Multi-Stage Framework to Boost Optical Character Recognition Performance on Low Quality Document Images

Nitin Gupta*
IBM Research, India

Shashank Mujumdar*
IBM Research, India

Abhinav Jain*
IBM Research, India

Douglas Burdick
IBM Research, Almaden, US

ABSTRACT

In order to extract text from good quality document images, the state-of-the-art (SOA) Tesseract Engine (TE) performs: (i) image processing, (ii) page segmentation to extract text lines and (iii) apply Optical Character Recognition (OCR) on text lines to extract the text tokens. However, TE fails miserably on complex document images with low resolution, colored text regions, tables, charts etc. which presents the need to optimize the TE performance. In this paper, we propose a novel multi-stage pipeline to address the shortcomings of the TE and boost the OCR performance for challenging document images. Specifically, we propose an approach - (i) for page segmentation to extract text lines, (ii) to detect and binarize colored text regions and (iii) to detect and correct the image quality. We rigorously test the pipeline on 5 datasets and show the improvement in the OCR performance against the standard TE and SOA baselines.

KEYWORDS

OCR, Page Segmentation, Tesseract, Low Quality Document Images

ACM Reference Format:
Nitin Gupta, Shashank Mujumdar, Abhinav Jain, and Douglas Burdick. . Multi-Stage Framework to Boost Optical Character Recognition Performance on Low Quality Document Images. In *Proceedings of The Second Document Intelligence Workshop at KDD (Document Intelligence Workshop at KDD).* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

OCR is the technique to extract printed or handwritten text from images and convert it to machine encoded text. For an input document image, TE segments it to extract text lines and applies OCR on the text lines to recognize the text tokens [13]. By default, TE is optimized to recognize sentences of words. The OCR phase of TE operates on a binary image where the text is black against a white background. Thus, for a given high quality document image, where the conditions are favourable for TE to optimally segment the text lines (page segmentation) and generate an accurate binary output, the OCR performance is high. However, factors such as low resolution, illumination changes, blur, noise, skew, character merging/fragmentation, colored regions with poor text contrast against background, broken graphical lines from tables etc. deter the ability of TE to reliably perform page segmentation to identify the text lines and generate the accurate binary representation. It is recommended to generate a high (300-600) dots per inch (dpi) scan of the text document from a flat scanner without illumination artefacts and noise to achieve optimal OCR performance. However, in the absence of adequate hardware, variability in the scanning process and complexity of the document input contents, achieving a high OCR performance on the resultant sub-optimal quality document image scans is a challenging task. Thus, there is a need to reliably extract the text lines and improve the image binarization from such images to improve OCR performance of TE.

Due to multiple image quality issues that need to be addressed for improving OCR performance of low quality documents, any approach catered to address a specific issue is sub-optimal. Beyond improving the image quality, handling colored regions with text and reliably extracting text lines from the image are crucial for optimal OCR performance. We argue that a multi-stage framework that independently improves the performance of TE at every stage is better suited for optimal OCR performance. In this paper, we propose a novel multi-stage pipeline to address the shortcomings of TE and boost the OCR performance specifically on complex low quality document images containing tables, colored regions etc. The main contributions of this paper are - (i) DPI Enhancement: novel un-supervised approach to improve overall document image resolution, (ii) Colored Region Detection: novel un-supervised approach to detect and binarize colored text regions, (iii) Page Segmentation: novel deep learning based page segmentation approach to extract text lines and (iv) Results on five challenging document image datasets.

2 RELATED WORK

The techniques proposed in literature to improve document image quality for better OCR performance can be broadly categorized into two groups - (i) unsupervised image processing techniques and (ii) supervised deep learning techniques. In the former category, due to the intrinsic assumptions of traditional image processing techniques, they are only suitable for specific applications of improving OCR performance [7, 12]. Sophisticated supervised deep learning techniques perform image super resolution (SR) to improve the document image resolution [3, 15, 18]. Although these approaches show improvement in the perceptual quality of the super-resolved document image, the OCR performance is not incorporated in the objective function of the proposed losses. Approach to simultaneously optimize perceptive quality and OCR performance is presented in [9], however, the approach lacks a mechanism to handle colored regions and authors do not discuss improvements in page segmentation. Hence, these approaches do not guarantee improved OCR performance for challenging real

*Authors contributed equally to this work.

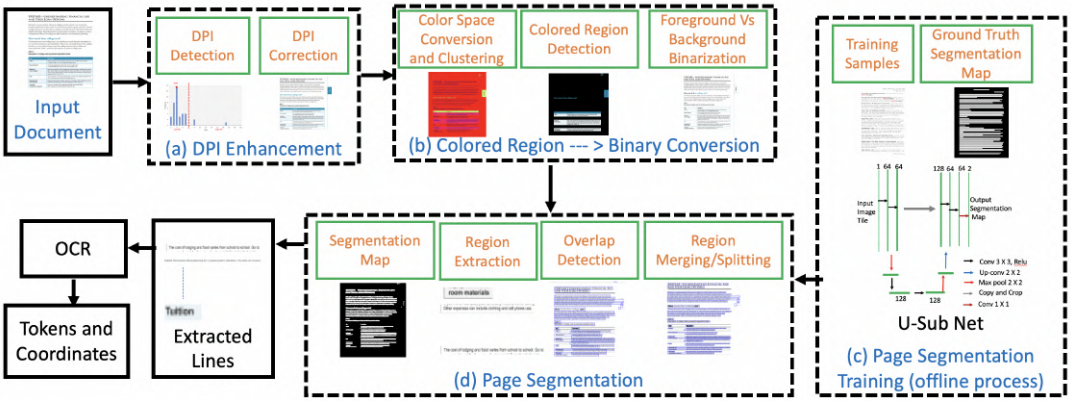

Figure 1: Proposed Multi-Stage Framework For Text Extraction Quality Improvement.

world documents with the presence of colored regions and low resolution. For image binarization, different approaches such as dual thresholding for background/pictures and text [11], local image entropy filtering [8], auto-encoders [1] etc. have been proposed. Approaches for document image segmentation include connected component analysis using linear filtering [6], pixel-based image segmentation [2] etc. Although such methods have been proposed for corresponding tasks, they have not been utilized to improve OCR performance of TE.

It is evident from the documented issues with TE performance that a single approach for image quality improvement will always generate sub-optimal OCR performance [14]. Thus, we argue that improving the results of text line extraction and addressing the image binarization for complex document images would result in improved TE performance. In this paper, we propose such a multi-stage framework to identify and correct quality of low quality documents, image binarization approach that is able to handle colored text region, and a page segmentation approach that accurately extracts the text lines from both low quality and complex document images, for boosting OCR performance of TE.

3 PROPOSED METHODOLOGY

Our multi-stage framework achieves tight bounding boxes around every text line and reliable document binarization with the help of following three stages: (1) DPI Enhancement, (2) Coloured Region Binary Conversion, (3) Page Segmentation. We invoke TE OCR module on the binarized text lines obtained from our pipeline.

3.1 DPI Enhancement

We propose a simple and novel approach for DPI detection and correction based on the observation that high and low DPI document images can be distinguished by estimating the height of the text lines present in them. Let $R = \{r_1, .., r_N\}$ be the set of the detected closed regions from the dilated thresholded image. The subset of closed regions that correspond to text ($R_t = \{\}$) is obtained as follows:

$$R_t = \begin{cases} R_t \cup r_i, & \text{if } height(r_i) > th \ \ where \ \ i \in \{1, .., N\} \\ R_t, & \text{otherwise} \end{cases} \quad (1)$$

where, $height(r_i)$ represents the maximum pixel height of r_i and th is threshold set to 40 pixels.

A histogram of 10 bins with a bin size of 4 is created to capture the distribution of the set R_t against pixel height. Experimentally, we observed that for low DPI images (72-100 DPI), the peak of the histogram lies below bin 5 (i.e. max number of text lines have pixel height < 20 pixels). On the other hand, for high DPI images (250-300) the peak lies in the bin range 5-10 (> 20 pixels). Thus, for a given input document image, we threshold the generated histogram on bin 5 to identify if the document has low/high DPI. For low DPI documents ($\sim$ 72-100 DPI), we re-scale the documents to 300 DPI using bi-cubic interpolation.

3.2 Coloured Region - Binary Conversion

Document image binarization is the pre-processing step required by TE to highlight text from the foreground. The local contrast observed between the text and the colored-region is different than what is observed in the rest of the document with mostly dark-text appearing against a light-background. Hence, we propose an effective binarization scheme (see Fig. 1 (b)) that first detects the coloured regions in a document image using clusters in HSV space and then locally performs otsu-thresholding in selected regions to perform binarization.

3.3 Page Segmentation

We formulate the page segmentation problem as an image to image transformation problem. Proposed framework (see Fig. 1 (c-d))) is motivated by the existing U-Net architecture, originally designed for medical image segmentation [10]. We make use of fewer layers to train our page segmentation model since the text lines are abundantly available in the training set which consists of coloured/grey scale images (from a dataset of 3000 images) rendered at multiple DPIs between a range of [72-300]. The network consists of a contracting path having two blocks of two 3x3 convolutions, each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for down-sampling. The expansive path consists of an up-sampling operation followed by 2x2 convolution, concatenation with the corresponding feature map cropped from the contracting path, and two 3x3 convolutions, each followed by a

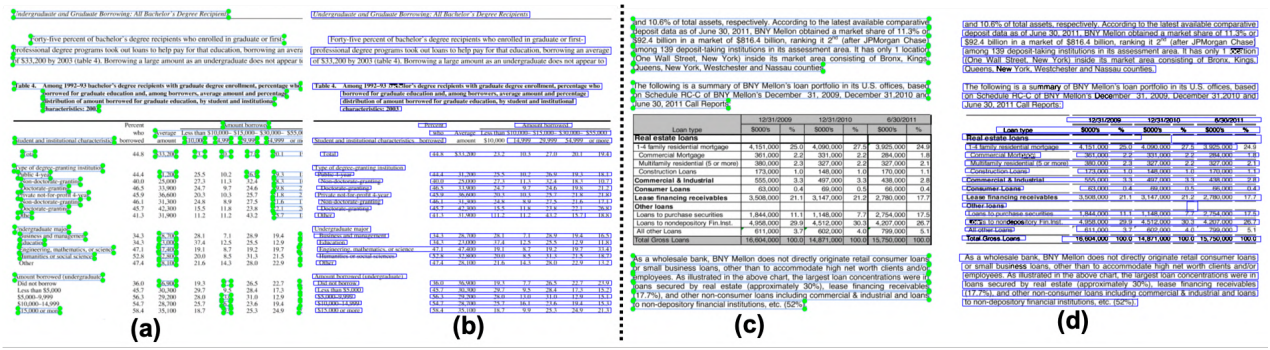

Figure 2: Visual Illustration of Tesseract (a and c) and Proposed (b and d) Page Segmentation Pipeline on 72 DPI Images.

	Tesseract	SRCNN	SAE-18	Model_P	Model_PH	Model_PD	Model_PHD
LCWA	**95.12**	91.95	92.78	94.99	94.41	94.92	94.54
ICDAR	**95.09**	92.27	91.13	94.68	95.08	94.66	94.72
UNLV-A	87.29	79.41	84.65	**88.16**	87.56	87.78	88.10
UNLV-B	94.40	80.77	93.41	**94.54**	94.28	94.13	94.47
CI	93.32	91.27	90.42	93.60	94.11	94.49	**95.10**

Table 1: OCR Accuracy Percentage on 300 DPI Images.

	Tesseract	SRCNN	SAE-18	Model_P	Model_PH	Model_PD	Model_PHD
LCWA	63.17	7.5	9.52	81.16	85.67	87.79	**87.89**
ICDAR	59.07	4.51	5.67	78.16	78.22	84.89	**85.13**
UNLV-A	18.76	6.43	10.28	32.46	32.74	33.70	**33.76**
UNLV-B	35.27	3.8	4.6	42.99	44.0	45.01	**45.16**
CI	57.63	12.92	29.71	73.95	74.71	80.43	**86.60**

Table 2: OCR Accuracy Percentage on 72 DPI Images.

ReLU. In the final layer, a 1x1 convolution is applied before calculating the loss between output layer map and binary image map obtained from the ground truth. Similar to U-Net architecture, the loss layer optimizes two energy functions: (i) pixel-wise soft-max over the final feature map $S_k(x) = exp(a_k(x))/(\frac{1}{n} * \sum_{t=1}^{K} exp(a_t(x)))$ and (ii) cross entropy loss function, $E = \sum_{x \in \lambda} w(x) log(p_{l(x)}(x))$. Here, $a_k(x)$ denotes the activation in feature channel k at the pixel position $x \in \lambda$, K is the number of classes, $p_k(x)$ is the approximated maximum-function, $\lambda \in \{1, .., K\}$ is the true label of each pixel and w is a weight map that assigns some pixels more importance in the training. These weight maps can be estimated using the concept of small separation borders as described in [10].

The output of the sub U-Net is a per-pixel probability estimate, representing the probability that the pixel is part of segmented text lines (see Fig. 1 (d)). We threshold the output probability map with an empirically determined threshold (0.9) to get the binary segmented image. To obtain tight bounding boxes on the text lines from these segmented maps we propose the following post-processing mechanism; (i) *Region Smoothing and Extraction* - extract the closed rectangular bounding regions to get all the segmented lines and perform morphological erosion to divide bigger regions into smaller regions if the pixel height is greater than pre-defined threshold (50 pixels) to help separate merged lines. (ii) *Overlap Detection* - overlap of segmented text lines is observed due to small noisy regions detected near actual text lines as well as due to the presence of skewed text lines within documents. To address this, we consider the single pixel line at the centre of each detected region and move the line in both directions along the height of the segmented region until no pixel belonging to a text character from the binarized image is present on the line. This results in tight bounding boxes for the segmented text regions. (iii) *Region Merging* - in some cases, for e.g., "1. The last issue in Segmentation", the segmented output consists of multiple detections "1.", "The last", "issue in Segmentation". We use simple heuristics such as, identifying if the horizontal pixel gap is smaller than few pixels (30 pixels), to merge such horizontal neighbours.

4 RESULTS AND EVALUATIONS

4.1 Datasets and Baselines

Evaluation is performed on 72 and 300 DPI images rendered from the following document datasets: **(1) UNLV-A**: contains 300 randomly selected PDFs from [17], **(2) UNLV-B** [16]: contains 427 table

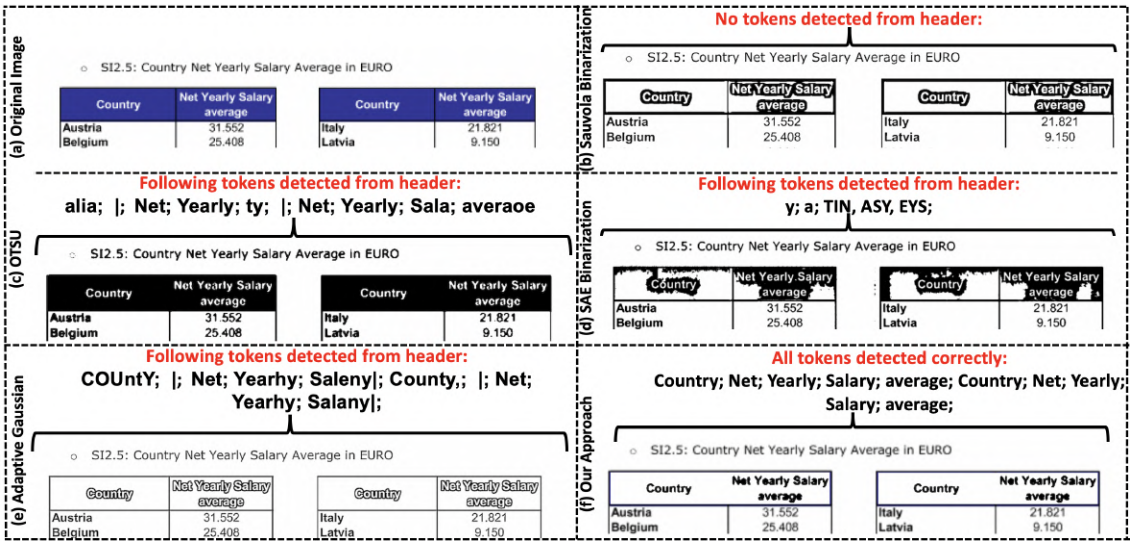

Figure 3: Visual Comparison of different standard thresholding mechanisms and proposed coloured region detection based thresholding mechanism. We have also provided the tokens extracted from the header region by the OCR system for comparison.

containing PDFs from [17], **(3) ICDAR 2013** [4], **(4) LCWA Govt. PDFs** [5]: contains 300 randomly selected PDFs, and **(5) Colored image dataset** (CI): contains 50 manually selected, challenging colored documents from aforemtioned datasets.

We selected following baselines for comparison: **(1) LSTM-based Tesseract** (offers no pre-processing), **(2) SRCNN** [3] which super-resolves low DPI documents, **(3) SAE-18** [1] which uses a selectional auto-encoder network for binarization. Additionally, to highlight improvements offered by each component of our framework, we compare them in configurations such as **(a) Model_P** which uses only page segmentation as pre-processing, **(b) Model_PH** uses both page segmentation and colored region detection modules. Similarly, we created **(c) Model_PD** and **(d) Model_PHD** configuration, where D refers to DPI enhancement module.

4.2 Discussion

An OCR system when deployed on a text image, returns the coordinates of the bounding boxes enclosing the tokens and strings of characters present in every token. For each of the baselines, we compute the OCR accuracy by matching the text tokens extracted from processed image with those in the ground truth at the corresponding location as returned by TE. As seen from Table 2, for 72 DPI datasets, our approach outperforms all other baselines by approximately 15-30%. TE segmentation pipeline assumes font size to be high for optimal performance while our Model_P is trained on both low and high DPI images and hence shows significant improvement in OCR accuracy on 72 DPI datasets where font size is considerably smaller. The poor performance of SRCNN is attributed to the fact that the SRCNN network learns to smoothen the text content by virtue of the MSE loss function which results in characters and tokens merging for low resolution document images (due to a gap of only few pixels). Similarly, SAE Binarization also assumes high DPI document images to operate on and fails miserably on low DPI images. Proposed approach and the baselines perform at par with each other on 300 DPI images, as seen from Table 1.

Qualitatively (see Fig. 2(a) and (c)), we can observe that Tessearct's built-in page segmentation misses several tokens (from table) in the output, unlike our approach as seen from Fig. 2(b) and (d). For evaluating colored-region conversion, Fig. 3(a) shows part of the original colored image, Fig. 3(b)-(e) represent baseline outputs and Fig. 3(f) represents output of our pipeline. Baselines like OTSU, Adaptive, Sauvola and recently proposed deep learning based binarization give optimal performance only on images with dark text against a light colored background and fail on colored document images as seen from Fig. 3(d) which further affects TE token extraction performance. From Table 1 and 2, specifically for CI dataset with colored images only, our colored region processing pipeline is able to boost performance of (a) proposed page segmentation pipeline (from Model_P to Model_PH), and also (b) the combination of proposed page segmentation and DPI Enhancement pipeline (from Model_PD to Model_PHD). With these experimental results, we can conclude that our proposed multi-stage pipeline is performing significantly better than the baselines on low resolution images, and having similar or marginal improvement in performance for high resolution images.

5 CONCLUSION AND FUTURE WORK

In this paper, we have proposed a multi-stage framework that improves the OCR performance of Tesseract on low-quality complex document images through proposed modules. Rigorous experiments against standard baselines on 5 challenging datasets showcase the OCR performance boost offered by our framework specifically on 72 DPI document images. In future, we will explore post-OCR correction techniques utilising language-based models for correcting erroneous tokens.

REFERENCES

[1] Jorge Calvo-Zaragoza and Antonio-Javier Gallego. 2019. A selectional auto-encoder approach for document image binarization. *Pattern Recognition* 86 (2019), 37–47.

[2] Kai Chen, Mathias Seuret, Jean Hennebert, and Rolf Ingold. 2017. Convolutional neural networks for page segmentation of historical document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE.

[3] Chao Dong, Ximei Zhu, Yubin Deng, Chen Change Loy, and Yu Qiao. 2015. Boosting optical character recognition: A super-resolution approach. *arXiv* (2015).

[4] ICDAR. [n.d.]. 2013 ICDAR Table Detection Dataset. http://www.tamirhassan.com/html/competition.html.

[5] LCWA. [n.d.]. LCWA Govt. Dataset. https://s3.us-east-2.amazonaws.com/lclabspublicdata/lcwa_gov_pdf_README.txt.

[6] Alesya Ishchenko Marina Polyakova and Natallia Huliaieva. 2018. Document image segmentation using averaging filtering and mathematical morphology. In *2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)*. IEEE.

[7] Matija Pul Matteo Brisinello, Ratko Grbić and Tihomir Anđelić. 2017. Improving optical character recognition performance for low quality images. In *2017 International Symposium ELMAR*. IEEE.

[8] Hubert Michalak and Krzysztof Okarma. 2019. Improvement of Image Binarization Methods Using Image Preprocessing with Local Entropy Filtering for Alphanumerical Character Recognition Purposes. *Entropy* 21, 6 (2019), 562.

[9] Shashank Mujumdar, Nitin Gupta, Abhinav Jain, and Douglas Burdick. 2019. Simultaneous optimisation of image quality improvement and text content extraction from scanned documents. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1169–1174.

[10] Olaf Ronneberger and Fischer. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*.

[11] Jaakko Sauvola and Matti Pietikäinen. 2000. Adaptive document image binarization. *Pattern recognition* 33, 2 (2000), 225–236.

[12] Mande Shen and Hansheng Lei. 2015. Improving OCR performance with background image elimination. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE.

[13] Tesseract. [n.d.]. Tesseract OCR. https://github.com/tesseract-ocr/tesseract.

[14] TesseractQuality. [n.d.]. Tesseract - Improve Quality of Input Image. https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality.

[15] Hanh Tran and Tien Ho-Phuoc. 2018. Deep Laplacian Pyramid Network for Text Images Super-Resolution. *arXiv* (2018).

[16] UNLV. [n.d.]. UNLV Dataset. https://code.google.com/archive/p/isri-ocr-evaluation-tools/downloads.

[17] UNLV-Tables. [n.d.]. UNLV Table Dataset. http://www.iapr-tc11.org/mediawiki/index.php/Table_Ground_Truth_for_the_UW3_and_UNLV_datasets.

[18] Haochen Zhang, Dong Liu, and Zhiwei Xiong. 2017. CNN-based text image super-resolution tailored for OCR. In *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE.