International Journal of Innovative Science and Modern Engineering (IJISME)
ISSN: 2319-6386 (Online), Volume-6 Issue-7, May, 2020

Enhancing Predictability of Handwritten Document Content using HTR and Word Substitution

Varshini Prakash, Keshav Moorthy, Jasmin T Jose

Abstract*: Handwritten Text Recognition (HTR) can become progressively abysmal when the documents are damaged with smudges, blemishes and blurs. Recognition of such documents is a challenging task. We, therefore propose a system to identify textual handwritten content in documents where the state-of-the-art Optical Character Recognition (OCR) existing at its full extent performs with low accuracy. By introducing word substitution using character and distance analysis for spell checking and word completion in such areas for giving out more accurate results using a word corpus, we improved our prediction results especially in cases where the OCR is prone to predict false positives on the smudge areas predominantly. Blur detection on every word before segmentation is also substituted with a new word by our OCR algorithm to avoid false positive results and are instead substituted with suitable words. This methodology is far more convenient and reliable since even state-of-the-art HTR technologies do not have more than 71% accuracy. The accuracy of the predicted test is measured using the text similarity metric - Fuzzy Token Set Ratio (FTSR).

Keywords*: Damaged documents, Fuzzy Token Set Ratio, Handwritten Documents, Spell Check, Word Replacement

I. INTRODUCTION

Smeared documents are those that are hand written or printed hard documents that are exposed to the environment and get destroyed because of foreign objects like liquids, dust and dirt. These foreign objects either cause smudges and blemishes or obfuscate certain characters which cannot be identified using Optical Character Recognition (OCR). This project focuses on identifying the text from these smeared documents. OCR for handwritten documents is still a growing challenge and one way to tackle this problem is by combining it with Natural Language Processing (NLP) for sentence completion until OCR can become mature enough to identify texts from various handwritings, symbols and styles of writing. This can take a long time to solve given the various ways in which humans write different characters. Cursive handwriting recognition [1] is the closest we have got to for OCR but that is not significant enough. Segmenting cursive handwriting recognition is a procedure for removing singular characters from handwritten words, which may vary in size and shape. A novel Binary Segmentation Algorithm (BSA) is presented in [2] that decreases the dangers of the chain failure problems during validation and improves the segmentation precision. Handwritten documents are abundantly available and have been the biggest problem with OCR. However, they are never uniform, aligned, well maintained and sometimes they can be so damaged that even us humans cannot interpret the writing. In such cases it is impossible to purely use OCR for developing solutions to recognizing the text content in these documents and require a far more superior Artificial Intelligence with multiple algorithms working on both OCR and word prediction for correcting corrupted OCR characters. There have been significant developments, especially for historical documents, in identifying handwritten manuscripts and texts using OCR, [3][4][5] and a few researches have used algorithms such as Part of Speech Tagging and some of them have collectively added features using Joint Feature Distribution in adding features to the OCR which could be specific and discriminative to only certain documents. There is a lack of research in significantly using NLP for analysing smudge areas and using NLP algorithms for completing words in these missing areas. There have also been significant improvements in segmentation technologies focusing on historical manuscripts but these are also very specific and the dataset these have been tested on are finite.

II. EXISTING SYSTEMS

There have been significant developments, especially for historical documents, in identifying handwritten manuscripts and texts using OCR and a few researches have used algorithms such as Part of Speech Tagging [6], which is evaluated on data that cannot be interpreted using Optical Character Recognition (OCR). German and English text data, NLTK taggers and recognized text to significantly increase the error rates. Recent Deep learning approaches that utilize lexicon-based architectures and recurrent neural networks have considerably enhanced handwriting recognition. A completely convolutional network architecture which outputs arbitrary length symbol streams from handwritten content in [7]. Sheng and Schomaker have collectively added features using Joint Feature Distribution in adding features to the OCR which could be specific and discriminative to only certain documents [8].

Revised Manuscript Received on April 22, 2019.
* Correspondence Author
Varshini Prakash*, Computer Science, Vellore University of Technology, Vellore, India. Email: varshini.prakash2016@vitstudent.ac.in
Keshav Moorthy, Computer Science, Vellore University of Technology, Vellore, India. Email: keshav.moorthy 2016@vitstudent.ac.in
Jasmin T Jose, Computer Science, Vellore University of Technology, Vellore, India. Email: jasminlijo@vit.ac.in

The ability to restore a degraded document to its ideal condition would be highly useful in a variety of fields such as document recognition, search and retrieval, historical document analysis, law enforcement. Wiener Filer algorithm is used for noise removal from degraded handwritten documents using techniques such as de-blurring and histogram equalisation [9].

There is a potential for NLP for predicting words which are based on their associative relations assessed in the Serbian language dataset [10]. This approach is to use a number of different written documents and materials, and sees if a possible neural network can extract the word associations just by reading these regular texts. There is a lack of research in significantly using NLP for analyzing smudge areas and using sentence completion algorithms for completing words in these missing areas. There have also been significant improvements in segmentation technologies focusing on historical manuscripts but these are also very specific and the dataset these have been tested on are finite. A simple coffee spill over such datasets can reduce the capabilities and accuracy of these algorithms.

III. PROPOSED FRAMEWORK

Document analysis primarily consists of four modules, namely text line segmentation, word extraction, character dissolution and script recognition. This project focuses on identifying the text from smeared documents. The dataset used for this experiment is generated from IAM Dataset consisting of English texts from LOB corpus, explained in detail in [11]. The handwritten documents from IAM is randomly superimposed with images of dirt and blur noises to generate data used to train and test handwritten text the recognizers and to perform writer identification and verification experiments. The proposed framework can be divided into four major steps, as follows:

A. *Pre-processing and Segmentation*

Since the images we use are said to be dirty and unclear some amount of pre-processing is required before sending the images to OCR for better accuracy. For this purpose, we perform dilation and then erosion (in the same order) after which we perform word segmentation. For word segmentation, we find contours of continuous characters that separate one word from another and draw bounding boxes across these contours.

B. *Blur Detection*

Usually the OCR performs abysmally on blurred words and sometimes returns false positive results. The accuracy of a blurred character can hence not be trusted and so we detect words that have a significant blur for word substitution. Blurs are detected uses a variation of Laplacian operators for sharpness detection across the edges using the formula in fig.1

Fig.I Performing blur detection on segmented words

C. *Handwriting Text Recognition*

For Handwritten Text Recognition, we propose a transfer learning methodology approach using an image-based sequence recognition algorithm [12] which runs on six layers of CNNs that help with feature extraction and two layers of RNNs. An additional layer of CNN with 32x256 is added for feature extraction. The HTR returns text from the entire document with unavoidable errors.

D. *Word Substitution*

Spellchecker is used to identify and correct misspelled words. This performs particularly well in correcting high frequency words. It can be customized to a specific corpora by understanding the nouns and the frequency of words [13]. There is a feature to exclude nouns in the list of unknown words from English Language, or from a particular corpus. When an unknown word is identified, it is substituted with the correct spelling by choosing the closest word with most similarity from a candidate word list, which it generates. The similarity rate of the spellchecker also be customized, making the system's performance more rigid by increasing or flexible by decreasing the similarity rate. Most often, a relatively rigid system is preferred to avoid inaccurate corrections with low confidence.

IV. THEORETICAL BACKGROUND

Bounding boxes are drawn over texts that are smudged and unclear to specify this region. The blur detection uses various Laplacian operators for identifying sharpness of edges.

$$L = \frac{1}{6}\left(\begin{array}{*{20}{c}} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{array} \right) \quad (1)$$

$$LAP\_VAR(I) = \sum\limits_m^M {{\sum\limits_n^N {\left[{\left| {L(m,n)} \right| - \overline L } \right]} }^2} \quad (2)$$

The second derivative is used for high spatial frequencies. So, the equation $[L]$ in (1) represents the operator. The higher frequencies correspond to sharper edges. (2) is used for pooling the operation data at each point of the image. $[L(m,n)]$ is the result of convolution of the given input image that is represented by $[I(m,n)]$ where m,n are the dimensions of the image.

The mean of the values is represented by \[\overline L] which can be represented by (3).

\[\overline L = \frac{1}{NM}\sum\limits m^M\sum\limits n^N{\left[\left| {L(m,n)| }]}] (3)

The architecture used to predict handwritten text is an extension of the base generally used for character recognition. The network is a combination of Convolutional Neural Networks and Recurrent Neural Network, integrating the advantages of both the networks.

RNNs can remember sequences because their architecture has feedback loops, which helps in retaining the information [14]. However, they can't retain long term dependencies and fail to learn this successfully. RNNs failure to overcome long term dependencies is overcome by LSTMs (Long Short Term Memory Networks). They differ in the repeating module in their architecture, which is more complex in LSTMs and is therefore capable of retaining long dependencies [15].

Furthermore, Fuzzy Token Set Ratio (FTSR) is used a text similarity metric for evaluation because of the flexibility of the approach in the context of dealing with certain damaged texts that are incorrectly recognized by the HTR. The strings are tokenized and split into intersection and reminder, the similarity ratio is high when the tokens common to both strings form a higher percentage of the string and when the string reminders are similar [16].

V. RESULT AND DISCUSSION

The results obtained use FTSR as the text similarity metric to calculate the similarity between the original text from IAM corpus and recognized test using HTR system. The mean text similarity is calculated for over 500 degraded documents from the generated dataset. Table I. compares the mean text similarity for these documents with text recognized by HTR, the recognized text then corrected with word substitutions and the text recognized using state-of-the-art Google Vision. We can observe that the proposed model performs monumentally better when directly compared with Google Vision's handwriting recognition model. In comparison to our model, the Google Vision model specifically fails to handle and recognize smudged documents accurately, as seen in fig. III.

As seen from the Fig II, Google Vision almost predicts every word accurately except for smudged words like "in fact" and "must Britain". When passed with documents with higher degree of blurry and smudgy characters the model fails entirely. In these places, our word substitution formula works better and the results are documented in table I.

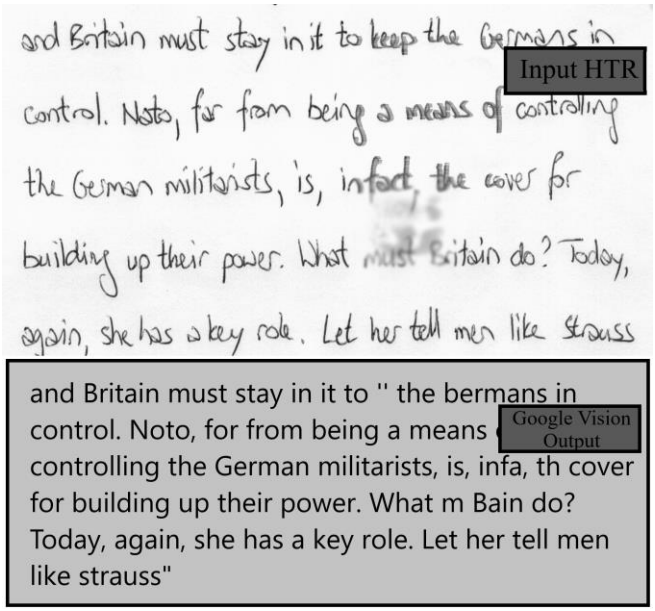

Fig.II An example of Google Vision Model's output for smudged handwritten document

Table I. Text similarity comparison of our model against others

Recognized Text	Mean Text Similarity
HTR	77.14
HTR after Word Substitution	80.13
Google Vision	18.51

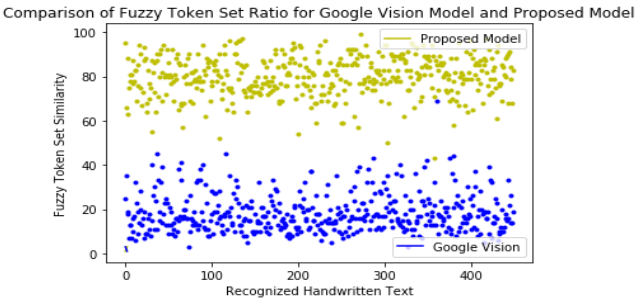

Fig.III Evaluating text similarity on state-of-the-art Google Vision Model's handwriting recognition and proposed model's recognition.

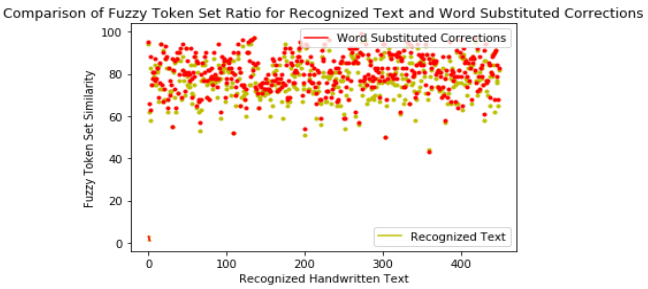

Fig. IV A comparison of accuracy for HTR recognized text and text corrected with Word Substitution.

The enhanced performance of the system once correction strategies are employed post OCR is depicted in Fig IV. This results in improved word level precision as shown in Table I.

The FTSR is the metric used for evaluating the text similarity of the output text along with the correct document text. It is more convenient to be used on handwritten text content considering how many of the handwritten documents may contain certain characters and strikeovers that are not meant to be recognized and our models would return garbage non-alpha and non-numeric results for these. The FTSR avoids such characters when testing.

VI. CONCLUSION

The proposed model is capable of recognizing text from degraded Handwritten documents using a combination of Image Processing techniques such as preprocessing, segmentation and blur detection. The recognized text is further rectified using Word Substitution to correct misspellings. Although our model cannot be relied upon for absolute accuracy towards handwritten data, it is a breakthrough into one of the biggest challenges of OCR in today's world, which is handwritten text recognition, specifically for degraded documents. The challenge also extends to historical document text recognition which has garnered attention from the academia in the recent times. Considering the unpredictability of human handwriting errors and other mistakes that can make it even harder for a human to interpret the data, it is very difficult to train a single OCR model and so approaches that make use of multiple NLP techniques such as word prediction and substitution can enhance the reliability of such systems. There has been limited research in combining grammar or spelling correction techniques along with text recognition, in general. Future work on this topic can be extended to grammatical error correction, a challenging topic which is an ongoing research theme in the linguistic community. Furthermore, the blurred words can be predicted using sentence prediction techniques.

REFERENCES

1. Choudhary, U., Bhosale, S., Bhise, S., & Chilveri, P. (2017, May). A survey: Cursive handwriting recognition techniques. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 1712-1716). IEEE.
2. Lee, H., & Verma, B. (2012). Binary segmentation algorithm for English cursive handwriting recognition. *Pattern Recognition*, *45*(4), 1306-1317.
3. Chammas, E., Mokbel, C., & Likforman-Sulem, L. (2018, April). Handwriting recognition of historical documents with few labeled data. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)* (pp. 43-48). IEEE.
4. Soni, R., Shekhawat, P. S., Jangir, R., & Saini, S. K. (2016, March). An efficient method to enhance the readability of historical handwritten artifacts. In *AIP Conference Proceedings* (Vol. 1715, No. 1, p. 020069). AIP Publishing.
5. Elfattah, M. A., Abuelenin, S., Hassanien, A. E., & Pan, J. S. (2016, November). Handwritten arabic manuscript image binarization using sine cosine optimization algorithm. In *International conference on genetic and evolutionary computing* (pp. 273-280). Springer, Cham.
6. Mieskes, M., & Schmunk, S. (2019, July). OCR Quality and NLP Preprocessing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Florence, Italy* (pp. 102-105).
7. Ptucha, R., Such, F. P., Pillai, S., Brockler, F., Singh, V., & Hutkowski, P. (2019). Intelligent character recognition using fully convolutional neural networks. *Pattern Recognition*, *88*, 604-613.
8. He, S., & Schomaker, L. (2017). Beyond OCR: Multi-faceted understanding of handwritten document characteristics. *Pattern Recognition*, *63*, 321-333.
9. Kaur, D. (2015). Remove Noise Effects From Degraded Document Images Using Matlab Algorithm.
10. Grujic, N. D., & Milovanovic, V. M. (2019). Natural Language Processing for Associative Word Predictions. IEEE EUROCON 2019 -18th International Conference on Smart Technologies. Presented at the IEEE EUROCON 2019 -18th International Conference on Smart Technologies. https://doi.org/10.1109/eurocon.2019.8861547
11. U-V Marti and Horst Bunke. The IAM-database: an English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition, 5(1):39–46, 2002.
12. Shi, Baoguang, Bai, Xiang, Yao, & Cong. (2015, July 21). An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. Retrieved from https://arxiv.org/abs/1507.05717
13. Rand, S., & Lall, R. (2019). Development of a Custom Spell-Checker for Emergency Department Data. *Online Journal of Public Health Informatics*, *11*(1).
14. Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.
15. Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM.
16. Wang, J., Li, G., & Fe, J. (2011, April). Fast-join: An efficient method for fuzzy token matching based string similarity join. In *2011 IEEE 27th International Conference on Data Engineering* (pp. 458-469). IEEE.

AUTHORS PROFILE

Varshini Prakash is an undergraduate Computer Science and Engineering student from Vellore Institute of Technology, Vellore. Her area of interests includes exploring Deep Learning based solutions for Computer Vision and Natural Language Processing problems.

Keshav Moorthy is a student studying at Vellore Institute of Technology, Vellore. His main interests lie in Image Processing and Natural Language Processing.

Jasmine T Jose is currently working as an Assistant Professor in School of Computer Science and Engineering, at Vellore Institute of Technology, Tamil Nadu, India. She received B.E. in Computer Science and Engineering from Anna University, Chennai, India in 2005, MTech. in Computer Science and Engineering from NIT Calicut, Kerala, India in 2013. She is pursuing Ph.D. from VIT University, Vellore, India. Her research interest includes computer vision, video processing and Image processing. She has published many research papers in international conferences and journals. She is a life member of ISTE.