TEXTLINE INFORMATION EXTRACTION FROM GRAYSCALE CAMERA-CAPTURED DOCUMENT IMAGES

Syed Saqib Bukhari and Thomas M. Breuel

Technical University of Kaiserslautern,
Germany
bukhari@informatik.uni-kl.de,
tmb@informatik.uni-kl.de

Faisal Shafait

German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany
faisal@iupr.dfki.de

ABSTRACT

Cameras offer flexible document imaging, but with uneven shading and non-planar page shape. Therefore camera-captured documents need to go through dewarping before being processed by traditional text recognition methods. Curled textline detection is an important step of dewarping. Previous approaches of curled textline detection use binarization as a pre-processing step, which can negatively affect the detection results under uneven shading. Furthermore, these approaches are sensitive to high degrees of curl and estimate x-line[1] and baseline pairs using regression which may result in inaccurate estimation. We introduce a novel curled textline detection approach for grayscale document images. First, the textline structure is enhanced by using match filter bank smoothing and then central lines of textlines are detected using ridges. Then, x-line and baseline pairs are estimated by adapting active contours (snakes) over ridges. Unlike other approaches, our approach does not use binarization and applies directly on grayscale images. We achieved 91% of detection accuracy with good estimation of x-line and baseline pairs on the dataset of CBDAR 2007 document image dewarping contest.

Index Terms— Curled Textline Detection, Grayscale Camera-Captured Document Image Segmentation.

1. INTRODUCTION

Digital cameras are low priced, portable, long-ranged and non-contact imaging devices as compared to scanners. These features make cameras suitable for versatile OCR related applications like, mobile OCR, digitizing thick books, digitizing fragile historical documents, etc. But fundamental obstacles like uneven light shading, low resolution, motion blur, under- or over-exposure, non-planar page shape and perspective distortions bring up new problems to the traditional OCR system. Therefore some pre-processing steps like "binarization" and "dewarping" are performed before text recognition. Curled textline detection and x-line and baseline pairs estimation are important steps for dewarping. Previous approaches of curled textline detection [1, 2, 3, 4, 5, 6, 7, 8] work on binarized images. These approaches can be divided into two categories: (a) heuristic search [1, 2, 3, 4, 5, 6] and (b) active contours (snakes) [7, 8].

Heuristic search based approaches start from a single component and search other components in a growing neighborhood region. Most of these approaches use rule-based criteria for textline searching. Our active contours (snakes) based baby-snakes [7] model introduced the use of many small open curved snakes and snakelets [8] model introduced many small growing coupled snakes pairs for curled textlines detection from binarized images.

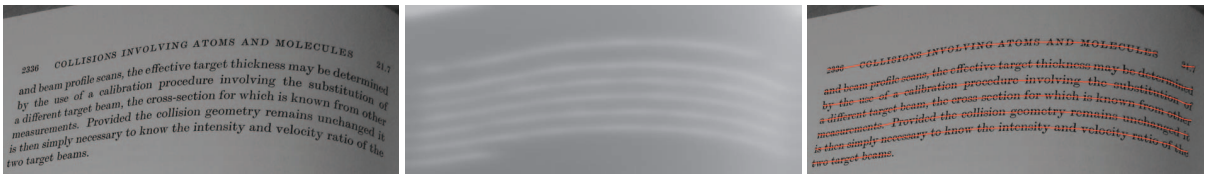
Heuristic search and active contours (snakes) based approaches depend upon binarization [9] of camera-captured document image before textline detection. In the presence of the challenging conditions like uneven shading, low resolution, motion blur and under- or over-exposure, binarization may give bad results. Therefore, binarization can negatively affect the textline detection results.

Moreover, most of these approaches perform x-line and baseline pairs estimation using regression over top and bottom points of binarized connected components of detected textlines, which may result in inaccurate estimation.

We have introduced a method for textline detection directly from grayscale document images in [10]. Our work in this paper is the extension of our previous work [10] with the additional estimation of x-line and baseline pairs. Here, our method starts by enhancing the grayscale curled textline structure using multi-oriented multi-scale anisotropic Gaussian smoothing based on matched filter bank approach [11]. Then ridges [12, 13] are detected from the smoothed image, where ridges defines the unbroken central lines structure which pass through the centers of the textlines. Then we model active contours (snakes) [14] over ridges for estimating the pairs of x-line and baseline.

We make the following contributions in this paper. The method presented here works directly on grayscale intensities of camera-captured document images and therefore inde-

[1] Line following the top of x-height of characters.



(a) Grayscale document image portion. (b) Enhanced curled textlines structure. (c) Ridges detection using smooth image, which intern results in textline detection

Fig. 1. Different steps of curled textline information extraction algorithm

pendent of binarization and its problems. Unlike regression over connected components points, adaptation of active contours (snakes) over grayscale document images gives precise and accurate x-line and baseline pairs information. Besides, our approach also works well in the presence of marginal noise [15], and does not require prior marginal noise removal.

The rest of the paper is organized as follows: Section 2 describes the technical and implementation details of curled textline information extraction algorithm. Section 3 comprises the performance evaluation and experimental results. Section 4 discusses the results and conclusion.

2. CURLED TEXTLINE INFORMATION EXTRACTION

Our curled textline information extraction algorithm comprises three steps: (1) grayscale textline enhancement using multi-oriented multi-scale anisotropic Gaussian smoothing, (2) detection of central lines of curled textlines using ridges, and (3) adaptation of active contours (snakes) over ridges for estimating x-line and baseline pairs. Steps 1 and 2 are based on the pre-processing steps of our reported approaches [10, 16]. All these steps are described below.

2.1. Textline Enhancement: Matched Filter Bank

In handwritten textline segmentation domain, Li et al. [17] introduced the concept of straight textline enhancement of binary image using anisotropic Gaussian smoothing. This textline enhancement approach can not be applied on camera-captured document images which contain non-straight textlines with high degrees of multi-oriented curl and variable font sizes. Matched filter bank approach has been used for enhancing the structure of multi-oriented blood vessels [11] and finger prints [18]. Here, multi-oriented multi-scale anisotropic Gaussian smoothing based on matched filter bank approach, is used for enhancing curled textlines. Here a set of oriented anisotropic Gaussian filters is generated by using different values of σ_x, σ_y, and θ from their ranges. For σ_x and σ_y single range has been used which is a function of document image height. For θ, -45 to 45 degrees range has been used. The set of filters is applied to each pixel of grayscale image and the maximum value among them is selected for a particular pixel in the resulting smoothed image. Figures 1(a) and 1(b) show the input and smoothed images respectively. Multi-oriented multi-scale anisotropic Gaussian smoothing enhances the curled textlines structure well, which is shown in Figure 1(b).

2.2. Textlines Detection: Ridges

Ridges detection has been used for describing the significant features of grayscale images [12] and speech-energy representation in time-frequency domain [13]. Here we are using ridges for detecting textlines regions. We have already seen that multi-oriented multi-scale anisotropic Gaussian smoothing enhances curled textlines structure well from grayscale image. Therefore, detection of ridges from smoothed/enhanced textlines image can produce central line features of textlines. Here, Horn-Riley [12, 13] based ridges detection approach is used. This approach is based on differential geometry, which uses local direction of gradients and second derivatives as the measure of curvature. Hessian matrix is used for finding direction of gradients and derivatives. By using this information, ridges are detected by finding the zero-crossing of the appropriate directional derivatives of smoothed image. Detected ridges over the smoothed image of Figure 1(b) are shown in Figure 1(c). These ridges cover the complete central line feature of textlines and result in textlines detection.

2.3. X-Line-Baseline Pairs Estimation: Snakes

Curled textlines have been already detected using previous steps. This section introduces the modified snakes model for estimating the information of x-line and baseline pairs from detected textlines. Our baby-snake [7] and snakelet [8] models are also based on snakes for curled textline detection, but from binarized document image. The modified snake model presented here is as an extension of [7, 8]. The features of modified snakes model for estimating x-line and baseline pairs are described below:

1. **External forces calculation from gradient of grayscale image:** The gradient of grayscale document image is computed by using Sobel filter. Positive magnitudes in the gradient image are dominated by the top parts

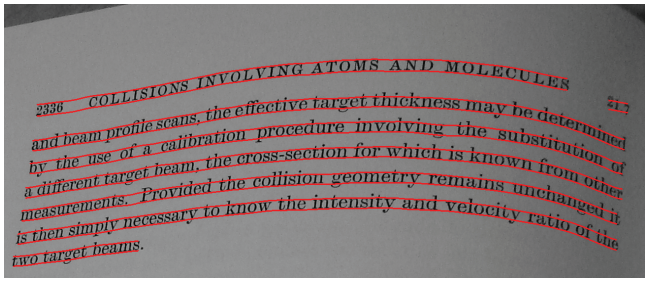


Fig. 2. Result of extracting x-line and baseline in Figure 1.

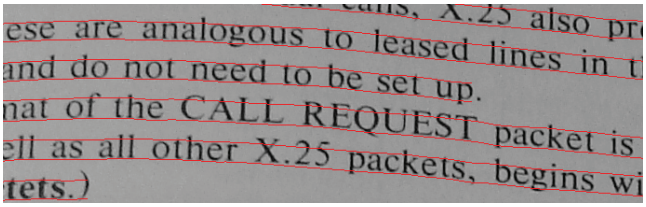
of curled textlines and similarly negative magnitudes of gradient image are dominated by the bottom parts of curled textlines. Here positive and absolute negative gradient images are referred as top and bottom gradient images respectively. The gradient vector flow (GVF) [19] forces are calculated by using these images.

2. **Weighted-coupled snakes pair for grayscale image:** In [8] we have used coupled snakes model (snakelets) for textline detection from binarized images. Here modified coupled snakes model is presented for estimating x-line and baseline pairs from detected textlines, referred to as weighted-coupled snakes model. Let's suppose, duplicated pairs of open-curve snakes pass through the centers of curled textlines. For each pair, one snake is deformed with respect to the vertical components of GVF of top gradient image and another one with respect to the vertical components of GVF of bottom gradient image. Large percentage of GVF of bottom gradient image and small percentage of GVF of top gradient image are used during coupling, because of the assumption that more characters lie on baseline than on x-line. After each deformation iteration, the distances between each pair of snakes are adjusted and made equal to average distance.

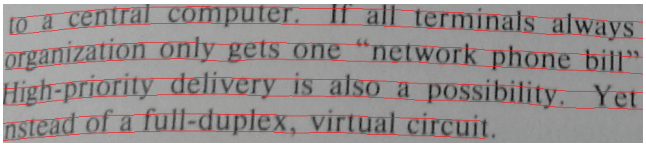
Based on the above defined features of snakes model, x-line and baseline pairs estimation is performed as follow. The top and bottom gradient images are generated from an input grayscale image. Then GVFs of both images are calculated. The duplicated ridges are used as initial open-curve snakes pairs for curled textlines. Each pair is deformed using GVFs images in the weighted-coupled snakes fashion described above. Figure 2 shows the estimated pairs of x-line and baseline for curled textlines.

3. EXPERIMENTS AND PERFORMANCE EVALUATION

We evaluate the presented approach on the real-world handheld camera-captured document images dataset used in CBDAR 2007 for document image dewarping contest [20]. This



(a) Presence of capital letters and numbers



(b) Presence of blur and quotation marks

Fig. 3. Sample results of x-line and baseline estimation.

Table 1. Performance evaluation results based on [21, 23] on CBDAR 2007 dewarping contest dataset. For more details, please refer to [10].

Correct segmentation accuracy [23]	90.50%
Correct segmentation accuracy [21]	91.05%

dataset contains 102 grayscale and corresponding binarized images with textline based ground truth of binarized images. We test our reported algorithm on grayscale images from this dataset but there is no ground truth and evaluation method for grayscale document images. Therefore, for performance evaluation purpose, we map ridges detected from grayscale images to their corresponding binarized images and do the label assignment based on the overlapping between connected components and ridges. Here, we use two standard textline detection evaluation methods [21, 22]. As mentioned earlier, presented method is the extension of our already reported method [10] with the addition of x-line and baseline information extraction using active contours (snakes). Performance evaluation results are same as reported in [10], which is shown in Table 1. Some sample results of our algorithm are shown in Fig. 3.

4. DISCUSSION

The paper describes a novel approach for curled textline information extraction from grayscale camera-captured document images which is independent of binarization. Textline detection accuracy of 91% on the dataset of CDBAR 2007 document image dewarping contest shows the effectiveness of the presented approach. Qualitative evaluation of our x-line and baseline estimation algorithm shows that we are able to accurately track x-line and baseline of curled textlines even in the presence of large number of capital letters, numbers, and quotation marks. Our approach is also robust against high degrees of curl and requires no post-processing. The presented

method can be tuned for textline information extraction from grayscale scanned document images or historical documents where binarization noise presents severe challenges to textline extraction.

5. ACKNOWLEDGMENT

We are grateful to Mr. B. Gatos of Computational Intelligence Laboratory,Athense, Greece, for giving us the textlines segmentation evaluation software.

References

[1] Z. Zhang and C. L. Tan, "Correcting document image warping based on regression of curved text lines," in *Proc. 7th Int. Conf. on Document Analysis and Recognition*, Edinburgh, Scotland, 2003, pp. 589–593.

[2] S. J. Lu and C. L. Tan, "The restoration of camera documents through image segmentation," in *Proc. 7th IAPR workshop on Document Analysis Systems*, Nelson, New Zealand, 2006, pp. 484–495.

[3] B. Fu, M. Wu, R. Li, W. Li, and Z. Xu, "A model-based book dewarping method using text line detection," in *Proc. 2nd Int. Workshop on Camera Based Document Analysis and Recognition*, Curitiba, Barazil, 2007, pp. 63–70.

[4] B. Gatos, I. Pratikakis, and K. Ntirogiannis, "Segmentation based recovery of arbitrarily warped document images," in *Proc. 9th Int. Conf. on Document Analysis and Recognition*, Curitiba, Barazil, 2007, pp. 989–993.

[5] N. Stamatopoulos, B. Gatos, I. Pratikakis, and S. J. Perantonis, "A two-step dewarping of camera document images," in *Proc. 8th IAPR Workshop on Document Analysis Systems*, Nara, Japan, 2008, pp. 209–216.

[6] A. Ulges, C. H. Lampert, and T. M. Breuel, "Document image dewarping using robust estimation of curled text lines," in *Proc. 8th Int. Conf. on Document Analysis and Recognition*, Seoul, Korea, 2005, pp. 1001–1005.

[7] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Segmentation of curled textlines using active contours," in *Proc. 8th IAPR Workshop on Document Analysis Systems*, Nara, Japan, 2008, pp. 270–277.

[8] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Coupled snakelet model for curled textline segmentation of camera-captured document images," in *Proc. 10th Int. Conf. on Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 61–65.

[9] F. Shafait, D. Keysers, and T. M. Breuel, "Efficient implementation of local adaptive thresholding techniques using integral images," in *Proc. Document Recognition and Retrieval XV*, San Jose, CA, USA, 2008, vol. 6815, p. 81510.

[10] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Ridges based curled textline region detection from grayscale camera-captured document images," in *Proc. 13th Int. Conf. on Computer Analysis of Images and Patterns*, Muenster, Germany, 2009.

[11] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, "Detection of blood vessels in retinal images using two-dimensional matched filters," *IEEE Transaction on Medical Imaging*, vol. 8, no. 3, pp. 263–269, 1989.

[12] B. K. P. Horn, "Shape from shading: A method for obtaining the shape of a smooth opaque object from one view," *PhD Thesis, MIT*, 1970.

[13] M. D. Riley, "Time-frequency representation for speech signals," *PhD Thesis, MIT*, 1987.

[14] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. Journal of Computer Vision*, vol. 1, no. 4, pp. 1162–1173, 1988.

[15] F. Shafait, J. van Beusekom, D. Keysers, and T. M. Breuel, "Document cleanup using page frame detection," *Int. Jour. on Document Analysis and Recognition*, vol. 11, no. 2, pp. 81–96, 2008.

[16] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Script-independent handwritten textlines segmentation using active contours," in *Proc. 10th Int. Conf. on Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 446–450.

[17] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, "Script-independent text line segmentation in freestyle handwritten documents," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1313–1329, 2008.

[18] L. O. Gorman, "Matched filter design for fingerprint image enhancement," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, New York, NY, USA, 1988, pp. 916–919.

[19] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," in *IEEE Transaction of Image Processing*, 1998, pp. 359–369.

[20] F. Shafait and T. M. Breuel, "Document image dewarping contest," in *Proc. 2nd Int. Workshop on Camera Based Document Analysis and Recognition*, Curitiba, Brazil, 2007, pp. 181–188.

[21] B. Gatos, A. Antonacopoulos, and N. Stamatopoulos, "ICDAR 2007 handwriting segmenentation contest," in *Proc. 9th Int. Conf. on Document Analysis and Recognition*, Curitiba, Brazil, 2007, pp. 1284–1288.

[22] F. Shafait, D. Keysers, and T. M. Breuel, "Pixel-accurate representation and evaluation of page segmentation in document images," in *Proc. Int. Conf. on Pattern Recognition*, Hong Kong, China, Aug 2006, pp. 872–875.

[23] F. Shafait, D. Keysers, and T. M. Breuel, "Performance evaluation and benchmarking of six page segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 941–954, 2008.