Ruling Line Removal in Handwritten Page Images

Daniel Lopresti
Department of Computer Science and Engineering
Lehigh University
Bethlehem, PA 18015, USA
lopresti@cse.lehigh.edu

Ergina Kavallieratou
Department of Information and Communication Systems Engineering
University of the Aegean
Samos, Greece
kavallieratou@aegean.gr

Abstract

In this paper we present a procedure for removing ruling lines from a handwritten document image that does not break existing characters. We take advantage of common ruling line properties such as uniform width, predictable spacing, position vs. text, etc. The proposed process has no effect on document images without ruling lines, hence no a priori discrimination is required. The system is evaluated on synthetic page images in five different languages.

1. Introduction

Line processing is a necessary task in many systems, including graphic/text discrimination [1], form or invoice processing [2-3], and engineering drawings [4]. Among these, handwritten documents are a special case where ruling lines are used as guides to make it easier to write neatly. Such lines generally share some common characteristics:

1. They are uniform in thickness.
2. Their position is predictable on the page.
3. They are lighter in color and thickness than the handwritten text. Because of this, they often appear broken in thresholded document images.
4. Even careful writers often overlap ruling lines.

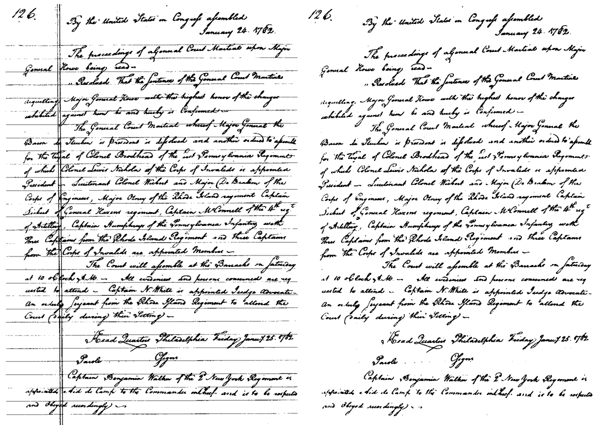
In previous work, Arvind et al. [5] proposed a method where the document image is cleaned of noise, segmented into blocks, and skew-corrected. Ruling lines are detected and then removed based on a run-length analysis. Finally, overlapping handwritten characters are repaired by detecting strokes and filling in the missing area. They quote a subjective evaluation with an accuracy of 86.33%.

Abd-Almageed et al. [6] introduced a page rule line removal algorithm based on linear subspaces. During a training phase, they incrementally construct linear subspaces representing horizontal and vertical lines using a set of rule line-only images. During the testing phase, they measure the distance between features extracted from the test image and the previously constructed subspace. To identify rule line pixels, the feature vector is projected onto the subspace model and the reconstruction error is computed. If the error is larger than an experimentally determined threshold, the pixels are considered foreground; otherwise they are ruling line pixels. They also introduce a scheme for evaluating noise removal which we use in our work as well. Their subspace method achieves approximately 88% for both recall and precision on 50 test images of Arabic handwriting.

In this paper, we present a methodology for line detection and removal capable of eliminating ruling lines (continuous, broken, or overlapping) from a page image without damaging the text significantly so that it will need an extra repairing task. Moreover, it does not require training or present a large computational cost. Since it has no effect on documents without ruling lines, the same procedure can be applied in all cases. Thus, no previous discrimination of the pages is required. In Fig. 1, we show an example of the kind of input that our system can handle.

A description of the proposed system follows in Section 2, while experimental results and a discussion of future work are presented in Sections 3 and 4, respectively.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.663



126.
By the United States in Congreſs aſsembled
January 24. 1782.
The proceedings of a General Court Martial upon Major
General Howe being read —
„ Resolved That the Sentence of the General Court Martial
acquitting Major General Howe with the highest honor of the charges
exhibited against him be and hereby is Confirmed —
The General Court Martial whereof Major General the
Baron de Steuben is President is dissolved and another ordered to aſsemble
for the tryal of Colonel Brodhead of the 1st Pennsylvania Regiment
of which Colonel Lewis Nichola of the Corps of Invalids is appointed
President — Lieutenant Colonel Wiebert and Major (De Brahm of the
Corps of Engineers, Major Olney of the Rhode Island regiment Captain
Seibert of General Hazens regiment, Captain McConnell of the 4th Regt
of Artillery, Captain Humphreys of the Pennsylvania Infantry with
three Captains from the Rhode Island Regiment and three Captains
from the Corps of Invalids are appointed Members —
The Court will aſsemble at the Barracks on Saturday
at 10 oClock A.M. — All evidences and persons concerned are req
uested to attend — Captain N. White is appointed Judge Advocate —
An orderly Serjeant from the Rhode Island Regiment to attend the
Court (daily during their Sitting) —
Head Quarters Philadelphia Friday Janry 25. 1782
Parole . Cygnus
Captain Benjamin Walker of the 2d New York Regiment is
appointed Aid de Camp to the Commander in Chief and is to be respected
and obeyed accordingly —

126.
By the United States in Congreſs aſsembled
January 24. 1782.
The proceedings of a General Court Martial upon Major
General Howe being read —
„ Resolved That the Sentence of the General Court Martial
acquitting Major General Howe with the highest honor of the charges
exhibited against him be and hereby is Confirmed —
The General Court Martial whereof Major General the
Baron de Steuben is President is dissolved and another ordered to aſsemble
for the tryal of Colonel Brodhead of the 1st Pennsylvania Regiment
of which Colonel Lewis Nichola of the Corps of Invalids is appointed
President — Lieutenant Colonel Wiebert and Major (De Brahm of the
Corps of Engineers, Major Olney of the Rhode Island regiment Captain
Seibert of General Hazens regiment, Captain McConnell of the 4th Regt
of Artillery, Captain Humphreys of the Pennsylvania Infantry with
three Captains from the Rhode Island Regiment and three Captains
from the Corps of Invalids are appointed Members —
The Court will aſsemble at the Barracks on Saturday
at 10 oClock A.M. — All evidences and persons concerned are req
uested to attend — Captain N. White is appointed Judge Advocate —
An orderly Serjeant from the Rhode Island Regiment to attend the
Court (daily during their Sitting) —
Head Quarters Philadelphia Friday Janry 25. 1782
Parole . Cygnus
Captain Benjamin Walker of the 2d New York Regiment is
appointed Aid de Camp to the Commander in Chief and is to be respected
and obeyed accordingly —

Figure 1. Example of Document Image processing (Precision=0.89, Recall 0.97).

2. System Description

The proposed procedure is presented in Fig.2, while in the following subsections we describe in detail the different stages. Although we make use of the above-mentioned properties (see §1), we attempt to make the procedure as robust as possible. Note also that the same procedure can be repeated for vertical lines.

2.1. Pre-processing

Currently, the only step used in pre-processing is scanning (border) noise removal by the examination of the four edges of the page.

2.2. Left line identification

In this stage, the first column containing a black pixel on the page, according to the vertical histogram, is examined. For each group of black pixels found in a column, the position *(x,y)* of its central pixel is kept, as well as its thickness. The tenth and twentieth columns are also examined for groups with similar thickness, taking into consideration that their positions can vary from the first by several pixels in either direction.

At the end of the procedure, the mean thickness of the groups is considered to be the Thickness of the ruling lines on the page.

2.3. Right line identification

This same procedure is followed (see §2.2), only in this case the final column with black pixels is examined, as well as the tenth and twentieth column before this. Moreover, the Thickness, estimated in §2.2, is assumed at this point, and only groups with Thickness ± 1 are considered, since the ruling lines are considered uniform in thickness (*Property 1*, §1) . However, the position (central pixel of group in the last column) is calculated and kept, as before.

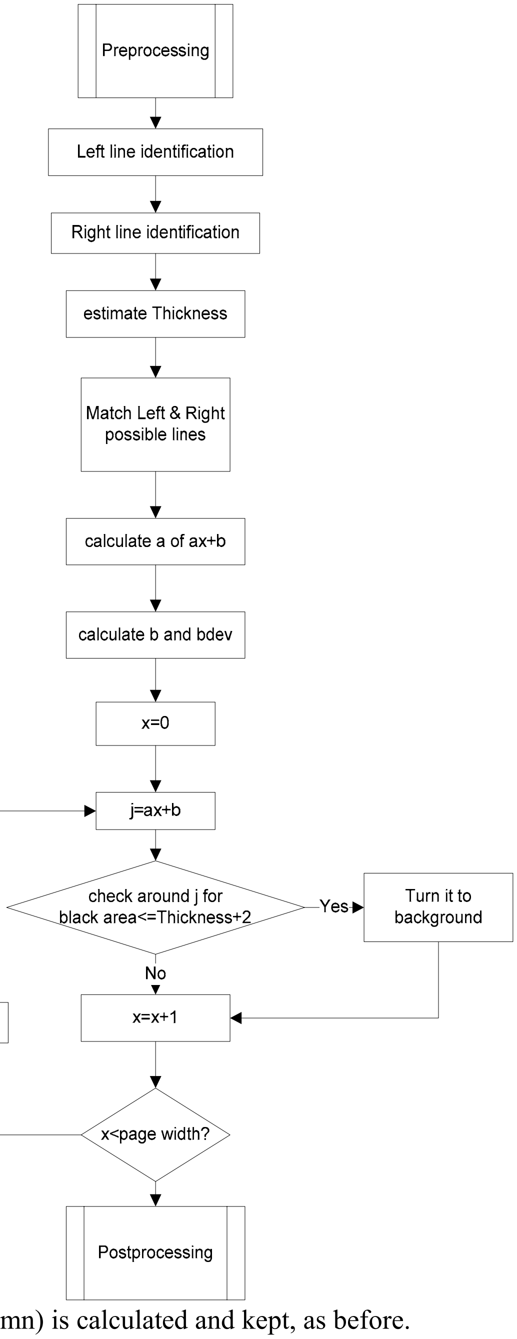


Figure 2. The proposed methodology.

2.4. Match Left & Right lines

Correspondence is established for each pair of left (y_l) and right (y_r) points such that:

$$\left| y_l - y_r \right| < 50 \qquad [1]$$

This difference of 50 pixels permits a small amount of skew and shift for the page and is safe since the distance between ruling lines is usually at least 100 pixels. Such a threshold makes our method robust even for skews of up to 20 degrees or more, although this is not a primary concern here.

2.5. Line removal

The mathematic formula for a line is:

$$y = ax + b \qquad [2]$$

Given two points (x_1,y_1) and (x_2,y_2), a and b can be calculated by the formulas:

$$a = \frac{y_1 - y_2}{x_1 - x_2} \qquad [3]$$

and

$$b = y_1 - ax_1 \qquad [4]$$

Thus, for each pair of corresponding left and right points, an *a* value is calculated from formula [3], and two *b* values from formula [4], one for a left point and one for a right. The difference in *b* is caused by the quantization to pixel values. This difference we call *bdev* and we split it between the columns of the pages.

Thus, scanning the page from left to right for each *x* (column), the *b* has to be adjusted (according to *bdev/page_width*), the estimated *y* is calculated by the formula [2] and quantized to a pixel value *j*. The area *[j-Thickness j+Thickness]* is searched for groups of black pixels, starting from its center *j* and moving towards its limits. If a group of black pixels is found in column x with thickness ≤ Thicknesss+2, it is considered part of a ruling line and changed to background. If the thickness is greater than the threshold, it is considered text and is not changed.

Notice that it is not necessary to find a group of black pixels, as broken ruling lines often arise in scanned pages.

2.6. Post-processing

There is always the possibility that a line is broken and either its left or its right edge, or both, do not exist (*Property 3*, §1). Thus, the potential for having broken parts of lines all over the page must also be considered. However, we have to be careful not to remove useful parts of the text.

At this point, we assume that the ruling lines on a page are likely to be less distinct than the handwritten text (i.e., less thick). Thus the whole page is scanned column by column and groups of pixels with thickness ≤ Thickness are removed.

3. Experimental results

To give objective and comparative results for our proposed approach, we use the evaluation methodology described in [6]. In that paper, synthetic data is employed to provide ground truth for each ruling line. The authors use 5 images of ruling lines and 10 images of Arabic documents, yielding 50 test images. Evaluation is via recall/precision and weighted harmonic mean F1 metrics, defined as:

$$\mathrm{Pr}\,ecision = \frac{tp}{tp + fp} \qquad [5]$$

$$\mathrm{Re}\,call = \frac{tp}{tp + fn} \qquad [6]$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \qquad [7]$$

Where:

- True positive pixel (*tp*): a pixel that exists in both the detection map and the rule line ground truth map, but not in the text only map.
- False positive pixel (*fp*): a pixel that exists in both the detection map and the text only map.
- False negative pixel (*fn*): a pixel that exists in the ground truth map, but in neither the detection map or the text only map.

In our case, 10 scanned page images with ruling lines from different pads were used with 10 images of text from different languages written by different persons (3 English pages, 2 Greek, 2 German, 1 French and 2 Arabic), resulting in 100 images of text with ruling lines.

Table 1 presents experimental results for our system, distinguished by language as we wish to check for language dependence.

Our experimental results show a slightly worse performance in the precision rate for French and German. However, further experiments using more data should probably be performed as the difference is small. We also observe that the missed handwriting (recall) mostly involves the endings of characters where writers lift the pen, resulting in thinner stroke widths. OCR is usually not affected by this.

Table 1. Experimental Results.

	English *30 images*	*Greek* *20 images*	*German* *20 images*	*French* *10 images*	*Arabic* *20 images*	*Total* *100 images*
Precision	0.81%	0.88%	0.68%	0.65%	0.96%	0.76%
Recall	0.93%	0.93%	0.90%	0.93%	0.90%	0.91%
F1	0.86%	0.89%	0.75%	0.75%	0.93%	0.81%

4. Discussion

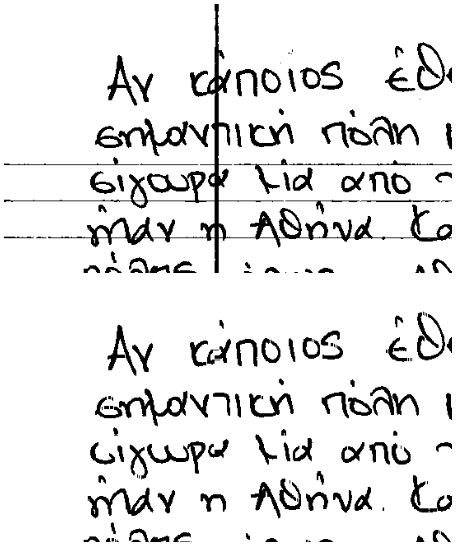


Figure 3. Detail of a Greek page, before and after its processing by the system.

In this paper, a method for ruling line removal has been presented. The proposed system takes advantage of standard properties of ruling lines, but at the same time is robust and less complicated than previous techniques. It does not break existing characters significantly and hence does not require an additional restoration step.

We reported an experimental evaluation using 100 synthetic pages, formed by 10 pages of ruling lines scanned from different pads in combination with 10 text images in 5 different languages.

In Fig. 3, we show detail of a Greek document demonstrating how well our system removes ruling lines. On the other hand, Fig.4 displays an Arabic document where the horizontal ruling lines are removed, but part of the scanned vertical line remains.

Our results are competitive with published methods, but with an approach that appears less complicated and is shown to work for 5 different languages.

In our future work, we plan further studies regarding language-dependence, as well as detailed experiments on skewed documents. Moreover we hope to address cases like the one depicted in Fig.4.

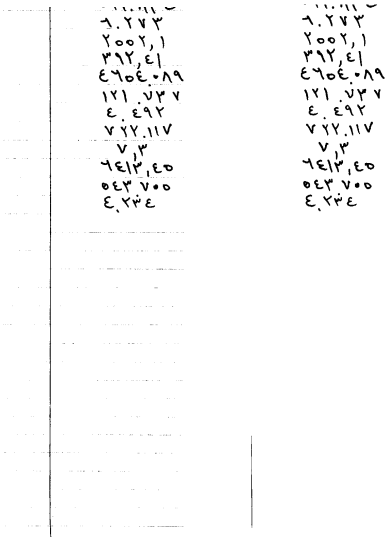


Figure 4. Part of Arabian document, before and after its processing by the system.

References

[1] K. Tombre, S. Tabbone, L. Plissier, and B. Lamiroy. “Text/graphics separation revisited.” Workshop on Document Analysis Systems, pp. 200–211, 2002.

[2] F. Cesarini, M. Gori, S. Marinai, and G. Soda. “Informys: A flexible invoice-like form-reader system” *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(7):730–745, 1998.

[3] Huaigu Cao, Venu Govindaraju, "Preprocessing of Low-Quality Handwritten Documents Using Markov Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1184-1194, July, 2008.

[4] P. Dosch, K. Tombre, C. Ah Soon, and G. Masini. “A complete system for the analysis of architectural drawings” *IJDAR*, 3(2):102–116, 2000.

[5] K.R. Arvind, J. Kumar and A.G. Ramakrishnan, “Line Removal and Restoration of Handwritten Strokes,” *International Conference on Computational Intelligence and Multimedia Applications*, vol. 3, pp. 208-214, 2007.

[6] Wael Abd-Almageed, Jayant Kumar, David Doermann, "Page Rule-Line Removal Using Linear Subspaces in Monochromatic Handwritten Arabic Documents," 10th International Conference on Document Analysis and Recognition, pp. 768-772, 2009.