International Journal of Management, IT & Engineering
Vol. 11 Issue 10, October 2021
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

A Narrative Content Detection Methodology for Handwritten English Alphabets

Prof. Rupa Rajakumari R. Peter
M.Sc., M.Phil., Asst. Professor
Dept. of Computer Science Hislop College, Nagpur.

Abstract

An ascend of Artificial Intelligence technology along with machine and deep learning are opening up almost limitless possibilities. There is also an element of fear towards this exponential growth. But it's the humanization of our machines and devices that is seldom discussed or reported. But we deem it acceptable to nominate an unfortunate volunteer to retype notes scribbled on tiny pieces of paper so they can send to other attendees. Repeatedly, again we need to ask ourselves if this is the most efficient method of managing our workload in this digital age. In order to solve this problem, we are proposing a solution to solve this. Instead of predicting by word, here we will be segregating the cursive English letter to individual characters and predicting it via trained Convolution Neural Network (CNN) model. By using this methodology, increase in rate of prediction of the individual characters is been increased and it could be implemented to digitalize the forms in companies which will be a greater level of automation.

Keywords:*Cursive handwritten letters, prediction, CNN model, automation*

INTRODUCTION

Digitization of the handwritten documents, forms are been done by manually and some proprietary software in offices. All the details are been entered manually into the system. Those records are been stored in their servers using some web application. But when migrated to a digital world and to transform all the previous data to digital, it will be a tedious process. When we have thousands of the data papers, we can't rely on the accuracy of humans. Also they have a heavy workload which may lead to decrease in accuracy levels of digital transformation. Once again we need to ask ourselves if this is the most efficient method of managing our workload in this digital age. In our proposed system, documents are been scanned to a portable document format. From the PDF we will be generating the images to perform template matching. Through the template matching, individual lines are been identified and been marked. These identified marked lines are been cropped out and been saved. These are been done (process of identify the columns filled) are been perform the contour detection to separate the lines. From the line images character segmentation algorithm is been applied to separate them into separate characters. This algorithm has some approximation methods to find out the cutting lines. Cutting lines are those points where need to split between two separate characters which are been joined in cursive letters. Then the trained model with 2000 of images will be able to predict the character. The model is been trained with nearly 2000 images of individual handwriting of characters from 10 different people's. In English alphabets we have two kinds of characters (i.e) open characters and the closed characters. In our character segmentation algorithm we will be handling the open characters and able to differentiate between u and two i's and two l's. With the separated character, we will be fed into input to the predictor model.

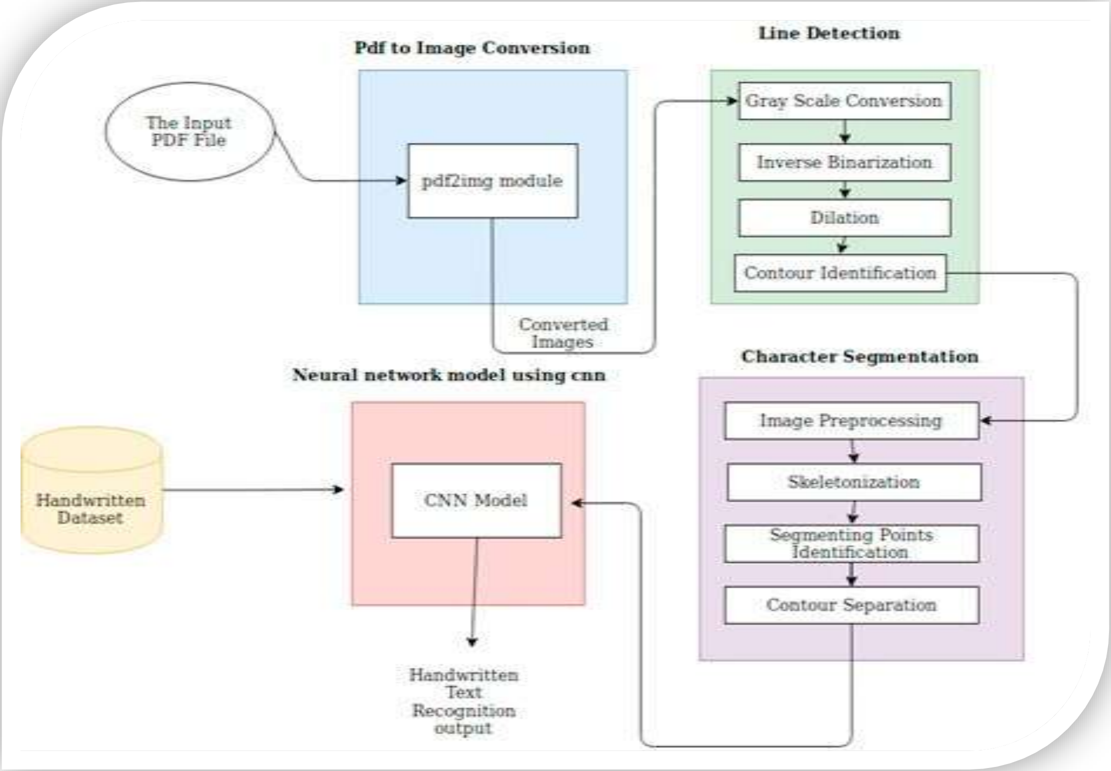

Fig.1. Digitization Architecture for Handwritten English Alphabets

Proposed system shown in Fig.1 is likely to have the following advantages:

- The proposed system will be efficient in character segmentation.
- It also helps in establishing identification of individual characters.
- Automatically will update the data to the database using web application.

Techniques and Methodologies

For conducting the segmentation experiment by the proposed segmentation technique, handwriting samples from 100 different people has been gathered. These samples have both alphabets and cursive sentence separately as well as together. Some of these samples were written on white paper and others on a colored or a noisy background.The image dataset are segregated and collected for training which contains all shapes of English characters. Some of the word image samples from the collected database. Handwritten characters recognition is been carried out by following steps:

1. PDF to Image Conversion
2. Line Detection (contour detection)
3. Image Preprocessing
4. Character Image Segmentation
5. Character Prediction Model

Digitalization of important paper documents creates a layer of Security, and often convenience. Scanned copy of the original document provides a fallback option in case the papers are lost or damaged, whereas controlledaccess to organized storage of the digital copies makes it much faster, easier and safer to find. Also, use the required documents, without physically touching / needing the original paper. Computer can process the script and recognized the characters [1]. Previous researches are state that there was a limitation of classical segmentation due to errors in recognizing unconstrained characters [10]. Later the extension of Optical Character Recognition (OCR) was applied in printed pattern which causes approximately 0.5% spacing errors [11]. The recent literatures are focuses accuracy by proposing innovative segmentation technique which is categories as explicit, implicit and segmentation free approaches [2]. After that final recognition is achieved by decision tree [2].In this proposed work segmentation follows Character Segmentation (Optical Character Recognition) System [3]. Here Algorithm used to separate the cursive characters by calculating distance between the letters using single line. The potential segmentation column is done by measuring height and width of scanned image segment and row & column wise pixel calculation.

Rajeev N. Verma. [4], proposed a technique to recognize anyone's handwriting, by removing skew in the characters. The recognition based segmentation was proposed [5], where images are divided into overlapping pieces. Also segmentation is applied in histogram of the images [6], where Characters are segmented by measuring minimal occurrence of pixels in histogram images. But these exiting algorithm fails in Character segmentation in open characters are affected and Ambiguous words are not been identified. In digital format is the first step of document digitization. The main aim of this work is to digitize the handwritten forms and the other handwritten materials to the digital form. Documents are been scanned and converted into individual images. By using the method of template matching on images with the comparison of the original plain documents with the details filled in documents. We are able to pick out the lines of details. For the process of template matching OpenCV provides us a way to perform. Then these individual images are been fed to the Neural Network Model and Machine learning algorithms to study the nature and understand the feature of the image. The usage of various algorithms to test and their accuracy and reasons why they have been failed are studied.

After training of the datasets the upcoming new words were predicted. Handwritten characters recognition is been carried out by following steps:

1)**PDF to Image Conversion** : In this module, conversion of the paper documents to the PDF, by scanning the individual papers and saving it as PDF. Various threshold activities on the PDF cannot be implemented. Conversion from PDF to the image using the PDF2img module in python is implemented. It's free of cost which has been able to take the individual page from the PDF and turns it into image in the desired format. To manipulate the image the PNG format is used.

2) Line Detection (Contour Detection) : In this module, process of finding the individual lines of the text is carried out by using the image manipulating library tool, openCV. Initially the image is been converted to gray scale in order to remove the noise with the help of COLOR_BGR2GRAY property. Now the gray scaled image is been transferred in to the binary format with thresh binary inverse method. To crop the individual lines a kernel with 5:500 ratio is used which is shown in Fig.2. Then the image is been set to find the contours that are available in the given page. In the scenario, the problem of ordering the contours occurs. Sort is done based on its location from top to bottom. Then the lines in the images will be bounded by rectangle box to crop it out. These manipulated images are used for both training and testing phases.

3) Image Preprocessing : The aim of pre-processing is to eliminate the inconsistency that is inherent in cursive handwritten words. The handwriting samples may be written on a noisy or colored background and also the quality of the word images may be degraded due to the noise that is introduced in the process of scanning or capturing the wordimages. It is necessary to remove the background noise to improve the quality of the word images to be used in the segmentation experiment. In this module, after the image has been obtained from line detection. They need to be preprocessed, before they are been passed to the predicting model. Initially need to remove the noise from the image. The image is been set to the binary threshold to remove the noise particle. Then apply the Inverse Binary Threshold. The image is been removed of some noise particle. This may lead to some distortion of the image particles, hence the image is dilated to 1px and it will retain the original letters. In order to provide segmentation, the entire letters in the image to be of single pixel. The skeletonization of the image pixels is performed. Since the image is binary (background: White, foreground: Black), skeletonization will affect only the black lines on the image. Then the image is been saved for character segmentation.

4) Character Image Segmentation : The process of Cursive Handwritten Segmentation follows the preprocessing and skeletonization of the image, will be fed into array of arrays (matrix). The matrix is been transposed, for easy calculation of the character segmenting points. Now traverse the entire matrix, row by row and calculate the average of the entire first element and the last elements. The average upper line and average lower line will be displayed. Also find the median of these two lines. Now during the segmentation of the cursive letters, the point of segmentation is the place where the sum of the entire column will be 255 (i.e.) a single black pixel will be present. Traverse the matrix again to calculate all the points where there will be only a single pixel.

There were some challenges in segmenting cursive handwritten word is a subject of much attention due to the presence of many difficulties such as: There can be variation in shapes and writing styles of different writers. Cursive nature of handwriting i.e. two or more characters in a word can be written connected to each other. In the view of cursive handwritten words, a ligature is a small link which is present between two successive characters to join them. Two consecutive 'i' Characters may give an illusion of the presence of a character 'u' and vice versa. Now the array of single points is been obtained. Need to find the consecutive array points and group them. From trial and error, if the group with points more than 7 points then their average is been taken and stored in another array. In order to save the lower cutting points in 'u', 'v', 'w'; we are using the average between average top line and median that we had calculated before as shown in Fig.7. In order to differentiate between two 'i' and 'u', we have a single point feature on upper end of 'i' and vice versa. Now, they are been passed to the next stage of prediction of each individual character. Handwritten characters can have more than one shape according to their position inside the word image.

Words may be written by a pen having ink of different colors. Some characters can give the illusion of presence of two similar characters. Letters "U" and "w" contain "intraletter ligatures," i.e., a subpart of these letters cannot be differentiated from a ligature in the absence of context. Artifacts sometimes cause erroneous Segmentation eg 'w' can be segmented into two 'v' and 'v' characters. There are two types of characters in English language. First type of characters are called Closed Characters and contain a loop or a semi-loop such as 'a' , 'b' , 'c' , 'd' , 'e' , 'g' , 'o' , 'p' ,'s' etc. Second type of characters are termed as Open Characters, it's very difficult to differentiate between ligatures and characters because of the cursive nature of handwriting. Generally, Letters "o,""b,""v," and "w" are followed by "upper" ligatures.

5) Character Prediction Model: In this module, the prediction of the individual character produced from the previous model. CNN is been used to train the model. For the training set of data with about 2K images of alphabets in both the cases from different people. Initially these images are been resized to same size and fed to the image preprocessing layer. Then the images with their labels are been shuffled and been trained with the CNN model. The model comprises of the 5 layers. First layer comprises of 2 Dimensional Convolution Layer of ratio (5,5) with relu activation then they fed in to max pooling of pool size (2,2) . Second layer and Third layer comprises of 2 Dimensional Convolution Layer of ratio (4,4) with the relu activation then they fed into max pooling of pool size (3,3). Fourth layer comprises of the Flatten module. Fifth layer comprises of softmax activation. Then the model is been compiled with the training set of images with batch size of 5 , epochs 100 and the training and testing split is 70 : 30 . With the trained model the characters from the previous steps are been tested and predicted

CONCLUSION

The proposed system has better character segmentation in cursive letters. The model is been trained with the 2k set of handwritten images which will enable the quality of the prediction methodology. This method could be used to digitize the handwritten English letters to digital format. There are some errors that occur in the character segmentation but this method is good for character detection in the application forms where they will have boxes to fill in. So by using template matching we can identify the forms and crop out the boxes and predict the values. Since the proposed system deals only with alphabets, we need a system to deal with numbers and symbols. In the future work an application that gets trained with the large data set of the alphanumeric letters so that the quality of the prediction will be higher. Improvement in the character segmentation algorithm is been needed so that the segmentation of the two 'i' and 'u' vice versa is been precise. With these combined technology forms will be template matched information is been transferred in to digital form.

REFERENCES

1) M. Gilloux, J.-M. Bertille, M. Leroux, "Recognition of Handwritten Words in a Limited Dynamic Vocabulary", Third Int'l Workshop Frontiers in Handwriting Recognition, pp. 417-422, 1993-May.
2) B.A. Yanikoglu, P.A. Sandon, "Off-Line Cursive Handwriting Recognition Using Style Parameters", Apr. 1993.A.W. Senior, Off-Line Cursive Handwriting Recognition Using Recurrent Neural Networks, Sept. 1994.
3) Gladence, L. Mary, M. Karthi, and V. Maria Anu. "A statistical comparison of logistic regression and different Bayes classification methods for machine learning." ARPN Journal of Engineering and Applied Sciences 10.14 (2015): 5947-5953.
4) R.G. Casey, E. Lecolinet, "Strategies in Character Segmentation: A Survey", Int'lConf. Document Analysis and Recognition, vol. 2, pp. 1,028-1,031, 1995-Aug.
5) Ch. N. Manisha ,E. Sreenivasa Reddy and Y. K. Sundara Krishna "Role of Offline Handwritten Character Recognition System in Various Applications" International Journal of Computer Applications 135(2):30- 33, February 2016
6) Amit Choudhary , Rahul Rishi , Savita Ahlawat , "A New Character Segmentation Approach for Off- Line Cursive Handwritten Words" Information Technology and Quantitative Management (ITQM2013).
7) Rajeev N. Verma, Dr. M. M. Raghuwanshi, "Comparative Study of Various Techniques on Handwriting Recognition and Analysis" 2018 3rd International Conference for Convergence in Technology (I2CT)
8) Pritam Dhande,Reena Kharat "A Survey of Methods and Strategies in Character Segmentation " on IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 7, July 1996

9) Fitrianingsih , Sarifuddin Madenda , Ernastuti , Suryarini Widodo,Rodiah . "Cursive Handwriting Segmentation Using Ideal Distance Approach", International Journal of Electrical and Computer Engineering (IJECE).

10) H.Akouaydi, S.Njah, and A.M. Alimi: Android Application for handwriting segmentation using PerTOHS theory, ICMV, Nice 2016.

11) H.Bezine, W.Ghanmi and A.M. Alimi: A HMM Model Based on Perceptual Codes for On-Line Handwriting Generation, COGNITIVE 2014: The Sixth International Conference on Advanced Cognitive Technologies and Applications, 2014.

12) J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

13) Gladence, L. Mary, M. Karthi, and T. Ravi. "A Novel Technique for Multi-Class Ordinal Regression- APDC." Indian Journal of Science and Technology 9.10 (2016): 1-5.

14) J. Schuermann, "Reading Machines," Proc. Sixth Int'l Conf. Pattern Recognition, Munich, Germany, 1982.

15) T. Nartker, ISRI 2993 Annual Report, Univ. of Nevada, Las Vegas, 1993