`

Handwritten Text to Editable Text Document

Ankur Garg
Btech, Department of CSE SRMIST Chennai,603203,India ankurgarg.0496@gmail.com
Payal Deora
Btech, Department of CSE SRMIST Chennai,603203,India payal.deora1997@gmail.com
D.Malathi
Professor, Department of CSE SRMIST
Chennai,603203,India malathi.d@ktr.srmuniv.ac.in

Abstract-

One of the major problem faced by every organization is management of old records of handwritten text that are likely to be subjected to further deterioration in the future. These old records are difficult to manage because of the sheer volume they exist in. This issue can be rectified if there existed a softcopy of those records but the Analysis of Handwritten text has been one of the major challenge in the field of image processing because of various writing styles, background lighting and text orientation. All the existing technique have failed when the letters in a text cannot be segmented properly. In our project, the text will be segmented and then segmented letters will be analysed and classified into appropriate class using a neural network .With the aid of Natural Language processing, we can analyse the context of a misclassified word and predict its occurrence in reference to the context. With a proper blend of image processing, Natural language processing and Convolutional Neural Network, we will recognize the text with a high degree of accuracy.

Keywords: *SVM, RNN, Shear Transformation, Gaussian filter, Skeleton coding, Otsu Thresholding, Concavity, Discourse analysis, Skew transformation, Hough tranform*

I. INTRODUCTION

Handwritten text recognition is the ability of a computer to interpret intelligible handwritten input from various sources such as a paper document, photographs touchscreen and other touch responsive devices. This problem can be tackled by using an online approach such as one used in smartphones, which are ultimately leading keyboard devices to the state of disuse. Online approaches require the use of a special digitizer or PDA, which helps in automatic conversion of text. However, this paper deals with the offline approach of text extraction and conversion because the major problem arises when the storage of offline documents becomes impossible as with time paper gets deteriorated.

The strategy used for offline recognition is an analytic strategy in which recognition is started from a character level. After that characters are combined and recognized as a whole word and words are combined into sentences. Most of the offline approach applied various pre-processing techniques to clean the data to the point such that it could be trained. A pre- processing technique will involve conversation of an image to greyscale, applying slant correction, baseline correction and then finding a segmentation region.

After that, the recognition of character is done. However, no recognition can output 100% accuracy because sometimes humans make errors while writing. Only reason humans are able to pick up these errors is that of our ability to sense the anomaly in the continuity of words. We have an idea of what might replace the word.

Some of the papers are having very few post-processing techniques to boost their accuracy. This paper proposes a post- processing technique by using a Natural Language Processing algorithm of discourse analysis. Discourse analysis will be the last stage of recognition, which will help to classify miss- classified words based on the stored history of trained data. Our idea is to integrate discourse analysis with neural network model to achieve high accuracy.

II. OVERVIEW

This paper depicts the working of a complete offline handwriting recognition system for general conversion of text into computer understandable form. The system divided is into multiple phases namely training, pre-processing, segmentation, recognition and post processing.

The data is captured by using any smart device capable of taking image of text document with optimal resolution. This image inputted to pre-processing stage to remove noises and deskew the text. The processed image is passed through multiple stages of segmentation namely line segmentation, word segmentation to finally extract individual character from each word. Then a series of image processing

`

operations is done to make the word invariant to some of the distortions that affect recognition of text. Each processed word is segmented into individual characters, which are fed to the recognition engine for output as character belonging to one of the many output classes.

Each recognised character is consolidated to form the entire text. To inhance the accuracy of the model, we are implementing discourse analysis. Any misclassification done by the algorithm is corrected by taking into account the context of the misclassified word. The results looks promising, as humans too adopt this technique when they come across an uncanny word.

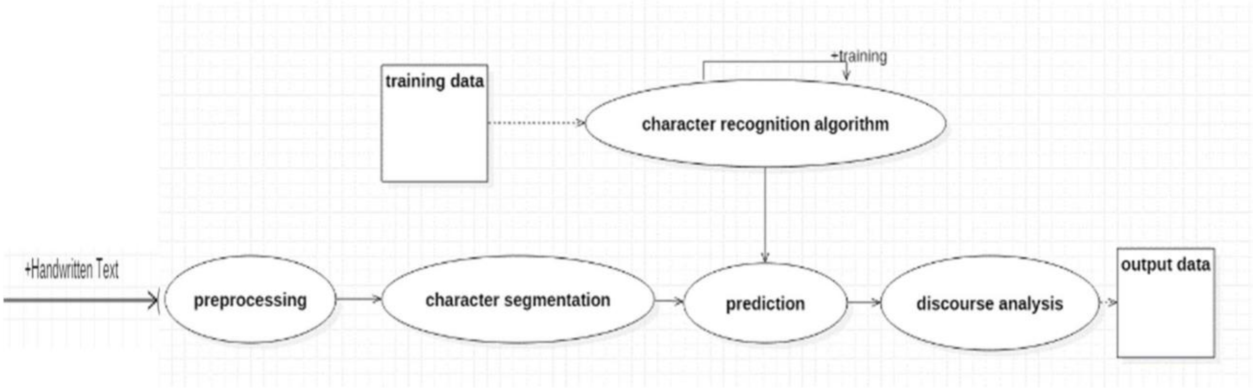


Fig. 1. A Schematic for Recognition System

III. PREPROCESSING

The system takes scanned image as input from smart devices such as smartphones .Each image fed to our algorithm is subjected to be vary depending on the image quality, ambience brightness, paper quality and ink color. Hence, pre-processing becomes a critical stage to greatly simplify the segmentation and boost up the accuracy of the system.

The input image is converted from RGB to Grayscale format. Then using Adaptive thresholding, Grayscale image is converted to Binary format. Adaptive thresholding find the threshold of different region in text separately instead of finding a global threshold and hence is extremely helpful when different area of text are illuminated differently.

Image opening is then performed to remove noise in the image.

The input text can be skewed from the horizontal axis because not all images are clicked perfectly horizontal. This skew must be eliminated for proper segmentation of text into individual lines. This is done by dilating the text in horizontal directing and finding the angle of minimum bounding rectangle around the dilated text. The Rectangle is then rotated according to the angle to dekew the original text.

IV. SEGMENTATION

Text segmentation is the process of dividing written text into smaller units, such as sentences, words and characters. Each segmentation phase has its own pre- processing.

1. Line Segmentation-

   Line segmentation is done to extract individual lines from the text. Problem can arise when the descender of upper text text joins the ascender of the lower text
   The heuristics used for line segmentation consist of following steps-

   1.1. Disjoin lines joint by descenders and ascenders in the text.

   1.2. Calculate the vertical density histogram, shown in Fig. 2, by counting the number of black pixels in each horizontal line in the image.

   1.3. Region between two lowest values in histogram is identified as a line and segmented from the text.

   1.4. Lines extremely small in width are classified as noise and not taken into account.

2. Word Segmentation

   After performing the line segmentation, each word should be segmented .The major problem faced while word segmentation is when distance between two characters in a word is comparable to distance between two words.

The heuristics used for word segmentation consist of following steps-

2.1. Dilating the words to join the characters within the same word in a line image

2.2. Calculate the horizontal density histogram, shown in Fig. 2, by counting

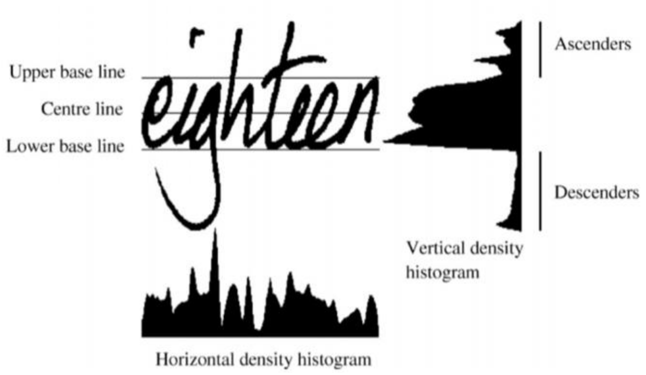


Figure 2 Histograms, center line, and baselines of a deslanted word

2.3. the number of black pixels in each vertical line in the image.

2.4. Region between two minimums in histogram is identified as a word and segmented from the line.

2.5. Boundaries are used to calculate the intermediate distance between words to recognize if two separated entities are actually words or largely spaced characters

2.6. Intermediate distance between all possible words is calculated and average of distance is taken

2.7. Percentage of difference between minimum and maximum word distance given the maximum distance is calculated

2.8. If the percentage is greater than a threshold set manually, we know that a word itself contain character which are separated by spaces.

2.9. Loop over the number of words and check the average distance between words is less than the distance between two intermediate word currently being looped over.

2.10. If the average distance is more, we consider the two separated words as two separated characters of the same word

Before the output of word segmentation can be provided as an input to character segmentation, baseline of a word must be estimated and the slant angle of it must be corrected.

3. Slant Correction and Baseline Estimation

The slant angle is estimated by finding the average angle of near-vertical strokes in a word. This is estimated by finding the edges of strokes, by using an edge-detection filter. This will output the image as a chain of connected pixels representing the edges of strokes. Each word is then divided into multiple parts equally. The mode of alignment of those edges close to the vertical on each part is used as an overall slant estimate (Fig. 5d). Edge orientations are estimated with a Canny edge detector.

The heuristic used for baseline estimation consists of the following steps:

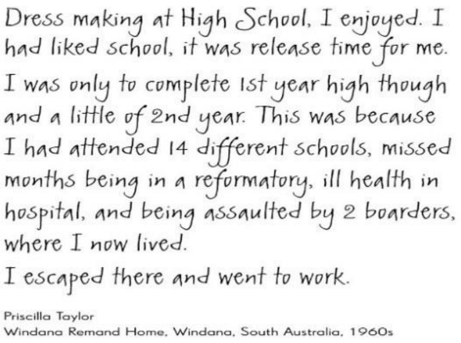
3.1. Calculate the vertical density histogram, shown in Fig. 2, by counting the number of black pixels in each horizontal line in the image.

3.2. Descenders in a word are identified by finding the peak in the histogram and minimum above this peak is found.

3.3. A straight line is drawn through this minimum which is now the lower baseline.

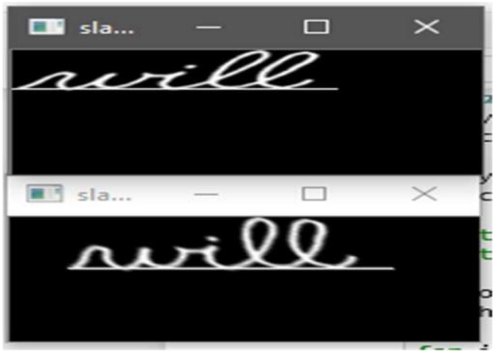
Upper baseline can be similarly estimated but it is less accurate due to presence of t strokes.

Slant of the word is corrected by shear transforming the image in horizontal axis.



Dress making at High School, I enjoyed. I
had liked school, it was release time for me.
I was only to complete 1st year high though
and a little of 2nd year. This was because
I had attended 14 different schools, missed
months being in a reformatory, ill health in
hospital, and being assaulted by 2 boarders,
where I now lived.
I escaped there and went to work.

Priscilla Taylor
Windana Remand Home, Windana, South Australia, 1960s



4. Character Segmentation and Recognition

Segmenting word into character is last phase of segmentation. The slant corrected word is obtained as an input in this phase-

Tesseract is used to perform character segmentation and recognition.

Following steps were performed to improve tesseract accuracy-

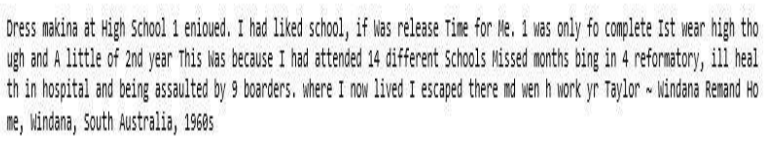
4.1. Tesseract configuration is set to recognize single word. This allow us to optimize the algorithm and get better results.

4.2. Background is isolated from foreground to reduce noise.

4.3. size of image is compressed to do faster processing

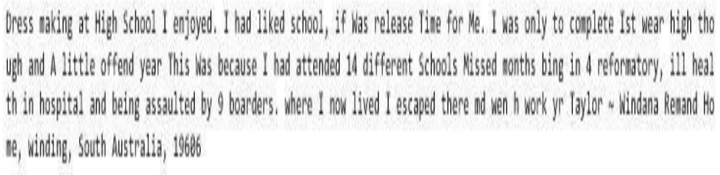
4.4. Adding border around the word image aids in segmentation.

4.5. Binarization - Tesseract does this internally (Otsu algorithm), but the result can be suboptimal, particularly if the page background is of uneven darkness. In our project, adaptive threshoding was yielding better results.



```
Dress makina at High School 1 enioued. I had liked school, if Was release Time for Me. 1 was only fo complete Ist wear high tho
ugh and A little of 2nd year This Was because I had attended 14 different Schools Missed months bing in 4 reformatory, ill heal
th in hospital and being assaulted by 9 boarders. where I now lived I escaped there md wen h work yr Taylor ~ Windana Remand Ho
me, Windana, South Australia, 1960s
```

Figure 5. Output of tesseract



```
Dress making at High School I enjoyed. I had liked school, if Was release Time for Me. I was only to complete Ist wear high tho
ugh and A little offend year This Was because I had attended 14 different Schools Missed months bing in 4 reformatory, ill heal
th in hospital and being assaulted by 9 boarders. where I now lived I escaped there md wen h work yr Taylor ~ Windana Remand Ho
me, winding, South Australia, 19606
```

Figure 6. Output of discourse analysis

V. DISCOURSE ANALYSIS

Discourse analysis is the last stage of the project. The recognized text may contain some misclassified words. Discourse analysis includes correction of words which are misclassified by the recognition model. This is done by having the knowledge of the history of the sentence. This misclassification can be corrected if we know the context of word in relation to the sentence.

A huge dataset of sentences are used to train our Recurrent Neural Network. A corpus of documents related to law is maintained. The sequence length of the Network is set to 50. It means that the network will use the previous 50 words to predict the 51st word. Each word is assigned some probability depending on its occurrence. Words that are not found in corpus are given a low probability instead of zero

Our text is passed to the trained algorithm in batch of 50 words. The 51st word is predicted and if it is different from the word in the predicted text, it is replaced. Hence, some of the misclassification are corrected using discourse analysis.

Also there can be some misspelled words which are recognised by the model. These misspelled words will also be corrected by the last stage of the project. This will improve the efficiency of the model by reducing the error of classification.

VI. CONCLUSION

This paper shows a complete offline handwriting recognition model. We have combined Natural Language Processing with Computer Vision to recognize text, which is similar to what humans do to read text. For now, the corpus contain words from documents concerned with law due to memory and processing power constraints. The results can be further improved if the word corpus can be expanded for different domains.

REFERENCES

[1]. Andrew W. and Anthony J. Robinson, "An Off-Line Cursive Handwriting Recognition System", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, March 1998.

[2]. Nafiz Arica, and Fatos T. Yarman-Vural,"Optical Character Recognition for Cursive Handwriting" IEEE Transactions on Pattern Analysis and Machine Intelligence. 24, No. 6, June 2002.

[3]. Sajjad S. Ahranjany, Farbod Razzazi , Mohammad H. Ghassemian "A Very High Accuracy Handwritten Character Recognition System for Farsi/Arabic Digits Using Convolutional Neural Networks" IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010.

[4]. Martin Rajnoha, Radim Burget, Malay Kishore Dutta "Offline Handwritten Text Recognition Using Support Vector Machines", 2017 4th International Conference on Signal Processing and Integrated Networks.

[5]. Gyeonghwan Kim1, Venu Govindaraju2, Sargur N. Srihari2 "An architecture for handwritten text recognition systems" International Journal of Document Analysis and Recognition, Volume 2, , pp 37–44 1999.

[6]. J.Pradeep, E.Srinivasan , S.Himavathi "Neural Network based Handwritten Character Recognition system without feature extraction", International Conference on Computer, Communication and Electrical Technology – ICCCET 2011, 18th & 19th March, 2011.

[7]. Anshul Gupta, Manisha Srivastava, Chitralekha Mahanta. "Offline Handwritten Character Recognition Using Neural Network", International Conference on Computer Applications and

Industrial Electronics, 2011.

[8]. Nafiz Arica and Fatos T. Yarman-Vural "An Overview of Character Recognition Focused on Off-Line Handwriting", IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 31, No. 2, May 2001.

[9]. Youssouf Chherawala, Partha Pratim Roy, and Mohamed Cheriet "Feature Set Evaluation for Offline Handwriting Recognition Systems: Application to the Recurrent Neural Network Model", IEEE Transactions on Cybernetics, Vol. 46, No. 12, December 2016.

[10]. C. Y. Suen, C. C. Tappert, and T. Wakahara, "The state of the art in on-line handwriting recognition," IEEE Trans. Pattern Anal. Machine Intelligence., vol. 12, pp. 787–808, Aug. 1990.