Performance Evaluation of Curled Textlines Segmentation Algorithms

Syed Saqib Bukhari
Technical University of Kaiserslautern, Germany
bukhari@informatik.uni-kl.de

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

Thomas M. Breuel
Technical University of Kaiserslautern, Germany
tmb@informatik.uni-kl.de

ABSTRACT

Curled textlines segmentation is a necessary initial step for the hand-held camera-captured document image processing. Curled textlines information is often used as an intermediate step for camera-captured document image dewarping. Curled textlines information can also be used for other camera-based document image processing tasks, like layout analysis etc. So far no work has been done for the performance evaluation and the comparison of already reported curled textlines segmentation algorithms on the standard datasets and performance evaluation metrics. In this paper, we use publicly available camera-captured document image dataset and vectorial performance evaluation metrics to compare different algorithms. We present performance evaluation results on previously reported curled textlines segmentation algorithms.

1. INTRODUCTION

The hand-held camera-captured document image analysis is an open field. There is growing application for the research in the field of camera-based document image OCR. Most of the current commercial and open-source OCR systems provide support for planar scanned document images. These OCR systems produce bad results when applied directly on the warped camera-captured images. There are two possible solutions for improving OCR results of camera-captured document images: i) design novel techniques for the camera-captured document images or ii) design dewarping techniques for the warped images so that current scanner-based OCR can be applied on dewarped (planar) images. So far camera-captured document image community has provided different dewarping algorithms. A typical monocular dewarping algorithm performs dewarping using curled textlines segmentation results [17, 9, 15, 6, 14, 3]. Textlines segmentation means finding textlines from document images. In the future there may be new layout analysis and text-recognition algorithms for solving camera-based document image OCR problem without using dewarping. Curled textlines information can also play an important role in these new algorithms.

DAS '10, June 9-11, 2010, Boston, MA, USA

Therefore curled textlines segmentation is an important initial step for camera-based document image analysis.

But till date there is no common dataset and performance evaluation metrics which have been used for the evaluation of different curled textlines segmentation algorithms. In this paper, we describe our publicly available CBDAR 2007 dewarping contest dataset [12] and vectorial performance evaluation metrics [13] for comparing curled textlines segmentation algorithms. We have already developed three different curled textlines segmentation algorithms [1, 4, 2] for camera-captured document images. Oliveria et al. [10] have also reported curled textlines segmentation algorithm. Here we present the comparative results of above mentioned curled textlines segmentation algorithms using CBDAR 2007 dewarping contest dataset [12] and vectorial performance evaluation metrics [13]. In future, researchers can use the above mentioned dataset and performance evaluation metrics and can compare their curled textlines segmentation results to our baseline.

The rest of the paper is organized as follows: Section 2 briefly describes curled textlines segmentation algorithms [1, 4, 2, 10]. CBDAR 2007 dataset [12] is defined in Section 3. Section 4 explains the performance evaluation metrics [13] and compares results of different curled textlines segmentation algorithms. Section 5 discusses the impact of our work.

2. CURLED TEXTLINES SEGMENTATION

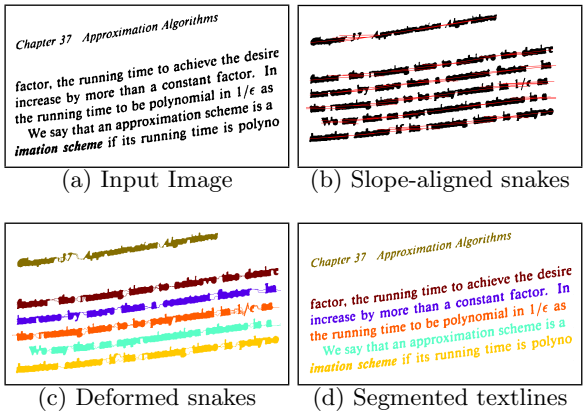
Short description of curled textlines segmentation algorithms [1, 4, 2, 10] are presented below.

2.1 Baby-Snakes Algorithm [1]

This algorithm is based on active contours (snakes) [8], which are traditionally used for the photographic image segmentation. Open-curve slope-aligned snakes are initialized over smeared connected components, referred to as "baby-snakes". Then GVF (gradient vector flow) [16] as an external energy is calculated from the smeared document image. This energy is used for baby-snakes deformation. After a few deformation steps, neighboring baby-snakes are joined together and resulted in textlines detection. A graphical representation of the algorithm discussed is shown in Figure 1.

2.2 Ridges-Based Algorithm [4, 5]

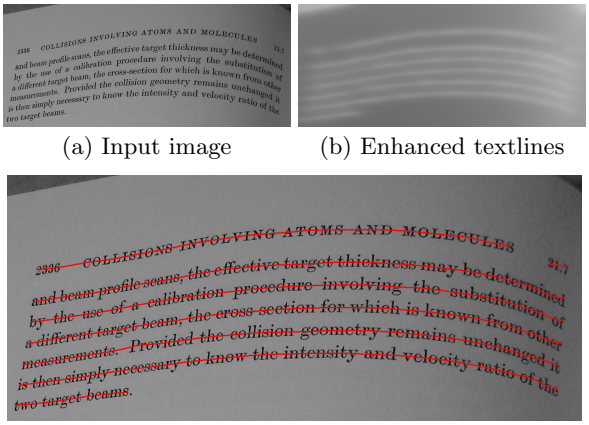
This algorithm is initially designed for segmenting curled textlines directly from grayscale camera captured document images, but is equally applicable on the binarized images



(a) Input Image (b) Slope-aligned snakes

(c) Deformed snakes (d) Segmented textlines

Figure 1: Baby-snakes algorithm snapshots [1].



(a) Input image (b) Enhanced textlines

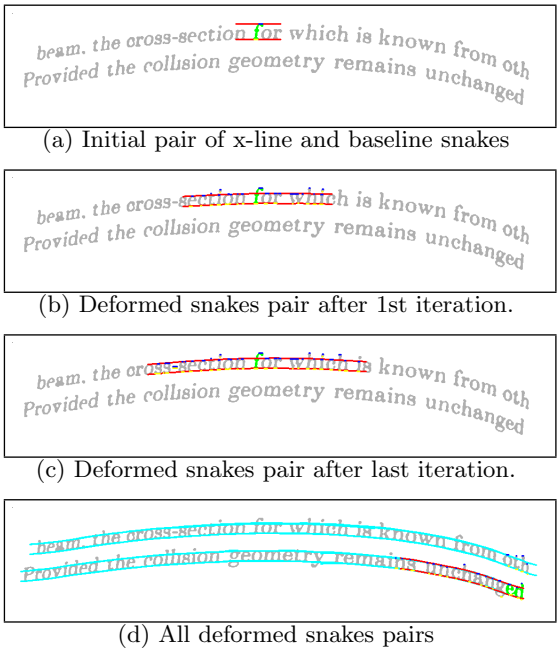
(c) Detected ridges from smoothed image

Figure 2: Different phases of curled textline information extraction algorithm [4, 5].

as well. The multi-scale multi-orientated anisotropic Gaussian smoothing is used for enhancing curled textline structures. Then the central line of textlines are detected by using Horn-Riley [7, 11] based ridges detection technique. These detected ridges are resulted in segmented textlines. A graphical representation of this algorithm is shown in Figure 2.

2.3 Coupled-Snakelets Algorithm [2]

This algorithm presents another approach for curled textline information extraction, referred to as "coupled-snakelets" [2]. This approach is also based on active contour (snakes) [8], but different from baby-snakes algorithms [1]. 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 of the vertical components of GVF of neighboring top points and bottom snake is deformed using 100% weights of vertical components of GVF of neighboring bottom points. The same procedure is repeated few more times with incremental increase in the snakes length and the deformation regions. The same procedure is repeated for all the connected components within an image. Overlapping pairs of snakes are resulted



(a) Initial pair of x-line and baseline snakes

(b) Deformed snakes pair after 1st iteration.

(c) Deformed snakes pair after last iteration.

(d) All deformed snakes pairs

Figure 3: Flow of Coupled-Snakelets algorithm [2].

as segmented curled textlines. A graphical representation of this algorithm is shown in Figure 3. The extended coupled-snakelets version[1] contains an additional step of removing badly deformed snakelets pairs on the basis of neighboring snakelets properties.

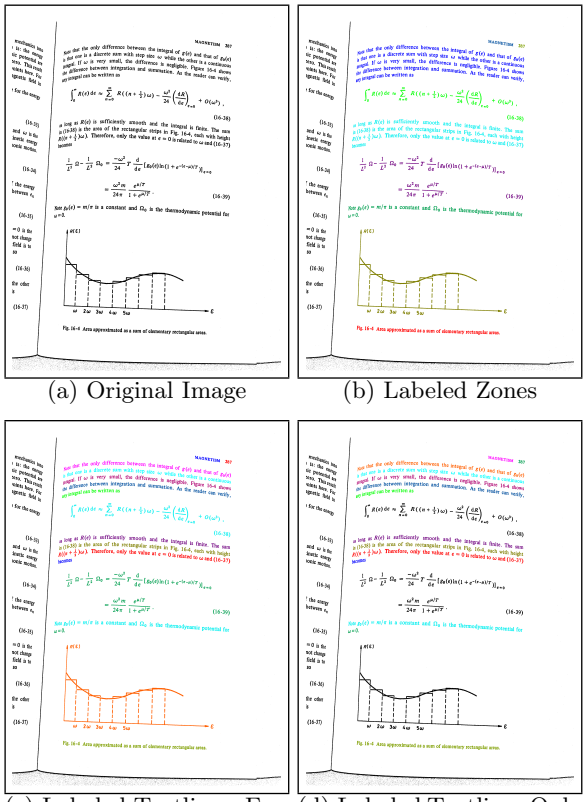
2.4 Rule-Based Algorithm [10]

Oliveria et al. [10] presented a curled textlines segmentation algorithm. In this algorithm rule based criteria is used for segmenting textlines, such that for each unvisited connected component, neighboring components are grouped together based on defined rules. After finding textlines, regression model is applied for finding x-line-baseline pairs of segmented textlines.

3. CBDAR 2007 DATASET [12]

CBDAR 2007 dataset contains 102 grayscale and binarized images of pages from several technical books captured by an off-the-shelf hand-held digital camera in a normal office environment. The captured documents were binarized using a local adaptive thresholding technique [15]. Together with ASCII-text ground-truth, this data also contains pixel-based ground-truth for zones, textlines, formulas, tables and figures. This pixel-based ground-truth are embedded in color-coding format, where red channel contains zone class information, blue channel contains zone number (in reading order) information and green channel contains textline number information. The value of green channel is zero for formulas, tables and figures. Marginal noise and foreground objects outside the page-boundary are marked with black color. We generated textlines-based ground-truth images from the actual ground-truth images for the performance

[1]The extended version of our coupled-snakelets algorithm [2] is in review phase of IJDAR special issue on ICDAR2009 selected papers.



(a) Original Image (b) Labeled Zones

(c) Labeled Textlines, Formulas and Figure (d) Labeled Textlines Only

Figure 4: An example image and its corresponding pixel-based color-coded ground-truth images from CBDAR 2007 dataset [12].

evaluation of curled textlines segmentation algorithms. The textlines-based ground-truth images contain labeling only for textlines and all the other foreground objects within page-boundary, like formulas, tables and figures, are marked as noisy pixels with black color. Textlines-based ground-truth images have been generated automatically by using actual ground-truth information as follows: in the actual ground-truth images, green color channel value is zero for all the foreground objects within page-boundary except textlines. All the foreground pixels are marked as noisy pixels having green channel equal to zero. An example image with different types of pixel-based ground-truth are shown in Figure4. Textlines-based ground-truth is shown in Figure 4(d).

4. PERFORMANCE EVALUATION [13]

The performance evaluation of curled textlines segmentation algorithms is based on vectorial metrics which are presented in [13]. 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 importance of vectorial metrics score is already highlighted in Shafait et al. [13]. 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 ta is a absolute threshold. In practice, $t_r = 0.1$ and $t_a = 100$ are good choices for textlines based performance evaluation [13]. On the basis of the above description the performance evaluation metrics are:

- **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 comparative performance evaluation results of different curled textlines segmentation algorithms are shown in Table 1. In general, the extended version of coupled-snakelets algorithm performs better than other curled textline segmentation algorithms with respect to the majority of performance evaluation metrics.

5. DISCUSSION

In this paper, we have described a common platform for evaluating the performance of curled textlines segmentation algorithms. For this purpose, we have used publicly available camera-captured document image dataset [12] containing 102 grayscale and binarized images. For the performance evaluation we have used vectorial performance metrics [13] instead of just a single score. These performance metrics are the representative of accuracy as well as crucial errors of curled textlines segmentation, like the missed components, under and over-segmentations. The performance evaluation of different curled textlines segmentation algorithms is

Table 1: Performance Evaluation Results of Curled Textlines Segmentation Algorithms [1, 2, 4, 10] on CBDAR 2007 dataset [12] by using vectorial performance evaluation metrics [13].

Algorithm	Metrics[a]								
	N_g	N_s	$P_{o2o}\%$	$P_{ocomp}\%$	$P_{ucomp}\%$	$P_{mcomp}\%$	N_{oseg}	N_{useg}	N_{falarm}
Coupled-Snakelets [2]	3091	2799	78.26%	**1.26%**	9.06%	**0%**	**39**	359	3251
Baby-Snakes [1]	3091	3371	87.58%	5.79%	2.91%	**0%**	294	117	13199
Ridges-Based [4, 5](gray[b])	3091	3045	89.10%	3.53%	3.85%	0.91%	115	131	1186
Ridges-Based [4, 5](binary[c])	3091	3115	89.65%	4.40%	3.30%	0.29%	144	110	2183
Rule-Based [10]	3091	2924	91.10%	21.71%	1.81%	4.43%	682	57	**785**
Extended Coupled-Snakelets [d]	3091	3093	**95.21%**	1.68%	**1.59%**	**0%**	54	**51**	3277

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

[b] detects textlines from grayscale images and than maps on binary images (as mentioned in [4, 5])

[c] detects textlines directly from binary images

[d] The extended version of our coupled-snakelets [2] is in review phase of IJDAR special issue on ICDAR2009 selected papers.

shown in Table 1. We hope that this paper will help the community to evaluate curled textlines segmentation algorithms on common grounds.

6. REFERENCES

[1] 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*, pages 270–277, Nara, Japan, 2008.

[2] 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*, pages 61–65, Barcelona, Spain, 2009.

[3] 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*, pages 34–41, Barcelona, Spain, 2009.

[4] 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*, volume 5702 of *Lecture Notes in Computer Science*, pages 173–180, Muenster, Germany, 2009.

[5] 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*, pages 2013–2016, Cairo, Egypt, 2009.

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

[7] 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.

[8] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int. Journal of Computer Vision*, 1(4):1162–1173, 1988.

[9] S. Lu and C. Tan. The restoration of camera documents through image segmentation. In *7th IAPR Workshop on Document Analysis Systems*, pages 484–495, Nelson, New Zealand, Feb. 2006.

[10] D. M. Oliveira, R. D. Lins, G. TorreÃčo, J. Fan, and M. Thielo. A new method for text-line segmentation for warped document. In *Proc. of Int. Conf. on Image Analysis and RecognitionInt*, Povoa de Varzim, Portugal, 2010.

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

[12] F. Shafait and T. M. Breuel. Document image dewarping contest. In *2nd Int. Workshop on Camera-Based Document Analysis and Recognition*, pages 181–188, Curitiba, Brazil, 20097.

[13] 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*, 30(6):941–954, 2008.

[14] N. Stamatopoulos, B. Gatos, I. Pratikakis, and S. J. Perantonis. A two-step dewarping of camera document images. In *Proc. of The Eight IAPR Workshop on Document Analysis Systems*, pages 209–216, 2008.

[15] A. Ulges, C. Lampert, and T. Breuel. Document image dewarping using robust estimation of curled text lines. In *Proc. Eighth Int. Conf. on Document Analysis and Recognition*, pages 1001–1005, Aug. 2005.

[16] C. Xu and J. L. Prince. Snakes, shapes, and gradient vector flow. In *IEEE Trans. of Image Processing*, volume 7, pages 359–369, 1998.

[17] Z. Zhang and C. L. Tan. Correcting document image warping based on regression of curved text lines. In *Proc. of the Int. Conf. on Document Analysis and Recognition*, pages 589–593, 2003.