Document Image De-warping Based on Detection of Distorted Text Lines

Lothar Mischke[1] and Wolfram Luther[2]

[1] Eduard Spranger Vocational School,
D-59067 Hamm, Vorheider Weg 8, Germany
lothar.mischke@esb-hamm.de
[2] Institute of Computer Science and Interactive Systems,
University of Duisburg–Essen,D-47048 Duisburg, Lotharstr. 65, Germany
luther@informatik.uni-duisburg.de

Abstract. Image warping caused by scanning, photocopying or photographing a document is a common problem in the field of document processing and understanding. Distortion within the text documents impairs OCRability and thus strongly decreases the usability of the results. This is one of the major obstacles for automating the process of digitizing printed documents.

In this paper we present a novel algorithm which is able to correct document image warping based on the detection of distorted text lines. The proposed solution is used in a recent project of digitizing old, poor quality manuscripts. The algorithm is compared to other published approaches. Experiments with various document samples and the resulting improvements of the text recognition rate achieved by a commercial OCR engine are also presented.

1 Introduction

In the context of a digitization project initiated in 2001 [2], for the German reception of the philosopher Nietzsche at the University of Duisburg–Essen, text documents composed between 1865 and 1945 have been digitally converted. The conversion process has been shown to be both complex and time-consuming. Many of the documents have been printed using one of the German fraktur typefaces which are not recognized very well by today's OCR programs and therefore have to be corrected manually. Due to the poor quality of some of the literary sources, which consist predominantly of photocopies of the original documents, the process of digitizing requires further human interaction. Shade, skew and warping of the document images markedly decrease OCR recognition accuracy.

To optimize the process of selecting and capturing the text documents [4] in the context of the funded project *eCampus Duisburg* a sub-project entitled "Distributed mobile selection and evaluation of documents in libraries and archives" was initiated in 2002 [3]. Mobile users digitize documents which are transmitted to a document server via WLAN. A team of experts accesses the documents via

F. Roli and S. Vitulano (Eds.): ICIAP 2005, LNCS 3617, pp. 1068–1075, 2005.

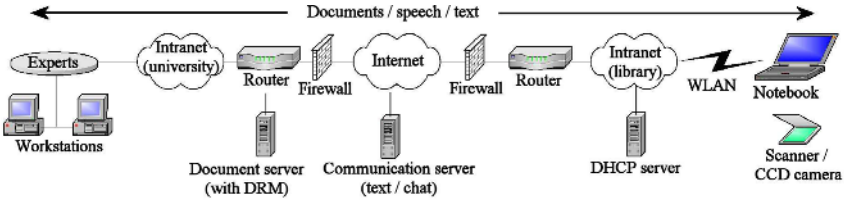


Fig. 1. Distributed document selection and evaluation (network view)

Internet, coordinates the selection, and decides about further proceedings. The network view of this process is shown in Fig. 1.

In this paper we address the preprocessing for successful OCR recognition in the case of warped document images. A brief discussion of recently published approaches to image de-warping can be found in section 2. In section 3 we present a new algorithm which is able to correct warping based on the detection of distorted text lines. Our method is capable of removing shade, correcting global skew and de-warping distorted text blocks within the document. Experiments with various document samples and the resulting improvements of the recognition rate received by a commercial OCR engine are presented in section 4.

2 Recent Approaches to Document De-warping

In the last few years several restoration methods for warped document images have been reported. These algorithms can roughly be divided into two categories:

- algorithms which make use of information derived from the source of the document distortion, and
- algorithms that simply detect the (horizontal) distortion by means of an analysis of the given document image.

The advantage of the former group of algorithms lies in the fact that knowledge of the kind of degradation within the document image can be used to model the geometrical type of distortion very well. In [5] Cao et. al. propose a cylindrical model for the restoration of camera-captured documents. Their correction algorithm is restricted to cases where the generatrix of the cylindrically assumed book surface parallels the image plane. Zhang et. al. suggest in [11] an algorithm which is capable of de-warping documents scanned from thick, bound volumes using an image scanner provided that the book spine lies parallel to the scanning light. Thus, today's model-based algorithms have the drawback that their usability is very limited as they need a lot of a priori knowledge. Currently there is no generic model which can be used for identifying and rectifying automatically all common types of warped document images.

Algorithms of the second kind do not require explicit knowledge regarding the source of distortion. Wu and Agam [9] developed a method which detects and traces curved text lines within single document images by minimizing the *local* cumulative projection within a given range of angles. The algorithm starts at the

left-hand border of a given region (which is assumed to be approximately vertical and must be provided manually) and gradually traces the curved lines. These lines are used to reconstruct a target mesh which can be used for de-warping. In [9] they apply their algorithm to perspectively degraded documents which have been captured using a digital camera. They do not impose any constraints regarding the angle between document and image plane of the camera. On the other hand results presented using this approach still show perspective distortion of the characters within the warped regions of the documents.

In [10] Zhang and Tan propose an algorithm which detects distortion by distinguishing between the light and the shaded area of a scanned gray-level document image from a bound volume. As the warped part of the image resides in the shaded area the alignment of the connected components which form a curved text line is approximated by two quadratic functions. Thus the line parts within the non-shaded area can be bound by straight reference lines. The relative position of the connected components within the shaded area with the two curves is used to move the components vertically to the corresponding straight reference lines within the non-shaded area. The orientation of the moved components is then corrected using the average angle of the tangent of the two reference curves.

Both the model-based and the analyzing type of algorithms have been shown to be suitable for increasing OCR accuracy. Although results presented indicate that the remaining distortion in the rectified documents is higher with algorithms of the second type we regard that their broader applicability is a distinct advantage. However, since within our project the original document sources are often not available and the process of digitizing and processing the documents is distributed between multiple individuals and locations (see Fig. 1), we chose to develop a method which does not rely either on explicit knowledge about the digitizing technique used (photographing, scanning, etc.) or on assumptions regarding the quality of the document image.

3 New De-warping Algorithm

The presented algorithm consists of three preprocessing steps and the document rectification step itself which are discussed in the following subsections. Fig. 2 briefly illustrates these four building blocks.

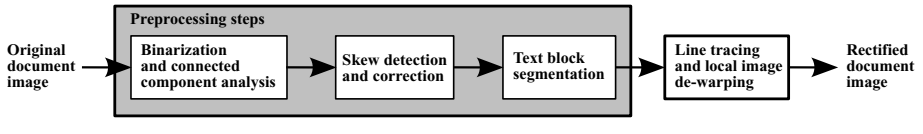


Fig. 2. Block diagram of the algorithm

3.1 Preprocessing Steps

Since most of the algorithms used to process the document image were designed for binary images we have to separate the background from the data. Warped

images often result from flatbed-scanning bound volumes. This leads additionally to a shaded region around the book spine. In this case global thresholding is ineffective. Savakis suggests in [8] an adaptive thresholding method where a document is binarized by moving an M-size window over the image, clustering all inner pixels into a foreground and a background cluster using a global threshold, and determining a local threshold for the window's central pixel by averaging these two clusters. Unfortunately this technique fails in the shaded regions of the document where the background level already exceeds the global threshold.

Therefore we divide the document image into a set of rectangular regions R_{ij}. For each R_{ij} we calculate a threshold $t_{c_{ij}}$ using Otsu's method [7] which we assign to the centroid c_{ij} of the region. Then a *localized* threshold t_{xy} for each pixel (x, y) obtained by a linear interpolation of the thresholds of the surrounding region centroids can be used for adaptive foreground/background clustering. This approach is justified by the fact that the background level decreases continuously with increasing distance to the spine area.

After binarizing the document the *connected components* must be determined. All eight-connected black pixels are grouped together in rectangular bounding boxes each of which identifies a single connected component. We use a modified version of the classical approach of Fletcher and Kastouri which is described in detail in [6] in order to separate text from graphics and to remove noise. Larger components are removed by an area/ratio filter. The remaining components are kept for subsequent use.

Skew is another standard problem in document processing and has a detrimental effect on the analysis of documents. A good survey on this topic can be found in [1]. Classical approaches make use of the Hough Transform or projection profiles in order to detect straight lines perpendicular to an accumulator line at different projection angles. As the text lines in a warped image are (partly) distorted the quality of these methods degrades. We use a skew correction method basing on projections of the centroids of the detected connected components provided that within the warped image there are still parts of the text lines which are approximately straight. Typical warping caused by scanning or photographing satisfies this constraint (see Fig. 5). This allows us in contrast to Zhang's and Tan's method [10] or the approach of Wu and Agam [9] to process document images which are *both* skewed and warped without further user interaction. The skew correction will not solve the distortion problem itself. But it aligns the nearly straight parts of the text lines parallel to the horizontal document bounds which is a precondition for the line tracing algorithm described in subsection 3.2.

After having separated text from graphics the *document structure* must be analyzed. The document is divided into text blocks which can be processed individually. For this task we use a pyramid algorithm which constructs a pyramid of less detailed images by successive reduction of the dimensions by half at each step. This leads to a natural clustering of the eight-connected components. The resulting connected components represent text blocks or isolated words which can be grouped to form text blocks.

3.2 Line Tracing and De-warping

In order to detect warped lines we identify the location of possible text lines and trace their run to the boundaries of the surrounding text block. The granularity parameter $g = 0.2$ and the angular range parameter $\theta = 15$ degrees which are used in the following steps are resolution and font independent and have been derived from experiments with multiple training documents containing varying types of distortion. Due to the preliminary skew correction we can assume that the non-distorted (straight) line parts approximately parallel the text block's horizontal borders. Thus we calculate a modified projection profile of the centroids of the connected components' bounding boxes. Let H_{avg} denote the average height of components in the text block. With text block height h we scale down the size of the accumulator to $h/(g \cdot H_{\text{avg}})$. For each bounding box a weighted vote v is calculated and added to the appropriate accumulator slot according to $v = w_j \cdot \max(0, 1 - |(h_j - H_{\text{avg}})/H_{\text{avg}}|)$, where w_j and h_j denote width and height of the bounding box, respectively. This is motivated by the observation that the line run can best be detected using average-sized bounding boxes which mostly contain the characters that lie *between* the text baselines. Notice that no further information (like manually provided vertical border lines or the location of shaded regions) is required. Fig. 3 illustrates this approach.

After processing each component the local maxima of the accumulator indicate the vertical position of the straight text line parts. As the warped parts typically lie next to the document borders we choose the horizontal middle of the text block to start the line trace. Hence for each line let (x_s, y_s) denote the starting point which we call *seed*. Then the line is built by creating an empty bounding box of height H_{avg} around (x_s, y_s) and extending it to the left and right by gradually adding all adjacent components provided that their height lies between $(1-g) \cdot H_{\text{avg}}$ and $(1+g) \cdot H_{\text{avg}}$ and y_s lies within the vertical extension of the component's bounding box. Let $(x_i, y_i), 1 \leq i \leq n$, denote the n recognized character box centroids sorted from left to right. The core zone of the text line is then set to a rectangle of width $w = r_n - l_1$ and height H_{avg}. l_1 and r_n denote the left and right bound of the leftmost and rightmost connected component, respectively. The rectangle is centered around $(x_{s'}, y_{s'})$ where $x_{s'} = (l_1 + r_n)/2$ and $y_{s'} = \text{median}_{1 \leq i \leq n}(y_i)$. The detected lines are now extended to the left and to the right by further adding adjacent components partly lying within an angular range of $\pm\theta$ degrees around the current text line orientation.

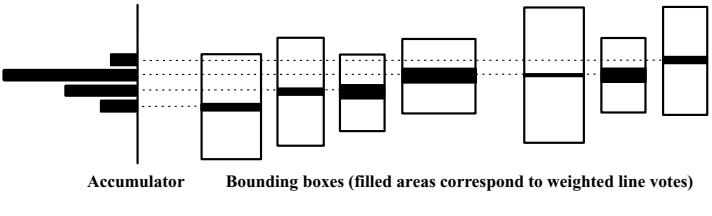
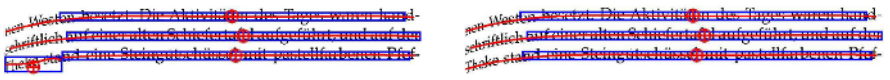


Fig. 3. Text line detection



(a) (b)

Fig. 4. Seed, core zone and approximated run of warped text lines before (a) and after (b) broken line correction

In regions of extreme line deformation it is possible that curved lines are broken by the detection algorithm (see Fig. 4 (a)). Therefore possibly broken line parts are melted provided that the extrapolated line run of at least one part hits the border component's bounding box of the adjacent line part.

The run of each distorted text line is then polynomially approximated. A k-th degree polynomial $y = a_0 + a_1 x + \cdots + a_k x^k$ can be determined using *least squares fitting*. Setting the partial derivatives $\partial E/\partial a_i$ of the residual $E = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} a_j x_i^j \right)^2$ to 0 leads to the following matrix equation which can be solved numerically:

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i & \cdots & \sum_{i=1}^{n} x_i^k \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \cdots & \sum_{i=1}^{n} x_i^{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_i^k & \sum_{i=1}^{n} x_i^{k+1} & \cdots & \sum_{i=1}^{n} x_i^{2k} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \vdots \\ \sum_{i=1}^{n} x_i^k y_i \end{pmatrix} \tag{1}$$

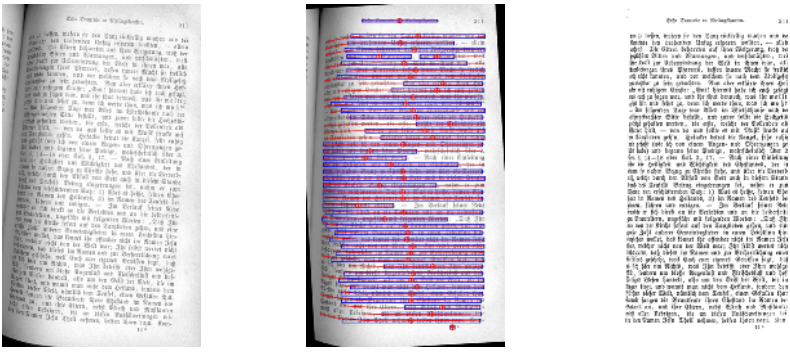
Note that we choose $k = \min(4, n-2)$ depending on how many centers have been localized.

The polynomial approximations of the text lines are now used to construct a dense source mesh. As the approximation is not reliable near the margins of short lines these grid values are substituted by averaging the run of surrounding text lines. The vertical grid lines are created by subdividing the lines into small segments of about H_{avg} width.

The target mesh can be assumed to be rectangular. Its vertical position is determined by the vertical position of the text block's non-distorted core zone which can be calculated from the detected text lines' core zones. Thus a rectified target (sub)image can now be created by applying any 2D-warping algorithm. For our tests we used a classical bilinear transformation.

4 OCR Results Comparison

We run our OCR tests twice before and after image binarization and restoration using the OCR software ABBYY FineReader 7.0. Like Zhang and Tan in [10] we use *precision* and *recall* as a measure for comparison of OCR improvement. Precision is hereby set to the quotient of the number of characters correctly detected and the total number of characters detected whereas recall denotes the quotient of the number of characters correctly detected and the total number of



(a) (b) (c)

Fig. 5. Skewed distorted (a), de-skewed (b), and binarized and rectified (c) sample fraktur document image

Table 1. OCR improvement (precision and recall)

	Original document		Bin. and rectified document	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
Ave. of 50 (300 dpi)	92.81	88.99	97.53	96.29
Ave. of 50 (150 dpi)	87.61	74.41	88.81	80.70
Ave. of 10 skewed (300 dpi)	32.77	1.62	98.03	94.94

characters in the document. It was not possible to compare the OCR improvements to the results presented in [10] because the text base was not specified. Wu and Agam did not provide any OCR test results in [9]. Thus, we chose a set of 50 different sample documents containing a wide variety of different font sizes, typefaces and layouts. These documents were scanned from bound volumes at resolution 150 and 300 dpi, respectively. The orientation of the documents was varied in order to create different types of distortion. Note that at 150 dpi resolution OCR accuracy is higher for grayscale than for binary images which explains the low increase in precision for the rectified documents. Additionally we experimented with a set of 10 severely skewed warped documents (5 to 40 degrees). Without manual correction the OCR program detected only 6.41 percent of the documents' content. The results of our tests are presented in Table 1.

5 Conclusion and Future Development

A novel algorithm for image warping detection and correction is presented. The advantage of this approach is that few assumptions are required relating to the type of distortion that can be processed. The algorithm was implemented to be used in a recent digitization project and can be used without human interaction. Our tests show that processing distorted documents with this algorithm can

significantly improve the recognition accuracy of a downstream OCR engine. Future research will focus on the improvement of the correction of perspective deformation which may have multiply causes, for example variation of distance between the document and the scanner or camera.

Acknowledgements

The project *eCampus Duisburg* was supported by the German Federal Ministry of Education and Research (BMBF).

References

[1] Amin, A., Fischer, S., Parkinson, A. F., Shiu, R., *Comparative Study of Skew Detection Algorithms*, Jour. of Electronic Imaging SPIE, USA, 1996, pp. 443-451

[2] Biella, D., Dyllong, E., Kaiser, H., Luther, W., Mittmann, Th., *Edition électronique de la réception de Nietzsche des années 1865 à 1945*, Proc. ICHIM03, Paris, France, Sept. 2003

[3] Biella, D., Luther, W., *Mobile verteilte Dokumentenrecherche in Bibliotheken und Archiven*, In: INFORMATIK 2003 - Innovative Informatikanwendungen, Vol. 1, GI 2003, Germany, pp. 298-302

[4] Biella, D., Luther, W., Pilz, Th., *A web-based System for Assisted Literature Research*, In: Proceedings of the 3rd European Conference on e-Learning, ECEL 2004, Nov. 2004, Paris, France, pp. 15-24

[5] Cao, H., Ding, X., and Liu, C., *A Cylindrical Surface Model to Rectify the Bound Document Image*, Ninth IEEE ICCV 2003 Vol. 1, Nice, France, Oct. 2003, pp. 228-233

[6] Fletcher, L. A., Kasturi, R., *A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images*, IEEE Trans. Pattern Anal. Mach. Intell. 10(6), 1988, pp. 910-918

[7] Otsu, N., *A Threshold Selection Method from Graylevel Histograms*, IEEE Trans. Sys. Man Cybern. 9(1), 1979, pp. 62-66

[8] Savakis, A. E., *Adaptive Document Image Thresholding Using Foreground and Background Clustering*, Proc. of ICIP 1998, 1998, pp. 785-789

[9] Wu, C., Agam, G., *Document Image De-Warping for Text/Graphics Recognition*, Proc. of Joint IAPR 2002 and SPR 2002, Windsor, Ontario, Canada, Aug. 2002, pp. 348-357

[10] Zhang, Z., Tan, C. L., *Correcting Document Image Warping Based on Regression of Curved Text Lines*, ICDAR 2003, Aug. 2003, Edinburgh, UK, pp. 589-593

[11] Zhang, Z., Tan, C. L., Fan, L., *Estimation of 3D Shape of Warped Document Surface for Image Restoration*, ICPR 2004, Aug. 2004, Cambridge, UK, pp. 486-489