PERFORMANCE EVALUATION OF CURLED TEXTLINE SEGMENTATION ALGORITHMS ON CBDAR 2007 DEWARPING CONTEST DATASET

Syed Saqib Bukhari[1], Faisal Shafait[2] and Thomas M. Breuel[1]

[1]Technical University of Kaiserslautern, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
bukhari@informatik.uni-kl.de, faisal.shafait@dfki.de, tmb@informatik.uni-kl.de

ABSTRACT

Camera-captured document images often contain curled textlines because of geometric and perspective distortions. Finding curled textlines, which is more difficult than straight textline detection, is a primary step in the processing of hand-held camera-captured document images. Detected textlines results can be used for dewarping of warped images, layout-analysis, etc. In this paper, we compare previously reported curled textline segmentation techniques by using the publicly available CBDAR 2007 dewarping contest dataset and vectorial performance evaluation metrics.

1. INTRODUCTION

Hand-held cameras are widely used for capturing document and screen-text images. Unlike scanned document image which consists of straight textlines, camera-captured document images are composed of curled textlines due to geometric and perspective distortions. Therefore, most commercial and open-source OCR systems, which are designed specifically for straight textlines document images, produce lot of garbage text for curled document images. OCR results of camera-captured document images can be improved either by designing novel recognition techniques for curled documents or by designing dewarping techniques for warped images so that current scanner-based OCR system can be applied on dewarped (planar) images. When only a single image is available for removing line curl, this is referred to as monocular dewarping. A typical monocular dewarping algorithm performs dewarping using curled textlines segmentation results [1]. Here textlines segmentation means finding textlines from document images. Curled textlines information can also be used for layout-analysis and text-recognition techniques for camera-captured document images.

Curled textline segmentation is an active research field in camera-based document analysis. There are variety of curled textline segmentation techniques in the literature [2, 3, 4, 5, 6, 7], but no work has been done for the performance evaluation and comparison of these techniques on common dataset and performance evaluation metrics. In this paper, we describe our publicly available CBDAR 2007 dewarping contest dataset [8] and vectorial performance evaluation metrics [9] for comparing curled textline segmentation algorithms. Here, we have selected following previously reported curled textline segmentation algorithms for performance evaluation and comparison: i) neighborhood-distance based approach [2], ii) baby-snakes [3], iii) coupled-snakelets [4], iv) smoothing and ridges based approach [5, 6] and v) skew detection based approach [7]. In future, researchers can use the above mentioned dataset and performance evaluation metrics and can compare their curled textline segmentation results to our baseline.

An initial version of this work was presented in international workshop on document analysis systems (DAS) 2010 [10]. This paper is an extended version of our previous work [10] as it has more curled textline segmentation algorithms as well as Docstrum [11] algorithm for comparison. Docstrum [11] is the state-of-the-art straight textline segmentation approach. It is included here to show that straight textline segmentation algorithm can not be directly used for curled textline segmentation.

The rest of the paper is organized as follows: Section 2 briefly describes curled textline segmentation algorithms. CBDAR 2007 dataset [8] is defined in Section 3. Section 4 describes the performance evaluation metrics [9] and results of different curled textline segmentation algorithms. Section 5 discusses the impact of our work.

2. CURLED TEXTLINE SEGMENTATION

Brief descriptions of selected curled textline segmentation algorithms are presented below.

2.1. Neighborhood-Distance based Algorithm [2]

This algorithm detects curled textlines by using nearest neighbor criteria. For each connected component, right successor is determined on the basis of minimum distance and overlap between bounding boxes of connected components. This type of grouping between a component and its right neighbor is resulted in forest of trees, where each tree represents

connected components belong to a single curled textline. After finding textlines, baselines and descender lines are determined using RAST [12] based geometric model fitting technique. For achieving better textline segmentation results, at first Voronoi based page segmentation approach [13] is used and then aforementioned textline finding technique is applied on each segmented block individually.

2.2. Baby-Snakes Algorithm [3]

Active contour (snakes) [14] is one of the state-of-the-art photographic image segmentation technique. Baby-snakes algorithm adapts active contour for curled textline segmentation from document images. Open-curve slope-aligned snakes are initialized over smeared connected components. External energy using GVF (gradient vector flow) [15] is calculated from smeared document image, that is used for baby-snakes deformation. Neighboring baby-snakes are joined together after a few deformation steps and are resulted in textlines detection.

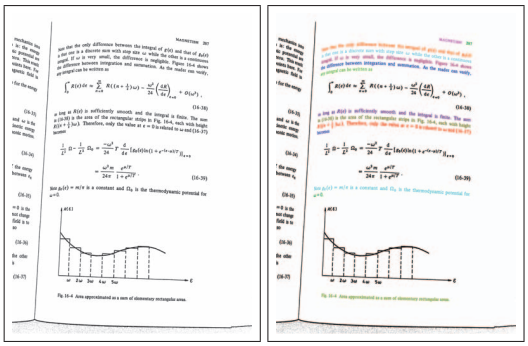
2.3. Coupled-Snakelets Algorithm [4]

This approach is also based on active contour (snakes) [14], but different from baby-snakes algorithm [3]. This approach solves both the problems of textlines segmentation and x-line-baseline pairs estimation. A pair of straight open-curve snakes is initialized over a connected component's top and bottom points, referred to as top- and bottom-snake. Then the top-snake is deformed using 50% weights and bottom-snake is deformed using 100% weights of the vertical components of GVF of neighboring top and bottom points respectively. The same procedure is repeated few more times with incremental increase in snakes length and deformation regions. The same procedure is repeated for all connected components within an image. Overlapping pairs of snakes are resulted as segmented curled textlines. The extended coupled-snakelets version[1] contains an additional step of removing badly deformed snakelets pairs on the basis of neighboring snakelets properties.

2.4. Smoothing and Ridges based Algorithm [5, 6]

In this approach multi-scale and multi-orientated anisotropic Gaussian smoothing is used for enhancing curled textline structure. Then central lines of textlines are detected by using Horn-Riley [16, 17] based ridges detection technique. These detected ridges are resulted in segmented textlines. This algorithm is designed for segmenting curled textlines directly from badly illuminated grayscale camera captured document images. It is also equally applicable on the binarized document images as well.

[1]The extended version of our coupled-snakelets algorithm [4] is in review phase of IJDAR special issue on ICDAR2009 selected papers.



(a) Original Image (b) Labeled Textlines Only

Fig. 1. An example image and its corresponding textline-based ground-truth image from CBDAR 2007 dataset [8].

2.5. Skew Detection based Algorithm [7]

This algorithm is based on the combination of skew detection [18] and simplified idea of coupled-snakelets [4] techniques. At first, by using these techniques, small local textlines are estimated and then these small lines are merged together to achieve segmented textlines. After finding textlines, regression model is applied for finding x-line-baseline pairs of segmented textlines.

3. CBDAR 2007 DEWARPING CONTEST DATASET

CBDAR 2007 document image dewarping contest dataset [8] consists of grayscale and binarized document images. These images are captured from several technical books by using hand-held camera in an office environment and are composed of curled textlines due to geometric and perspective distortions. This dataset contains ASCII-text ground-truth and pixel-based ground-truth for zones, textlines, formulas, tables and figures. Pixel-based color-coded ground-truth is defined as follows: i) red channel contains zone class information, ii) blue channel contains zone number (in reading order) information, iii) green channel contains textline number information which is equal to zero for formulas, tables and figures and iv) marginal noise and foreground objects outside page boundary are marked with black color (all three color channels are set equal to zero).

For curled textline segmentation performance evaluation, textline-based ground-truth images are generated automatically by using color-coded information. A textline-based ground-truth image contains labeling only for textlines such that all other foreground objects within page boundary where green channel equals to zero like formulas, tables and figures are marked as noisy pixels with black color. An example image with its textline-based ground-truth is shown in Figure 1.

4. PERFORMANCE EVALUATION

Performance evaluation of curled textline segmentation algorithms is based on vectorial metrics which are presented in [9]. These metrics are not only the representative of one-to-one segmentation accuracy, but also of the most important classes of segmentation errors (over-, under-, and miss-segmentation). The descriptions of performance evaluation metrics are as follows. Consider we have two segmented images, the ground-truth G and hypothesized segmentation H. We can compute a weighted bipartite graph called "pixel-correspondence graph" between G and H for evaluating the quality of the segmentation algorithm. Each node in G or H represents a segmented component. An edge is constructed between two nodes such that the weight of the edge equals the number of foreground pixels in the intersection of the regions covered by the two segments represented by the nodes. The matching between G and H is perfect if there is only one edge incident on each component of G or H, otherwise it is not perfect, i.e. each node in G or H may have multiple edges. The edge incident on a node is significant if the value of $w_i/P \geq t_r$ and $w_i \geq t_a$, where w_i is the edge-weight, P is the number of pixels corresponding to a node (segment), t_r is a relative threshold and t_a is a absolute threshold. In practice, $t_r = 0.1$ and $t_a = 100$ are good choices for textlines based performance evaluation [9]. Performance evaluation metrics are defined as follows:

- **Total correct segmentation (N_{o2o}):** the number of one-to-one matches between the ground-truth components and the segmentation components.
- **Oversegmented components (N_{ocomp}):** the number of ground-truth lines having more than one significant edge.
- **Undersegmented components (N_{ucomp}):** the number of segmented lines having more than one significant edge.
- **Missed components (N_{mcomp}):** the number of ground-truth components that matched the background in the hypothesized segmentation.
- **Total oversegmentations (N_{oseg}):** the number of significant edges that ground-truth lines have, minus the number of ground-truth lines.
- **Total undersegmentations (N_{useg}):** the number of significant edges that segmented lines have, minus the number of segmented lines.
- **False alarms (N_{falarm}):** the number of components in the hypothesized segmentation that did not match any foreground component in the ground-truth segmentation.

The performance evaluation results of curled textline segmentation algorithms and Docstrum [11] on CBDAR 2007 dewarping contest dataset are shown in Table 1. Docstrum is one of the sate-of-the-art page segmentation algorithm for scanned document images with straight textlines. The main reason of including it in the performance evaluation of curled textline segmentation is to show that how challenging the dataset is, and that straight textlines segmentation algorithms can not be directly applicable on curled document images, which is clearly shown in Table 1. Other than textlines, document images in the dataset also contain large number of noisy text components outside page boundary, figures, formulas, tables and marginal noise. Often curled textline segmentation algorithms detect most of the non-textline components as textlines and produce large number of false alarm (N_{falarm}) errors, as shown in Table 1. False alarm errors can be reduced by size and page boundary based post-processing filtering operation.

As compared to other curled textline segmentation techniques, the extended version of coupled-snakelets algorithm (Section 2.3) gives better compromise between one-to-one segmentation accuracy and over- and undersegmentation errors, as shown in Table 1.

5. DISCUSSION

In this paper, we have presented the common platform for the performance evaluation and comparison of five different curled textline segmentation algorithms using standard dataset and performance evaluation metrics. We have used publicly available camera-captured document image dataset [8] containing 102 curled or warped document images, which were introduced in CBDAR 2007 document image dewarping contest. We have used vectorial performance evaluation metrics [9] instead of just a single score. These performance metrics are good representative of accuracy as well as crucial errors of curled textline segmentation, like missed components, under and oversegmentations. Extended version of coupled snakelets algorithm (Section 2.3) is better than others in terms of the best one-to-one textlines finding accuracy of 95.21% with 0% missed textlines and fewer over- and undersegmentation errors, as shown in Table 1. We have also shown in Table 1 that the straight textlines segmentation approaches can not be directly applicable on curled textlines finding. We hope that this paper will help the community to compare further curled textline segmentation algorithms to our baseline.

6. REFERENCES

[1] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Dewarping of document images using coupled-snakes," in *Proc. of 3rd Int. Workshop on Camera-Based Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 34–41.

Table 1. Performance evaluation results of curled textline segmentation algorithms on CBDAR 2007 dataset [8] by using vectorial performance evaluation metrics [9].

Algorithm	Metrics[a]								
	N_g	N_s	N_{falarm}	N_{useg}	N_{oseg}	$P_{ucomp}\%$	$P_{comp}\%$	$P_{mcomp}\%$	$P_{o2o}\%$
Extended Coupled-Snakelets [b]	3091	3093	3277	51	54	1.59%	1.68%	**0%**	**95.21%**
Neighborhood-Distance [2]	3091	3134	2491	**18**	134	**0.55**	3.33	2.59	93.47%
Skew-Detection [7]	3091	2924	**785**	57	682	1.81%	21.71%	4.43%	91.10%
Ridges [5, 6] (binary)	3091	3115	2183	110	144	3.30%	4.40%	0.29%	89.65%
Ridges [5, 6] (grayscale)	3091	3045	1186	131	115	3.85%	3.53%	0.91%	89.10%
Baby-Snakes [3]	3091	3371	13199	117	294	2.91%	5.79%	**0%**	87.58%
Coupled-Snakelets [4]	3091	2799	3251	359	**39**	9.06%	**1.26%**	**0%**	78.26%
Doctsrum [11] [c]	3091	6852	6066	2096	4383	51.50%	66.90%	**0%**	21.26%

[a] N_g:ground-truth components; N_s:segmented components; N_{o2o}:one-to-one matched components; $P_{o2o}\% = N_{o2o}/N_g$;N_{oseg}: oversegmentations;N_{useg}: undersegmentations; N_{ocomp}:oversegmented components; $P_{ocomp}\% = N_{ocomp}/N_g$; N_{ucomp}: undersegmented components; $P_{ucomp}\% = N_{ucomp}/N_g$; N_{mcomp}: missed components; $P_{mcomp}\% = N_{mcomp}/N_g$; N_{falarm}: false alarms;

[b] The extended version of our coupled-snakelets [4] algorithm is in review phase of IJDAR special issue on ICDAR2009 selected papers.

[c] Docstrum [11] is used for straight line scanned document image segmentation. Here it is used for curled textline segmentation to show that: i) straight textlines algorithm can not be directly application on camera-captured documents and ii) the CBDAR 2007 dataset is challenging with respect to curled textlines.

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

[3] 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.

[4] 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.

[5] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Ridges based curled textline region detection from grayscale camera-captured document images," in *Proc. The 13th Int. Conf. on Computer Analysis of Images and Patterns*, Muenster, Germany, 2009, vol. 5702 of *Lecture Notes in Computer Science*, pp. 173–180.

[6] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Textline information extraction from grayscale camera-captured document images," in *Proc. The 13th Int. Conf. on Image Processing*, Cairo, Egypt, 2009, pp. 2013–2016.

[7] D. M. Oliveira, R. D. Lins, G. Torreo, J. Fan, and M. Thielo, "A new method for text-line segmentation for warped document," in *Proc. of Int. Conf. on Image Analysis and Recognition*, Povoa de Varzim, Portugal, 2010, pp. 398–408.

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

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

[10] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Performance evaluation of curled textlines segmentation algorithms," in *9th IAPR Workshop on Document Analysis Systems (short paper)*, Boston, MA, USA, 2010.

[11] L. O'Gorman, "The document spectrum for page layout analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1162–1173, 1993.

[12] T. M. Breuel, "Recognition by adaptive subdivision of transformation space: practical experiences and comparison with the hough transform," in *Hough Transforms, IEE Colloquium on*, 7 1993, pp. 7/1 –7/4.

[13] K. Kise, A. Sato, and M. Iwata, "Segmentation of page images using the area Voronoi diagram," *Computer Vision and Image Understanding*, vol. 70, no. 3, pp. 370–382, 1998.

[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] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," in *IEEE Trans. of Image Processing*, 1998, vol. 7, pp. 359–369.

[16] 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.

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

[18] B. T. vila and R. D. Lins, "A fast orientation and skew detection algorithm for monochromatic document images.," in *Proc. of the ACM Symposium on Document Engineering*, Bristol, UK, 2005, pp. 118–126.