Extracting Text from Degraded Document Image

Radhika Patel and Suman K. Mitra
Dhirubhai Ambani Institute of
Information and Communication Technology
Gandhinagar, India 382007
Email: radhika.patel4391@gmail.com, suman_mitra@daiict.ac.in

***Abstract*—The recent era of digitization is expected to digitized many old important documents which are degraded due to various reasons. Degraded document image binarization has many challenges like intensity variation, background contrast variation, bleed through, text size variation and so on. Many approaches are available for document image binarization, but none can handle all types of degradation at once. We proposed an approach which consists of three stages such as preprocessing, Text-Area detection and post-processing. Preprocessing enhances the contrast of the image. Next stage involves identifying Text-Area. Postprocessing technique takes care of false positives and false negative based on intensity values of preprocessed and gray image. The Performance is evaluated based on various quantitative measures and is compared with the method regarded best so far. The algorithm is also expected to be independent of the script, hence is tested on Gujarati degraded document images.**

Keywords: Binarization, edge detection, Text extraction

I. Introduction

Old historical documents now-a-days are digitized to preserve the information. Digitized document images can be used for text retrieval, image analysis or optical character recognition. With the boom of hand-held devices like kindle and i-pads, people prefer to have all text data in digital form. Image binarization is also an essential stage in automatic text extraction from documents. Binarization of digitized documents leads to less storage requirement. Image binarization can be done using various approaches. Segmentation and thresholding are intensity-based methods; morphological methods uses the shape of the object to binarize images along with other segmentation methods. Segmentation and thresholding can be explained as a partition of an input image into two classes based on pixel values in an input image. Morphological operators operate based on neighborhood set information to decide foreground from background. Document binarization is employed wherever text extraction/retrieval is necessary e.g. applications like information retrieval from a text document, language translation of the text, digitizing old books and so on. With these aims in mind a worldwide competition is organized known as DIBCO [1]. The present work is an attempt to binarize the data sets released by DIBCO. It has three main components vize pre-processing where contrast of background and foreground (mainly text) is carried out. Identify text area from enhance image is presented next. Finally post-processing is carried out to handle false positive and true negative. The rest of paper organized in following way Section 2 covers proposed approach, Section 3 discovers result of proposed approach and compared with state of art method, and Section 4 concludes the proposal.

978-1-4673-8564-0/15/$31.00 © 2015 IEEE

II. Proposed Approach

While binarizing a document, it is important to preserve maximum possible text. We observed that degraded images have highly overlapped text and background intensity range, thus hard thresholding can not separate the text from the background. Proposed approach concentrates on the philosophy of enhancing edges between text and background. However, due to degradations, the edges do not have very high intensity differences. The aim is to increase the local contrast in an image, in turn enhancing text as a preprocessing step. The proposed approach detects possible text area in preprocessed image and further decides to classify the pixel as either foreground or background. To identify possible text area, edge detection based on rough set [2] is used. Postprocessing technique is developed to remove the undesired noise to improve overall binarization performance. Now three main components of the proposal are discussed one by one.

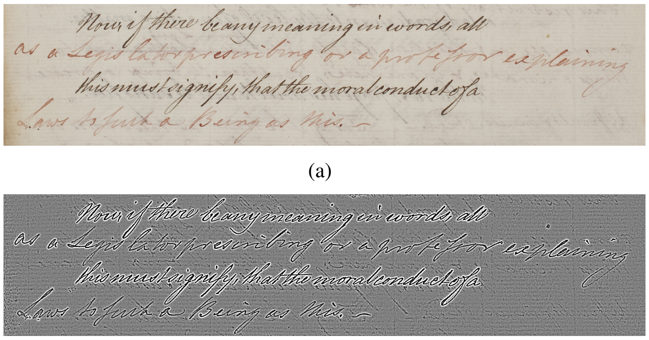
A. Preprocessing

The first step of preprocessing is converting the RGB (colour) image to grayscale. A PCA based conversion [3] is used for this purpose. Next is a chain of basic image processing techniques to improve local contrast and suppressing the noise from background texture in order to efficiently detect text region. This step is motivated by [4] where illumination variations is addressed.

Gamma correction: Gamma correction replaces gray-level I with I^{γ}(for $\gamma > 0$) or $\log I$ (for $\gamma = 0$) where $\gamma \epsilon [0, 1]$ is a user defined a parameter. It enhances the local dynamic range of the image in dark while compressing the range of bright regions. For given text documents, experiments were carried out with different γ values. The value $\gamma = 0.2$ is set considering the end result.

Difference of Gaussian (DoG): Gamma correction does not remove the influence of overall intensity variations. Document images have slow/smooth background variations contributing to low-frequency component. Text edges contributing middle range frequency components and noise contributing high-frequency component. DoG filtering [4] is a convenient way to achieve the bandpass behavior. DoG is basically a difference of $2D$ Gaussian filter having different variances. Experimenting on different images, values of variances are set to 1 and 4.

Equalization: Processed image still typically contains extreme values produced by highlights, small dark regions, garbage at the image borders. Two-stage approximation as describe in [4] is used to rescale the gray values present in processed image. Preprocessing result with input image is shown in Fig 1.



(a)

(b)

Fig. 1. (a) Input RGB image and (b) preprocessed image

B. Text Area Detection

Preprocessed image contains locally high contrast that separates text region from background with large intensity difference as shown in Fig 1(b). At the same time, it suppresses local texture of background and removes several degradations. Preprocessed image is now more suitable for edge detection.

Edge Detection Using Rough-Set Theory: In most of the preprocessed documents, histogram shows one peak (due to mid-range gray values belonging to background). Intensities lesser than the peak values generally belong to text, and intensities more than the peak value generally belong to background. But in practice the intensities lying nearer to peak intensity may contain pixels from text as well as background. Thus, applying hard thresholding based edge detection method will not work. To exploit local high contrast around text in preprocessed image rough set based edge detection as presented in [2] is used. Roughset method, approximate thresholds are obtained from the histogram by sliding window based approach. It is adapted to find local minimum in the histogram and use multiple minimum as thresholds. Here, the window size is fixed to 5 empirically. Image is binarized separately with each threshold and combined to get desired result.

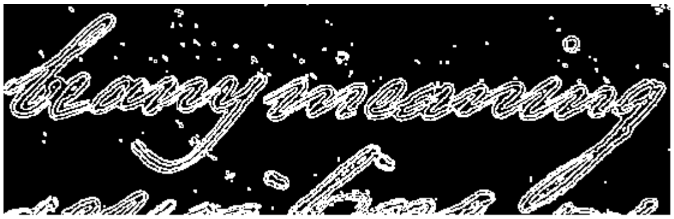


Fig. 2. Zoomed in edge detection result by Rough Set on preprocessed image

Region Filling: Possible text area is surrounded by edges as shown in Fig 2. Note that corrected edges are obtained by the Rough Set based method. Region Filling is used to fill these text area. Some very small group of pixels are wrongly detected as foreground. Even, in some of the images noise present in background (bleed through or very high textural background) are wrongly detected as text area (false positive). Some of these small connected components are removed by deciding the size of small connected components. To retain as much as possible text in the output image, it is observed in DIBCO dataset [1] that the components having the area less than 50 pixels are noise only.

Text detection: It is observed that thresholding of preprocessed image with some high value will separate background region surrounding the text. From foreground pixels (pixels those are identified as text in previous stage) whose intensity value in preprocessed image is higher than 173 is labeled the pixels as background in resulting image. The threshold value is set to 173 experimentally.

C. Postprocessing

Previous two stages generate documents with text. A few false positives (non text identified as text) and false negatives (text portion not detected) region are also generated. The false positive case is regarding as black blob and false negative case is regarded as white blobs. Fig 3 is showing example cases of black blob and white blob. Again black blobs are two types. One is large blob arising due to region filing and another is mostly one pixel width blob around the border of text, regarded as border noise. All three cases are addressed in post processing.

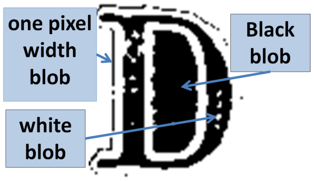
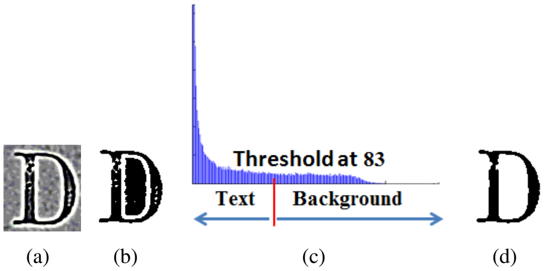


Fig. 3. Result of text detection process with black and white blob

Border Removal: Majority of the pixels which are neighbors of border noise are background. Hence 8 connected neighbors of all text pixels (this obviously include border noise) are considered and identified as background (non text) if 60% of its neighbors are background. This successfully removed border noise.

Black Blob Removal: Some letters such as 'D','R','P','e','o' have closed regions which get filled during region filling and hence remain as black till the text detection. These accumulated as black blobs which are undesired. To remove such blobs, help of image which is preprocessed (output of II-A) has been taken. Note that preprocessed image is gray scale image having pixel values in the range [0,255]. Text pixels are first identified from the so far binarized image (these obviously include black blobs if present). Positions of such pixels are identified within the preprocessed image. Now pixels values are plotted. It is expected that actual text will have values near 0 and pixel values of black blob will be far from 0. So the plotted histogram will be positively skewed. Now from this histogram a threshold value equal to $\mu + \sigma$ is set, where μ is mean and σ is the standard deviation of the histogram. It was observed that the black connected components (in so far binarized image) with value greater than threshold ($\mu + \sigma$) were black blobs. Hence those black blobs in the so far binarized image are changed to background. Thus most of the balck blobs are removed. Fig 4 shows an example of the above mentioned process.



(a) (b) (c) (d)

Fig. 4. Black blob removal (a) Preprocessed image (b) So far Binarized image (c) Plot of text intensity in preprocessed image (d) Output image

White Blob Removal: There are some degraded images with some of the text areas having very high intensities, however it was expected that those areas should have low intensity. Such intensities get enhanced in preprocessing, and thus classified as background during binarization. These accumulated in white blobs within the text area. To remove the white blobs following steps are performed

- Binarizing the gray scale (Output of RGB to gray) image such that no text region is lost (False Foreground can be tolerated, but False Background is not allowed). Observing all the available images from database 128 intensity is set experimentally such that no text area is lost. This image is now considered as reference image for the white blob removal.
- Find all white connected component from the output of previous step (Black Blob Removal), except the largest connected component which is clearly a background in the document image.
- If more then 30% pixels of any white connected component in the reference image are classified as background then these remain background in output image else it is set to foreground.

Image is now almost correctly segmented into text and background with some small components present due to noise in the input image. Small object removal is again performed as explained in section II-B. The size of the object is decided empirically such that it does not remove text pixels (true positive) such as punctuation symbols.

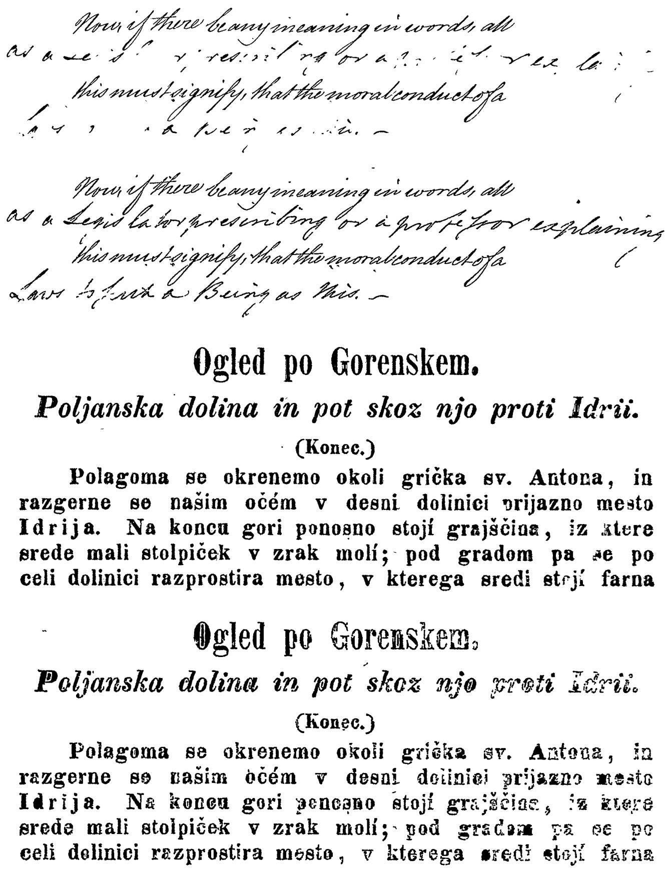
III. Results and discussion

The proposed approach is tested extensively on DIBCO dataset [1] published for the year 2009 to 2014. One of the best results and worsed results obtained are shown in Fig 5. For the purpose of comparison, corresponding results by the best approaches (as publised by DIBCO) are also shown. The Performance evaluation is based on various quantitative measures like F-Measure (FM), pseudo F-measure (psFM), Peak Signal-to-Noise Ratio (PSNR), Distance Reciprocal Distortion (DRD), pseudo-Recall (Rps) and pseudo-precision (Pps) as mention in [5]. Note that some of the measures defined at later point of time and hence not tested for data base of earlier days. F-measure and PSNR are commonly used measures and hence need not to explain. DRD measures visual distortion of output image with ground truth image. Pseudo F-measure uses pseudo recall and pseudo-precision instead of recall and precision in F-measure. Pseudo recall and pseudo-precision use weight matrix according to the contour of ground truth. In pseudo-Recall (Rps), the weights of the ground truth foreground are normalized according to the local stroke width and in pseudo-Precision the weights are constrained within an area that expands to the ground truth background, taking into account the stroke width of the nearest ground truth component as described in [6]. Performance is measured using DIBCO evaluation tool which is available in [1]. Quantitative results of proposed algorithm on DIBCO datasets are compared with year wise winner of DIBCO as shown in table I. DIBCO 2009, DIBCO 2010, DIBCO 2011, DIBCO 2012, DIBCO 2013, and DIBCO 2014, dataset has 10 (5 printed, 5 handwritten), 9 (Handwritten), 16 (8 printed, 8 handwritten), 14 (Handwritten), 16 (8 printed, 8 handwritten) and 10 (Handwritten) document images respectively. Rps measure of DIBCO 2013 dataset of current algorithm is high compare to the winner of DIBCO2013 (Bolan Su et al) [5]. Note that Rps measures efficiency of foreground (text) retrieval which is the main target. Proposed algorithm does not depend on text size and stroke width so it can be used on image containing text of almost all scripts. To test the suitability of proposed approach, dataset of Gujarati degraded document image is created. The algorithm performs equally well on Gujarati document images, that justifies the claim. Results on Gujarati data set are shown in Fig 6. Ground truth of data set is not generated so performance evaluation is carried out qualitatively by visual appearance. Proposed approach can efficiently binarize English numerical and Gujarati numerical which are included in same document.

TABLE I
Evaluation Comparison with State Of The Art Method. [Here, BRP stands for Best Result Published by DIBCO and PM stands for Proposed Method]

DIBCO 2009 DataSet			DIBCO 2010 DataSet		
Measure	BRP[7]	PM	Measure	BRP[8]	PM
FM(%)	91.24	86.28	FM(%)	91.50	87.39
PSNR	18.66	16.49	PSNR	19.78	17.77
DIBCO 2011 DataSet			DIBCO 2012 DataSet		
Measure	BRP[9]	PM	Measure	BRP[10]	PM
FM(%)	91.36	84.64	FM(%)	89.47	84.70
PSNR	16.13	16.39	PSNR	21.80	17.65
DRD	108.48	5.41	DRD	3.440	5.49
DIBCO 2013 DataSet			DIBCO 2014 DataSet		
Measure	BRP[5]	PM	Measure	BRP[11]	PM
FM(%)	92.12	88.22	FM(%)	96.88	92.29
psFM(%)	94.19	92.27	psFM(%)	97.65	94.29
PSNR	20.68	18.20	PSNR	22.66	18.54
DRD	3.10	4.60	DRD	0.902	2.66
Rps	94.15	95.94	Rps		98.09

Discussion: In some images proposed algorithm works better than the state of the art method (Winner of DIBCO 2013) [5], [12] and in some images the state of the art method works better. The advantage of proposed algorithm is that it gives maximal text retrieval. Statistically overall results is close to state of the art method results, but text detection in all type of degraded images using proposed algorithm is good in the sense that almost no text region is classified as background (false background). As shown in Fig 5, state of the art method losses many texts (Foreground classified as background resulting into false background) compare to the result of proposed algorithm. The only disadvantage of the method is the presence of undesired blobs (black and white). The challenge of removing such blobs is attempted, however more sophisticated method may improve the result.



Now if there be any meaning in words, all

[illegible]

this must signify, that the moral conduct of a

[illegible]

Now if there be any meaning in words, all

as a Legislator prescribing or a professor explaining

this must signify, that the moral conduct of a

Laws to such a Being as this. —

Ogled po Gorenskem.

Poljanska dolina in pot skoz njo proti Idrii.

(Konec.)

Polagoma se okrenemo okoli grička sv. Antona, in razgerne se našim očém v desni dolinici prijazno mesto Idrija. Na koncu gori ponosno stojí grajščina, iz ktere srede mali stolpiček v zrak molí; pod gradom pa se po celi dolinici razprostira mesto, v kterega sredi stojí farna

Ogled po Gorenskem.

Poljanska dolina in pot skoz njo proti Idrii.

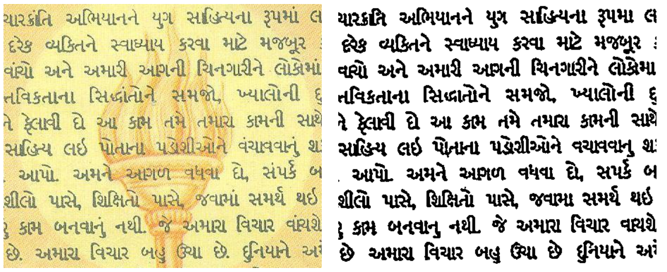
(Konec.)

Polagoma se okrenemo okoli grička sv. Antona, in razgerne se našim očém v desni dolinici prijazno mesto Idrija. Na koncu gori ponosno stojí grajščina, iz ktere srede mali stolpiček v zrak molí; pod gradom pa se po celi dolinici razprostira mesto, v kterega sredi stojí farna

Fig. 5. 1^{st} and 3^{rd} are best result published by DIBCO-13, whereas 2^{nd} and 4^{th} respectively are best and worst results by proposed approach

IV. Conclusion

This paper presented a method that is collection of image processing techniques for text extraction from degraded document. The binarization using rough set approach on preprocessed image provides the closed edges with almost zero loss of the text area which is the main advantage of the current algorithm. The proposed method gives satisfactory result in most of the degraded images in the sense of text preservation. Many existing techniques fail in case where text intensity is in large range or different types of ink used for text or text intensity is near to background. It is not claimed that proposed algorithm is the panacea of all problems, but able to work satisfactorily for all types of images for retrieving text near perfectly. The main disadvantage of the proposal is that it is highly dependent on parameters which are in turn dependent on the set of images used.



ચારક્રાંતિ અભિયાનને યુગ સાહિત્યના રૂપમાં લ
દરેક વ્યક્તિને સ્વાધ્યાય કરવા માટે મજબૂર :
વર્ષો અને અમારી આગની ચિનગારીને લોકોમાં
નવિકતાના સિદ્ધાંતોને સમજો, ખ્યાલોની દુ
ને ફેલાવી દો આ કામ તમે તમારા કામની સાથે
સાહિત્ય લઈ પોતાના પડોશીઓને વંચાવવાનું શ
. આપો. અમને આગળ વધવા દો, સંપર્ક બ
શીલો પાસે, શિક્ષિતો પાસે, જવામાં સમર્થ થઈ
કામ બનવાનું નથી. જે અમારા વિચાર વાંચશે
છે અમારા વિચાર બહુ ઉંચા છે દુનિયાને અં

Fig. 6. Result of Gujarati dataset images using proposed approach, Input image and corresponding output image

References

[1] Dr. Basilis G. Gatos. Konstantinos Ntirogiannis Dr. Ioannis Pratikakis. DIBCO. http://users.iit.demokritos.gr/~kntir/#Competitions, 2013. [Online; accessed 20-March-2015].

[2] Ashish Phophalia, Suman K Mitra, and Ajit Rajwade. A new denoising filter for brain mr images. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, page 57. ACM, 2012.

[3] Jon Parker, Ophir Frieder, and Gideon Frieder. Robust binarization of degraded document images using heuristics. In *IS&T/SPIE Electronic Imaging*, pages 90210U–90210U. International Society for Optics and Photonics, 2013.

[4] Xiaoyang Tan and Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on*, 19(6):1635–1650, 2010.

[5] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. Icdar 2013 document image binarization contest (dibco 2013). In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1471–1476. IEEE, 2013.

[6] Konstantinos Ntirogiannis, Basilios Gatos, and Ioannis Pratikakis. Performance evaluation methodology for historical document image binarization. *Image Processing, IEEE Transactions on*, 22(2):595–609, 2013.

[7] Basilis Gatos, Konstantinos Ntirogiannis, and Ioannis Pratikakis. Icdar 2009 document image binarization contest (dibco 2009). In *2009 10th International Conference on Document Analysis and Recognition*, pages 1375–1382. IEEE, 2009.

[8] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. H-dibco 2010-handwritten document image binarization competition. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 727–732. IEEE, 2010.

[9] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. Icdar 2011 document image binarization contest (dibco 2011). In *2011 International Conference on Document Analysis and Recognition*, pages 1506–1510. IEEE, 2011.

[10] Ioannis Pratikakis, Basilis Gatos, and Konstantinos Ntirogiannis. Icfhr 2012 competition on handwritten document image binarization (h-dibco 2012). In *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, pages 817–822. IEEE, 2012.

[11] Konstantinos Ntirogiannis, Basilis Gatos, and Ioannis Pratikakis. Icfhr2014 competition on handwritten document image binarization (h-dibco 2014). In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 809–813. IEEE, 2014.

[12] Bolan Su, Shijian Lu, and Chew Lim Tan. Robust document image binarization technique for degraded document images. *Image Processing, IEEE Transactions on*, 22(4):1408–1417, 2013.