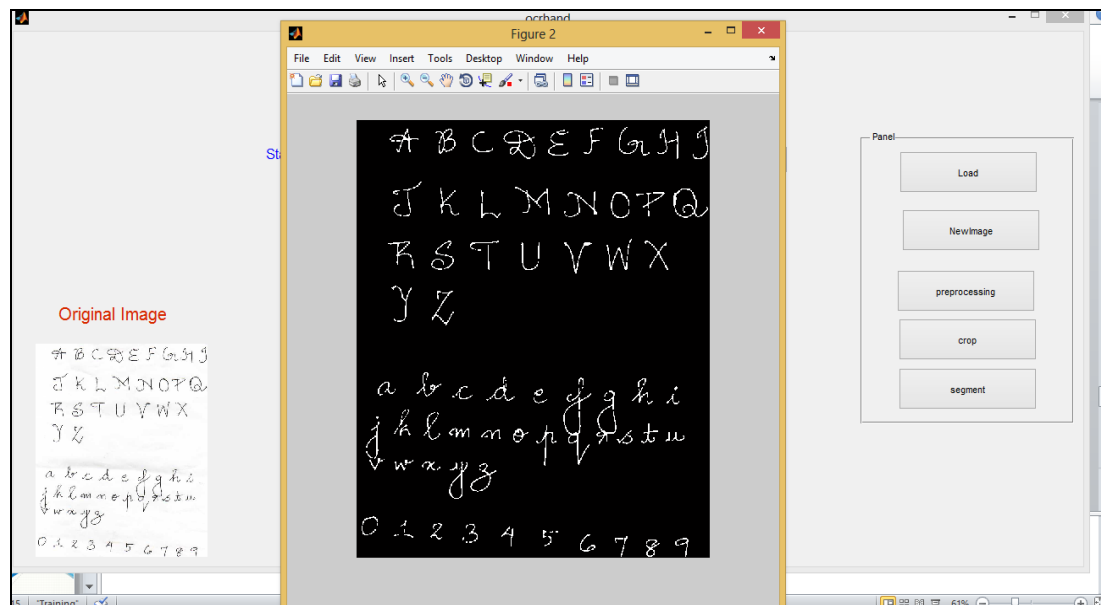


Figure 4 :The preprocessed image



Figure 5: Shows the segmented images

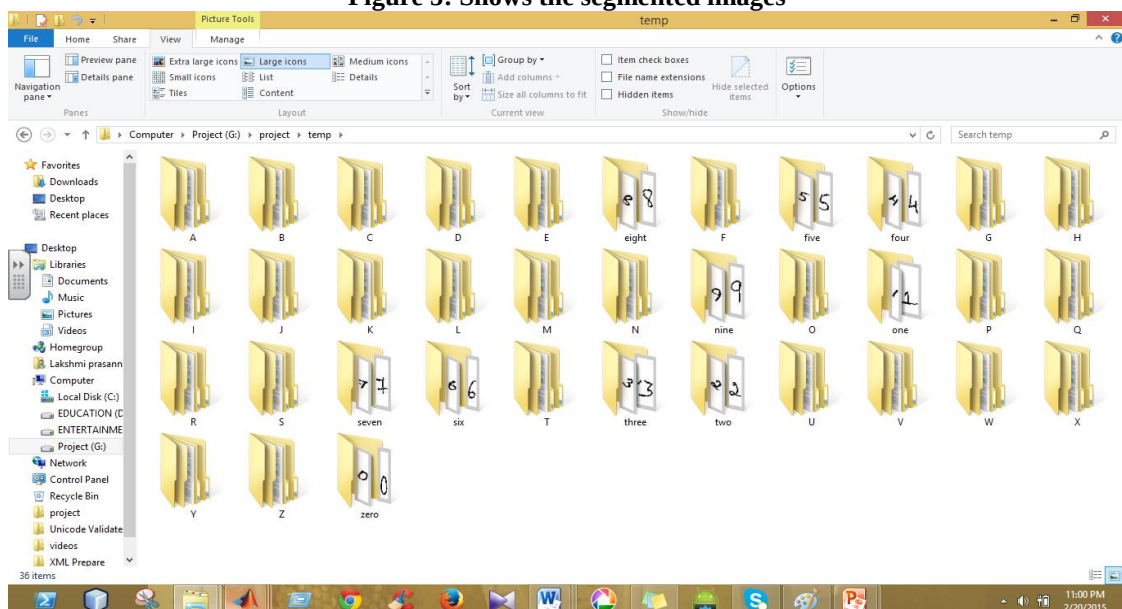


Figure 6: The storage of segmented images

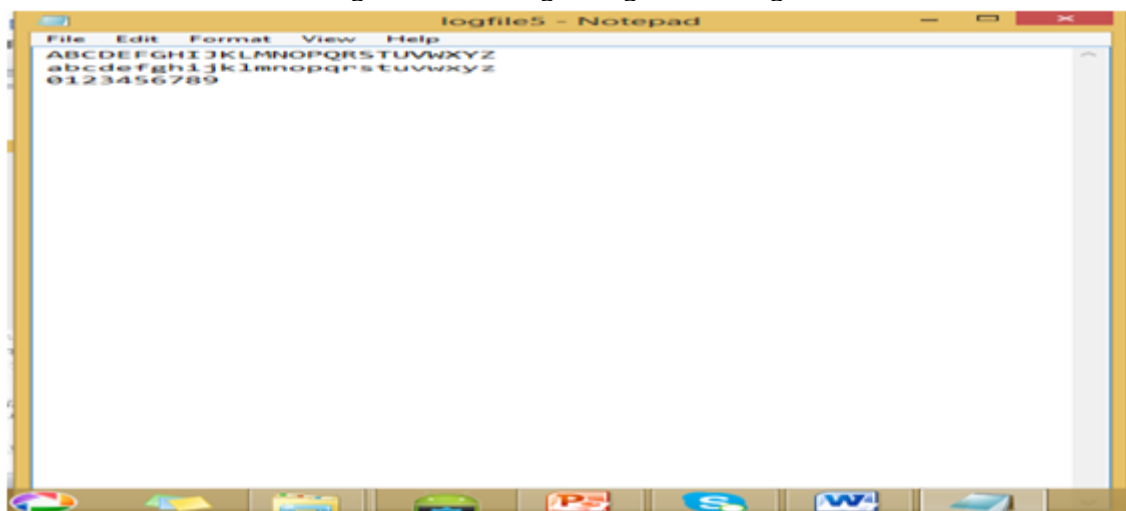


Figure 7: The output results in text format.

IV. Experimental Analysis

The proposed system is experimented with more than 50 images for testing. The algorithm has provided satisfactory results. The use of projection profile for determination of line and character segmentation. To recognise template convolution has been used which is simple and Unicode for each character is written to get the accuracy. The accuracy of algorithm is around 89%.

V. Conclusion

The proposed methodology has achieved very good recognition rates of around 89% in case of both printed as well as handwritten documents. This is a generic method which can be applied on either handwritten and printed documents and can perform automatic segmentation and recognition of the English documents. The projection profiles are suitable especially for the documents of Roman languages like English since documents of this type usually possess clear segmentation boundaries between the lines and words.

References

- [1] Tirtharaj Dash and Tanistha Nayak "Non-Correlated Character Recognition using Artificial Neural Network"
- [2] Aparna.A and Prof.I.Muthumani "Optical Character Recognition for Handwritten Cursive English characters" ,(IJCISIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, 847-848.
- [3] Sandeep Saha ,Sayam Kumar Das, Nabarag Paul ,Sandip Kundu "Optical Character Recognition using 40-point Feature Extraction and Artificial Neural Network" ,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013.
- [4] Anita Pal & Dayashankar Singh "Handwritten English Character Recognition Using Neural Network " ,International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 141-144.
- [5] Shalin A. Chopra[1], Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi, Prof. Gandhali S. Gurjar "Optical Character Recognition " , International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014.
- [6] SandipRakshit, AmitavaKundu, MrinmoyMaity, SubhjitMandal, SatwikaSarkar, SubhadipBasu "Recognition of handwritten Roman Numerals using Tesseract open source OCR engine". Proc. National Conference on NAQC (2008) 141-145.
- [7] P. P. Chaudhari, K.R.Sarode "Handwritten Digits Recognition Special point "International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 3,March 2014 .
- [8] Nisha Vasudeva, Hem Jyotsna Parashar and Singh Vijendra "Offline Character Recognition System Using Artificial Neural Network "International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012.
- [9] Chirag I Patel,Ripal Patel, Palak Patel "Handwritten Character Recognition using Neural Network", International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011.
- [10] AshutoshAggarwal, Rajneesh Rani, RenuDhir "Handwritten Character Recognition Using Gradient Features "
- [11] Kauleshwar Prasad, Devvrat C. Nigam, AshmikaLakhotiya andDheerenUmre "Character Recognition Using Matlab's Neural Network Toolbox "
- [12] Hanmandlu M, Murali Mohan K.R,Kumar H "Neural based handwritten character recognition" Document Analysis and Recognition, 1999. ICDAR '99. Proceedings of the Fifth International Conference on 20-22 Sep 1999.
- [13] Dinesh Dileep"A feature extraction technique based on character geometry for character recognition"
- [14] Sheng Wang "A Review of Gradient-Based and Edge-Based Feature Extraction Methods for Object Detection"