A Printed And Handwritten English Scanned Document Text Segmentation Using Deep Learning Method

[1]Mr Roshan D Suvaris , [2]Dr. S Sathyanarayana

[1]Research Scholar Bharathiar University Coimbatore, Asst. Professor AIMIT, St Aloysius College Mangalore.

[2]Srikantha First Grade Womens College Mysore.

Abstract:
The detection of text in a document is critical step in all text recognition systems. The state of the art techniques to locate the words are grounded on the handcrafted heuristics which is fine tuned by the image processing communities' experience. They only work if specific conditions are met, such as a relatively consistent background. We propose deep learning technique to recognize the text in printed and handwritten text documents. Our method can be used to process damaged documents or documents with variety of backgrounds. Experimental results show that our method perform better than older methods with slight engineering efforts and fewer parameter tuning. The method is tested on different types of scanned image documents and it we got an accuracy of 96.5%.

Keywords: word segmentation, text segmentation, deep learning.

1. Introduction

Text segmentation in computer world is a method which divides the written text in newspapers, banners, images videos etc. into meaningful units such as lines, words. The term is applied to mental processes while reading the text by humans and to artificial processes which is implemented by the computers. The text segmentation process is very complex in languages in which there is no gap between the words such as Chinese, Vietnamese and Japanese. In languages which supports spaces to delimit the words such as English and Russian depending on white spaces alone does not result in satisfactory segmentation because the noise present in the background.

We are in the fourth industrial revolution, digitising text document has become a critical undertaking. The text from these digital documents are extracted for different purposes such as to make editable documents so that people can search required information, to assist blind people by converting words into speech etc.

Text segmentation not only used to extract the text from documents, it also used in images, videos, camera captured pictures, banners, sign boards, number plates, Captcha in websites etc. To segment the text different methodologies have been developed

There are lot of methods already developed by different researchers to segment the words in the document. In this method we are applying deep learning method to segment the words in the images with different backgrounds. Our method can be applied to any document, image, banners, and sign boards to extract the words.

2. Literature Survey

Sunanda Dixit and Kanchan Keisham [1] proposed a method to segment the words in handwritten English documents. The method uses the information energy of the pixel to segment the line and words and characters are segmented using vertical histogram plotting and by finding the local minima. Finally the characters segmented are predicted using artificial neural network(ANN).

Batuhan Balci et.al [2] proposed a method to recognise the hand written text using Convolutional Neural Network (CNN) to train a model that can classify the words. Once the words are segmented then the bounding boxes are created around the character and character recognition is performed using Long Short Term Memory networks (LSTM) with convolution.

Van-Linh Pham et.al [3] proposed method to recognize the words in newspapers. Firstly the image is converted to grey scale mode which is followed by the Hough line transformation to find the candidate angles. The segmentation of the lines in the newspaper is performed using U-net which is built using encoder and decoder principle. The logo varies for different newspapers to overcome template matching is used in this research paper.

N priyanka et al.,[4] proposed method to segment line and words from Indian script such as Devanagari, Bangla, Telugu and Kannada. First the image is binarized and then lines are segmented using run length smearing is applied to increase the histogram strength. Horizontal histogram of text lines are generated and text lines are segmented. After text line segmentation the lines are vertically and vertical projection profile is generated and the words are segmented. Using this method the researchers have achieved an good overall accuracy.

A. M. Vil’kin et al.,[5] proposed a method to segment the text blocks in scanned documents. In this method the scanned document is first divided into blocks and texture features has calculated for each block. In the first phase different locations and dimensions of blocks, 26 texture features and for the classification of the blocks four algorithms are used namely adaboost, SVM, ANN and kNN. In the second phase blocks are readjusted on the basis of neighbouring regions or blocks.

Mark Johnson, Anne Christophe et al.,[6] proposed a method for word segmentation. In this they have modified existing method adapter grammar based Bayesian model to permit it to study sequences of monosyllabic words at the start and at end of collocation of words. Using this modified method they have achieved a 4% increase in accuracy.

Zeeshan Bhatti, Imdad Ali Ismaili et al., [7] proposed method segments Sindhi text. In this method sentences are extracted using fully stop and question mark. The words in these sentences are segmented using break iterator which finds the hard space between the words in

Sindhi language and tokens are generated. The extracted tokens are once again checked to remove any special characters and stored in the repository.

To segment the words in multiple languages Yan Shao et al., [8] proposed a method which they have used sequence tagging model. The segmentation framework BiRNN-CRF is adopted which is complemented with attention based sequence to sequence transducer no segmental multiword tokens. In this method six typological factors are used to characterise the trouble of segmentation between various languages. They have achieved good accuracy for languages with spaces as delimiter but relatively more dataset is required for languages without spaces.

Bastien Moysset et.al [9] proposed a method to segment paragraph text with recurrent neural network. To describe the dispersed and local phenomena of the input images a LSTM network is used. The LSTM network is applied in parallel in all the four directions and the output of it is merged. The method also used Stochastic Gradient Descent for training the recurrent neural network. The method doesn’t label the boundaries of segmented lines and the method is applied on the Maurdor database.

Dongyang wang et al., [10] proposed hybrid deep learning network to segment the English words. The method used Bidirectional Gated Recurrent Unit (BI-GRU) for segmentation of words and Conditional Random field to interpret sentence in sequence. This method reduces the prediction time and the training time. The methods processing effect is similar to BI-LSTM-CRF but processing speed is higher than BI-LSTM-CRF.

3. Our Methodology:

To recognize the words in the text document we have proposed the following methodology which is given in the fig 1.

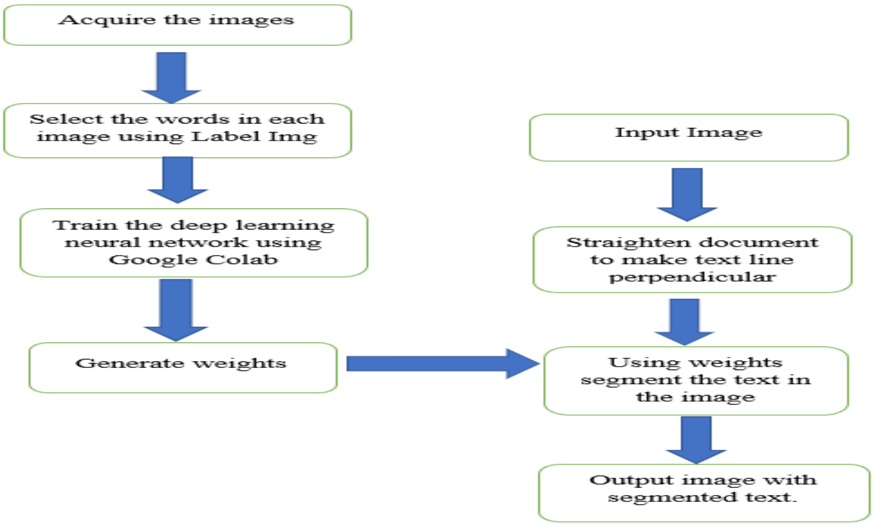
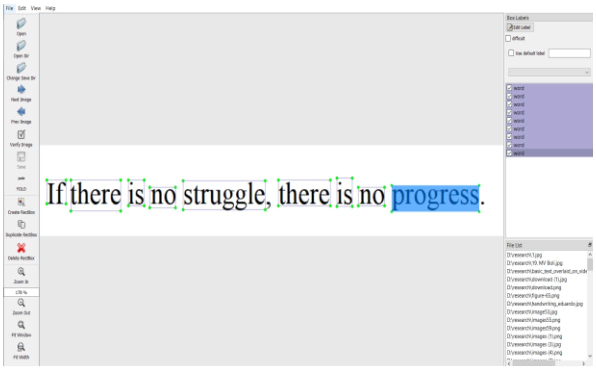

Acquire the images
Select the words in each image using Label Img
Train the deep learning neural network using Google Colab
Generate weights
Input Image
Straighten document to make text line perpendicular
Using weights segment the text in the image
Output image with segmented text.

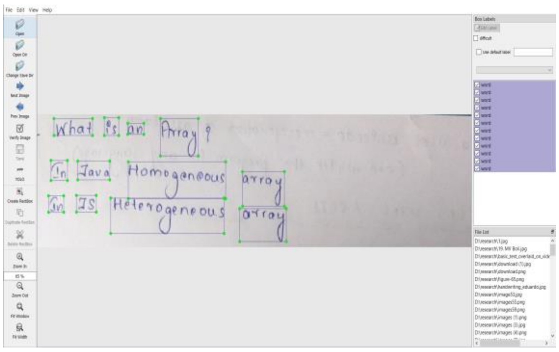
Fig 1: Flowchart diagram of our proposed system.

To segment the words in printed or handwritten scanned documents using deep learning method dataset should be created. The dataset is created using LabelImg tool. As to create our dataset we collected nearly 400 images that contains both printed and handwritten text in the document images. The LabelImg tool creates text file which contains the x, y, x+w and y+h coordinates for the each selected text regions. The text file contains entry for each individual word in the document and for each document image it contains the corresponding text file.



If there is no struggle, there is no progress.

Fig 2: Labelling the words in Printed document



What is an Array ?
In Java Homogeneous array
In JS Heterogeneous array

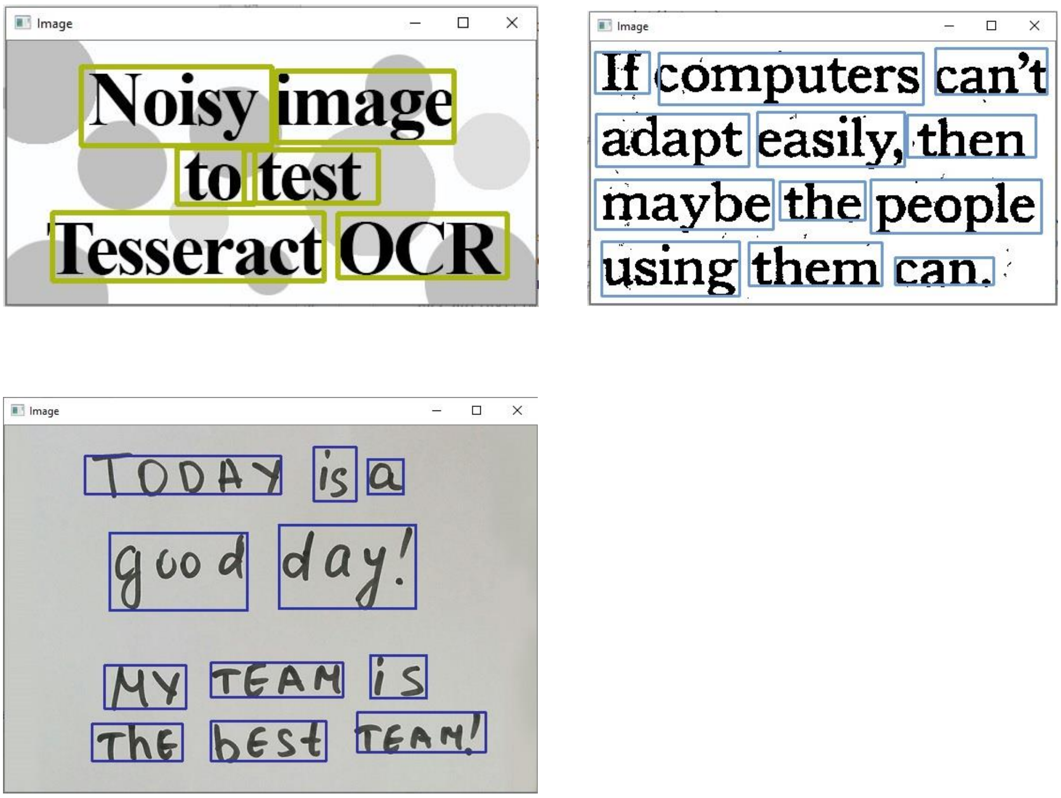
Fig 3: Labelling the words in handwritten document

Once the text files are created for all the text images it is fed into the deep neural network yolov3. Yolov3 is an algorithm which is used to detect the objects and it is developed by Joseph Redmon and Ali Farhadi [11]. This deep learning algorithm detects the score for each object bounding box using logistic regression. The prediction of the object is ignored if it doesn't overlap the ground truth object by more than some threshold. The class prediction performed using binary cross entropy. The prediction of the object is performed at three scales. The extraction of the features is performed using Darknet-53. When the dataset is trained using yolov3 it generates the weights.

The scanned documents contains may be slanted so we have performed de skew operations to make it straight. For de skewing open-cv rotation method is used. Next the input images are read and by using the weights generated by the deep learning network the document with segmented text is generated.

4. Result and Discussion

We have tested our method on different printed and handwritten documents and we have achieved a good recognition result using this method. For testing forty printed and forty handwritten documents are taken and the text segmentation is performed on these documents and we have achieved an accuracy of 96.5% using our methodology.



Noisy image
to test
Tesseract OCR

If computers can't
adapt easily, then
maybe the people
using them can.

TODAY is a
good day!

MY TEAM is
ThE bESt TEAM!

Fig: Recognition of words in printed and handwritten documents.

4. Conclusion

The method uses the deep learning method which identifies the text in the documents using the bounding box methodology. The method takes much time to extract the features and to train the network but it is very fast in recognizing the words in the document. We have achieved good accuracy of 96.5% in word recognition in scanned printed and handwritten documents. We can increase the accuracy of word segmentation by increasing the training data that is number of images in the dataset should be increased.

5. References

1. Keisham, Kanchan, and Sunanda Dixit. "Recognition of handwritten English text U minimisation." Information Systems Design and Intelligent Applications. Springer, New Delhi, 2016. 607-614.

2. Balci, Batuhan, Dan Saadati, and Dan Shiferaw. "Handwritten text recognition using deep learning." CS231n: Convolutional Neural Networks for Visual Recognition, Stanford University, Course Project Report, Spring (2017): 752-759.

3. Pham, Van-Linh, et al. "A Deep Learning Approach for Text Segmentation in Document Analysis." 2020 International Conference on Advanced Computing and Applications (ACOMP). IEEE, 2020.

4. Priyanka, Nallapareddy, Srikanta Pal, and Ranju Mandal. "Line and word segmentation approach for printed documents." IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition 1 (2010): 30-36.

5. Vil'kin, A. M., I. V. Safonov, and M. A. Egorova. "Algorithm for segmentation of documents based on texture features." Pattern recognition and image analysis 23.1 (2013): 153-159.

6. Johnson, Mark, et al. "Modelling function words improves unsupervised word segmentation." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014.

7. Bhatti, Zeeshan, et al. "Word segmentation model for Sindhi text." American Journal of Computing Research Repository 2.1 (2014): 1-7.

8. Shao, Yan, Christian Hardmeier, and Joakim Nivre. "Universal word segmentation: Implementation and interpretation." Transactions of the Association for Computational Linguistics 6 (2018): 421-435.

9. Moysset, Bastien, et al. "Paragraph text segmentation into lines with recurrent neural networks." 2015 13th international conference on document analysis and recognition (ICDAR). IEEE, 2015.

10. Wang, Dongyang, Junli Su, and Hongbin Yu. "Feature extraction and analysis of natural language processing for deep learning English language." IEEE Access 8 (2020): 46335-46345.

11. Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).