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 datasets 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 a 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. 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 CBDAIR 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 CBDAIR 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], CBDAIR 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
Figure 1: Baby-snakes algorithm snapshots [1].

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

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 as segmented curled textlines. A graphical representation of this algorithm is shown in Figure 3. The extended coupled-snakelets version 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 as mentioned above.

1The extended version of our coupled-snakelets algorithm [2] is in review phase of IJDAR special issue on ICDAR2009 selected papers.
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 differ-
et types of pixel-based ground-truth are shown in Figure 4.
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 impor-
tant 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 hypo-
thesized 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 compo-
nent 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 cor-
responding to a node (segment), \( t_r \) is a relative threshold
and \( t_a \) is an 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** \( (N_{mcomp}) \): the number of
  one-to-one matches between the ground-truth compo-
nents and the segmentation components.

- **Oversegmented components** \( (N_{oseg}) \): the num-
  ber of ground truth lines having more than one signif-
  icant edge.

- **Undersegmented components** \( (N_{useg}) \): the num-
  ber 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 oversegments** \( (N_{oseg}) \): 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** \( (N_{false}) \): the number of components
  in the hypothesized segmentation that did not match
  any foreground component in the ground-truth seg-
  mentation.

The comparative performance evaluation results of different
curled textlines segmentation algorithms are shown in Table
1. In general, the extended version of coupled-snakelets algo-

rithm performs better than other curled textline segmenta-
tion 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 avail-
able camera-captured document image dataset [12] contain-
ing 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 evalua-
tion 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].

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$N_g$</th>
<th>$N_s$</th>
<th>$P_{seg}$%</th>
<th>$P_{comp}$%</th>
<th>$P_{falarm}$%</th>
<th>$N_{oseg}$</th>
<th>$N_{useg}$</th>
<th>$N_{false}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coupled-Snakelets [2]</td>
<td>3091</td>
<td>2799</td>
<td>78.26%</td>
<td>1.26%</td>
<td>9.06%</td>
<td>0</td>
<td>39</td>
<td>359</td>
</tr>
<tr>
<td>Baby-Snakes [1]</td>
<td>3091</td>
<td>3371</td>
<td>87.58%</td>
<td>5.79%</td>
<td>2.91%</td>
<td>0</td>
<td>294</td>
<td>117</td>
</tr>
<tr>
<td>Ridges-Based <a href="gray">4, 5</a></td>
<td>3091</td>
<td>3045</td>
<td>89.10%</td>
<td>3.53%</td>
<td>0.91%</td>
<td>0</td>
<td>115</td>
<td>131</td>
</tr>
<tr>
<td>Ridges-Based <a href="binary">4, 5</a></td>
<td>3091</td>
<td>3115</td>
<td>89.65%</td>
<td>4.40%</td>
<td>0.29%</td>
<td>0</td>
<td>144</td>
<td>110</td>
</tr>
<tr>
<td>Rule-Based [10]</td>
<td>3091</td>
<td>2924</td>
<td>91.10%</td>
<td>21.71%</td>
<td>4.43%</td>
<td>682</td>
<td>57</td>
<td>785</td>
</tr>
<tr>
<td>Extended Coupled-Snakelets</td>
<td>3091</td>
<td>3093</td>
<td>95.21%</td>
<td>1.68%</td>
<td>1.59%</td>
<td>0</td>
<td>54</td>
<td>51</td>
</tr>
</tbody>
</table>

* $N_g$: ground-truth components; $N_s$: segmented components; $N_{oseg}$: one-to-one matched components; $P_{seg}$% = $N_{seg}$/$N_g$; $N_{useg}$: oversegmented components; $P_{comp}$% = $N_{comp}$/$N_g$; $N_{false}$: false alarms
* $N_{oseg}$: missed components; $P_{false}$% = $N_{false}$/$N_g$; $N_{false}$: undersegmentations; $N_{falarm}$: false alarms
* detect textlines from grayscale images and than maps on binary images (as mentioned in [4, 5])

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