Restoration of Arbitrarily Warped Historical Document Images Using Flow Lines[†]

Maryam Rahnemoonfar and Apostolos Antonacopoulos
Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Computing, Science and Engineering, University of Salford, United Kingdom
E-mail: rahnemoonfar@ieee.org, a.antonacopoulos@primaresearch.org

Abstract —Historical documents frequently suffer from arbitrary geometric distortions (warping and folds) due to storage conditions, use and to, some extent, the printing process of the time. In addition, page curl can be prominent due to the scanning technique used. Such distortions adversely affect OCR and print-on-demand quality. Previous approaches to geometric restoration either focus only on the correction of page curl or require supplementary information obtained by additional scanning hardware – not practical for existing scans. This paper presents a new approach to detect and restore arbitrary warping and folds, in addition to page curl. Warped text lines and the smooth deformation between them are precisely modelled as primary and secondary flow lines that are then restored to their original linear shape. Preliminary, but representative, experimental results, in comparison to a leading page curl removal method and an industry-standard commercial system, demonstrate the effectiveness of the proposed method.

Keywords-arbitrary warping; geometric correction; de-warping; page curl removal; text line modelling; flow lines

I. INTRODUCTION

A challenging problem in digitising historical documents is the presence of geometrical distortions that may be introduced at various stages during the life cycle of a document, from when it was first printed to the time it is digitised by an imaging device. Such distortions include arbitrary warping resulting from the printing process or storage conditions (see Fig. 1), folds resulting from use, and page curl resulting from scanning tightly bound books. Such geometric distortions, depending on their severity, can have detrimental effects to recognition (OCR) and readability e.g. for print-on-demand.

Several restoration techniques have been proposed for geometric artefacts, in general. Those can be broadly classified into two categories: restoration approaches for the correction of page curl [1-7] and those approaches that address arbitrary warping [8-13].

Methods in the first category either make some assumptions in modelling the page curl or require specialised scanning hardware. The latter can be cumbersome and not cost-effective for large-scale digitisation and is certainly not applicable to the millions of documents already digitised. For instance, Cao *et al.* [1] used a cylindrical surface to model the shape of the document image and used the skeleton of horizontal text lines in the image to estimate the model parameters. In the method of Zhang and Tan[2],page curl is corrected by dividing the page image to shade and non-shade regions. The alignment of text is modelled by straight reference lines in the non-shade area, and polynomial regression is used to model the warped text lines in shade area. In general, methods that use polynomial curve fitting [2][3][4][5]to represent text lines can only model relatively smooth variations of the direction of baselines. This is sufficient for page curl but will fail in documents with severe arbitrary warping or folds. Finally, methods using active contours [6][7] based on energy minimisation algorithms do not recognise descenders, and cannot distinguish between descenders and non-descender characters which belong to the baseline and should be preserved.

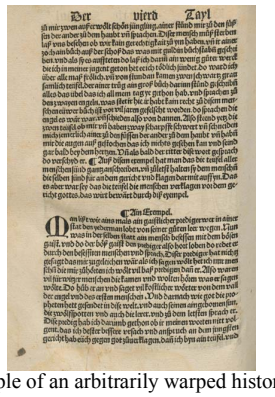


Figure 1. Example of an arbitrarily warped historical document.

Methods in the second category, such as those that use specialised hardware like laser projector [8],structured light 3D acquisition [9] or two-camera stereo vision [10] seem to be able to handle arbitrary geometric distortions, albeit only those resulting from warping of the paper (not from inexact printing methods). Other methods [11]in the second category obtain information directly from the images. However, they need paper pattern under image material. The general method idea proposed by Gatos *et al.*[12] and Stamatopoulos *et al.*[13] seems to be able to handle moderate arbitrary warping in historical documents, although as it is shown in experimental results it does not have enough accuracy as it does not model precisely enough the baselines or the space between them.

This paper presents a precise and flexible dewarping system, which, by using only the information in a single image, and without any assumptions about the type of distortions in the image, such as smooth curl, for example, will produce a simulated flat image of the original document.

[†]*This work has been partly supported through the EU 7th Framework Programme grant IMPACT (Ref: 215064).*

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.184

The remainder of this paper is structured as follows: In Section 2, the proposed technique is described in detail. Experimental results are presented and discussed in Section 3. Finally, conclusions are drawn in Section 4.

II. THE PROPOSED METHOD

In order to restore geometric distortions in historical documents including page curl, arbitrary warping and folds, the proposed method is based on (i) text line segmentation, (ii) precise baseline modelling, and (iii) dewarping by means of transforming primary and secondary flow lines.

The proposed method does not require any external information but instead works purely on the image data. The most apparent representation of deformation on a page is the (non) linearity of text lines. Therefore detecting text lines and precise modelling of their baselines plays an important role.

A. Text Line Detection

In this step, connected components are first detected in a bitonal image and analysed to remove noise and (to the extent possible) restore broken characters. Very large and very small components (remaining noise, dots, punctuation) are removed by examining the histogram of component height. Then the bottom-centre points of all remaining bounding boxes, including those of characters with descenders, are identified as potential points forming a line. Starting with potential points, the left and right neighbours of each potential point are detected and are connected based on the connecting criteria of minimum weighted L^2norm and projected overlap. Connecting criteria such as nearest-neighbour examine the neighbours within a small radius from a given point while our proposed weighted L^2 norm searches neighbours in an ellipsoidal space and in addition it considers the maximum overlap of the components. When there are text lines with high slope such as in the case of pronounced page curl or the existence of folds, the projected overlap feature helps to segment text lines while the minimum weighted L^2 norm is the key feature in segmenting very dense paragraphs with touching components.

B. Precise Baseline detection

In large-scale digitisation of historical documents due to the arbitrary (and possibly abrupt) nature of distortions, it is not always practical to model baselines (especially in the presence of folds) with a smooth curve (e.g. spline, polynomial curve fitting.

In the proposed method, having the segmented text lines from the previous step, the precise baselines are now detected. The procedure consists in correcting the simple piecewise-linear curve connecting connected-component points by ensuring that descenders do not distort the desired baseline.

After the text lines have been identified, each baseline consists of n segments (straight lines connecting points), where $Z_i = (x_i, y_i)$ are the coordinates of the beginning position of each segment. Each line on the page can be represented by a piecewise-linear function *f(t)* as follows

$$f : (\frac{i}{n}, \frac{i+1}{n}) \longrightarrow (Z_i, Z_{i+1}), \qquad i = 0 \text{ to } n$$

$$f(t) = (1 - (nt - [nt]))Z_{[nt]} \quad (nt \quad [nt])Z_{[nt+1]} \tag{1}$$

where n is the number of partitions and [ ] represents the well known *floor* function. The angle between two segments in the piecewise linear curve can be written as:

$$\theta = \arccos(-\frac{1 + f'^{-} f'^{+}}{\sqrt{1 + (f'^{-})^2}\sqrt{1 + (f'^{+})^2}}) \tag{2}$$

where f'^{+} and f'^{-} are right and left derivatives at each point respectively.

To calculate the desired angles, upward looking angles, the concept of convexity and concavity of a curve is used. Function f is convex if $f'' > 0$ and concave if $f'' < 0$, where f'' represents the second derivative of f. The desired angles are calculated according to Formula 3.

$$\phi = \begin{cases} \theta & if\ f'' \le 0 \\ 2\pi - \theta & if\ f'' > 0 \end{cases} \tag{3}$$

Based on the desired angle for each point, the new coordinate point is calculated as following (equation 4).

$$y_{i,j} = \quad _{i,j} . y_{i,j} + (1 \quad _{i,j}) - \quad \times \frac{(x_{i+1,j} - x_{i,j}) y_{i-1,j} + (x_{i,j} \quad x_{i-1,j}^{-}) y_{i \quad 1,j}}{(x_{i+1,j} - x_{i-1,j})} \tag{4}$$

with

$$_{i,j} = \text{sgn}(1 - \quad ^{\prime}_{i,j})$$

$$^{\prime}_{i,j} = \frac{1 - \text{sgn}(\phi_{i,j} - \pi \quad \varepsilon)}{2} \times + \frac{1 + \max(\text{sgn}(\phi_{i+1,j} - \pi - \varepsilon), \text{sgn}(\phi_{i-1,j} - \pi - \varepsilon))}{2} \tag{5}$$

where $\phi_{i,j}$ is the angle of i^{th} point in j^{th} line calculated according to formula 3 and sgn is the sign function.

The first term in equation (5) considers the angle of each point and evaluates if the point has the potential to be a descender, while the second term evaluates the right neighbour's angle $\phi_{i+1,j}$ and the left neighbour's angle $\phi_{i-1,j}$ to assess if the convexity in each point is due to a descender or occurs due to geometric distortion at the point by considering the fact that descenders tend to produce shorter wavelength, compared to the result of geometric distortions. If the evaluated point is a descender, would be 0 and therefore the new *y*-coordinate of the point is calculated according to the second term in formula (4) in which $X_{i+1,j} = \begin{pmatrix} x_{i+1,j} \\ y_{i+1,j} \end{pmatrix}$ is the

right neighbour and $X_{i-1,j} = \begin{pmatrix} x_{i-1,j} \\ y_{i-1,j} \end{pmatrix}$ is the left neighbour of the point in line j. If the point is not a descender, $_{i,j}$ would be 1 and the y-coordinate of the current point is preserved. The parameter ε is an experimentally established threshold, which provides enough flexibility to account for small variations due to special typefaces used in historical documents, noise and binarisation degradations and also permits characters to sit slightly below the baseline. In this experiment ε is set to 0.035. The term *max* in formula (5) indicates that if at least one neighbour of a point is concave, then the point corresponds to a descender, providing enough flexibility for two descenders to be located next to each other.

To detect the baseline precisely when more than *2n* descenders are consecutively located, the above algorithm will be repeated *n* times.

A similar procedure is performed for descenders located at the end of the line (and for noisy ascenders, which are out of the scope of this paper). Fig. 2 shows example results of the proposed method in modelling the precise baselines in the case of pronounced page curl, arbitrary warping and a fold.

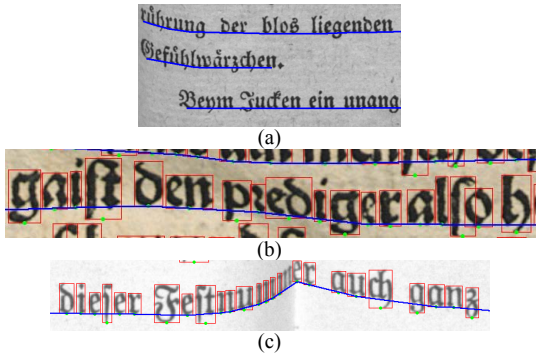


Figure 2. Precise baseline modeling by the proposed method for (a) pronounced page curl, (b) arbitrary warping and (c) fold

C. Dewarping

Based on the precise baseline detection proposed in the previous section, a new algorithm for dewarping of various historical documents is proposed here using both local and global features of the image.

In this approach, first a velocity vector field is constructed based on the baselines detected. Before constructing the velocity vector field, very short lines are excluded from the computation because they are not reliable. The main idea of the dewarping method is the construction of a primary and secondary velocity vector field. It must be flexible to describe the variety of different deformations, but simple enough to be computed speedily. Let $f(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$ be a baseline function obtained from equation (1) after updated with new coordinate points; $f'_+(t_k)$ is the first right derivative of $f(t)$ at point k which is velocity vector of $f(t)$ at that point. The velocity vector field $V : \Re^2 \longrightarrow \Re^2$ is constructed according to the following equation:

$$V(f(t)) = \frac{f'_+(t)}{|x'_+(t)|} \tag{6}$$

where $x'(t)$ is the first right derivative of $x(t)$. Vector field V maps each point k in the baseline in the image coordinate system to a vector $V(f(t_k))$. Each point on the primary flow line of the vector field can be estimated by the recursion

$$Z_{k+1} = Z_k + \delta V(Z_k) \tag{7}$$

where Z_k is the k^{th} point on the estimated flow line and δ is the distance step length, defining the accuracy of the approximation.

To calculate the secondary flow lines more accurately, first the primary flow lines are aligned to the left and right margins based on the flow lines below and above. To achieve this, the flow lines corresponding to the short baselines such as titles, lines at the end of the paragraph, etc. must be elongated to reach the same length as their neighbour lines. The shape of each flow line is determined based on the line above and below. Therefore, if the current line is closer to the upper line than the lower line, the upper line has the higher influence on determining the shape of the current flow rather than the lower line. Having detected the primary flow lines, the secondary flow lines will be calculated based on the primary flow lines.

Because of the arbitrary nature of the geometric deformations on a page, the shape of text lines is not the same on the whole page and even two subsequent lines are not necessarily parallel. With this approach, each line is assumed to be the deformation of the subsequent line and the shape deformation between two lines is modelled with a divergence-free velocity vector field. In this way, the amount of flow between two baselines is preserved in the whole page and primary flow lines deform smoothly during the time. To construct the smooth variation between two boundaries, two baselines in this case, the secondary flow line between two primary flows is defined as the following:

$$Z_{j,k+1} = Z_{j,k} + \delta(tV(Z_{i,k}) + (1 \quad t)V(Z_{i+1,k})) \tag{8}$$

where $Z_{j,k}$ is the k^{th} point on the j^{th} line, δ is the length step and t is the time step ranging from 0 to 1, denoting the ratio of the distance between the primary flow line i and secondary flow line j (to be determined) and the distance between two primary lines i and $i+1$. Since each line has an arbitrary shape, the distance between them is not constant on the whole page and therefore the parameter t is not constant and varies during length step δ (Fig. 3). According to Equation 8, the secondary flow lines are deformed smoothly from the primary flow line i to the primary flow line $i+1$ during time t. At the time 0 the secondary flow line corresponds to the first primary

flow line and at the time 1 it corresponds to the second primary flow line while between two boundaries it is the mixture of two primary flow lines. It is worth mentioning that one of the advantages of defining the primary and secondary flow lines using this technique is that it compensates any possible problem caused by the segmentation step described in section II.A; therefore the segmentation result does not need to be perfect. In other words, the primary flow lines help to complete a line if it is not detected or segmented completely, while the secondary flow lines construct the lines that are not detected at all in the segmentation step.

Having modelled the precise flow lines with the above technique, the corresponding flat (dewarped) lines are calculated from the local affine transformation of the secondary flow line. The transformation consists of the rotation of each vector of the flow line in piecewise manner and the translation of each vector so that all transformed vectors are located at the same level.

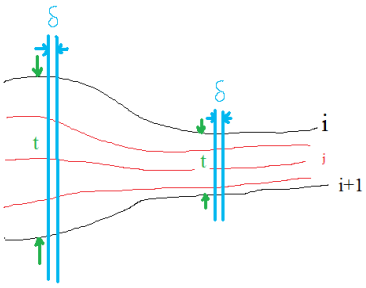


Figure 3.Primary and secondary flow lines

III. EXPERIMENTAL RESULTS

The proposed method has been evaluated on a dataset of 52 representative historical document images from major European libraries participating in the IMPACT project [14], including documents with page curl, arbitrary warping and folds. To facilitate comparison of methods without the requirement of a preceding complex segmentation step, pages containing mostly text in a single column were scanned in colour or greyscale at 300dpi. The proposed approach is compared with a leading dewarping method, albeit primarily designed for page curl removal [12][13] (NCSR) and the Book Restorer commercially available software [15]. Example results of correcting page curl, arbitrary warping and folds are shown in Fig. 4 to 7. To assist visual inspection of the distorted and corrected text lines, a grid is superimposed on the images.

As an evaluation metric, the deviation of each resulting line from a straight line was measured (area of difference between lines) and this value was compared with that of the same metric measured on the original image. The results in the indicative example of Fig. 4 demonstrate that the proposed method performs better than the two other methods. The evaluation results for this example show an average accuracy of 91.08% for the proposed method. The NCSR method and Book Restorer reached an average accuracy of 45.69% and 50.72% respectively. Fig 5 shows the result of the proposed method on a document image with arbitrary warping. Fig 6 shows the same result for an enlarged part of the image of Fig 5 and the corresponding results of the NCSR method and Book Restorer. It can be observed from Fig. 6 that the proposed method performs better than both Book-Restorer and the NCSR method in removing the distortion effectively. The evaluation results for the whole page are 92%, 29%, 16% for the proposed method, NCSR method and Book Restorer, respectively. Figure 7 shows the result of the proposed method on a page with a fold. The NCSR method was not designed to handle abrupt distortions such as folds but Book Restorer has accuracy of 42% while the proposed method has accuracy of 93%.

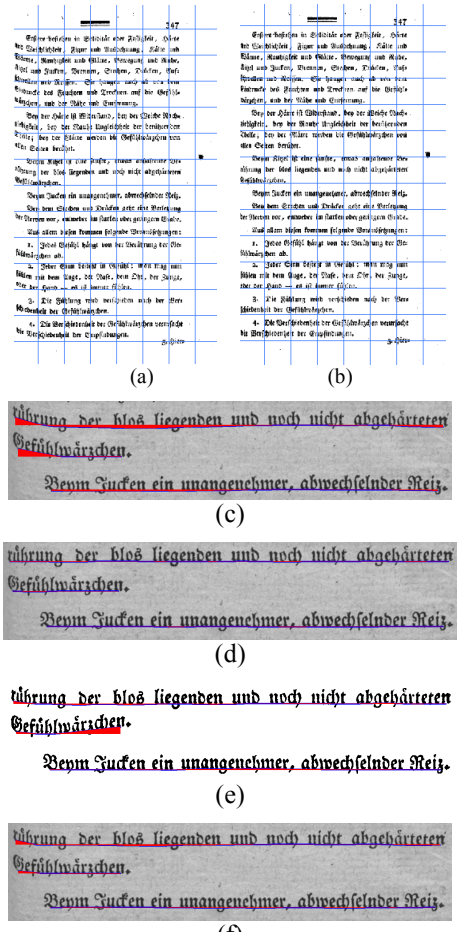
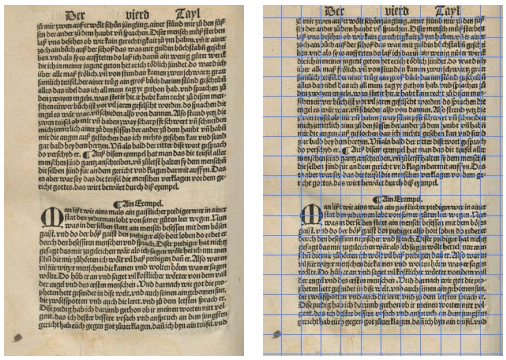


Figure 4. (a) Original image, (b) dewarped image by the proposed method, enlarged part of (c) original image (d) proposed method (e) NCSR method (f) Book-Restorer and their deviations fom the straight line

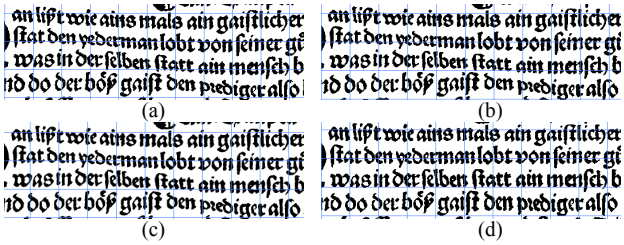
Figure 8 shows the results of the evaluation performed for all 52 images for the proposed method, the NCSR method and Book Restorer. Since the NCSR method is not applicable on document images with folds, the box plot for that method is provided only with 47 images. Overall,

the proposed method yields an average accuracy of 93.94%, which is a very promising result. The notches in the box plot represent the expected range of median. Since the notches of the NCSR method and Book Restorer overlap, it can be concluded that neither of them is statistically significantly better than the other algorithm. The fact that the notch on box-plot of the proposed method does not overlap with that of other methods indicates proposed method performs significantly better than two other algorithms.



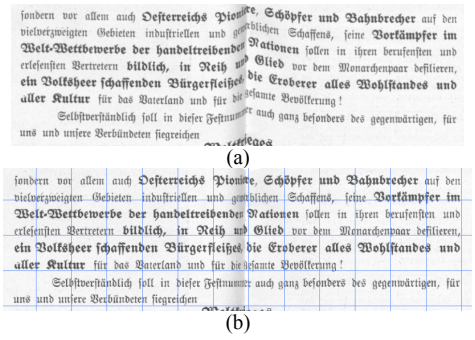
(a) (b)

Figure 5. (a) Original image, (b) dewarped by the proposed method



(a) (b) (c) (d)

Figure 6. Enlarged part of the (a) original image, De-warped image by (b) book restorer (c) NCSR (d) the proposed method



(a) (b)

Figure 7. (a) original image (fold) (b) the dewarped image by the proposed method

IV. CONCLUDING REMARKS

In this paper, a new dewarping method was proposed for historical document images that suffer from variety of geometric distortions including page curl, arbitrary warping, folds or any combination of them. The system does not require any prior information or special hardware and is able to achieve a high average accuracy of 93.94% on a diverse dataset of historical document images.

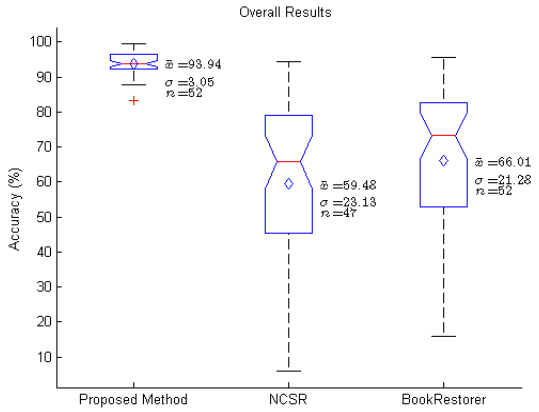


Figure 8. Overall accuracy for all images in the dataset

REFERENCES

[1] Cao, H., Ding, X., and Liu, C., "A cylindrical surface model to rectify the bound document image", *Int'l Conf. Computer Vision*, 2003, pp. 228-233.

[2] Zhang, Z., and Tan, C., "Correcting document image warping based on regression of curved text lines", *Int'l Conf. Document Analysis and Recognition*, 2003, pp. 589-593.

[3] Wu, M., Li, R., Fu, B., Li, W., and Xu, Z., "A Model Based Book Dewarping Method to Handle 2D Images Captured by a Digital Camera", *Int'l Conf. Document Analysis and Recognition*, 2007.

[4] Lu, S., and Tan, C.L., "Document flattening through grid modeling and regularization", *Int. Conf. on Pattern Rec.*, 2006, pp. 971–974.

[5] Zhang, Z., and Tan, C.L., "Straightening warped text lines using polynomial regression", *Int. Conf. on Image Processing*, 2002, pp. 977–980.

[6] Lavialle, O., Molines, X., Angella, F., and Baylou, P., "Active contours network to straighten distorted text lines", *Int. Conf. on Image Processing*, Thessaloniki, Greece, 2001, pp. 1074-1077.

[7] Bukhari, S., Shafait, F., and Breuel, T., "Dewarping of Document Images using Coupled-Snakes", *3rd Int. Workshop on Camera-Based Document Analysis and Recognition,* 2009, pp. 34-41.

[8] Doncescu, A., Bouju, A., and Quillet, V., "Former books digital processing: image warping", *Workshop on Document Image Analysis,* 1997, pp. 5-9.

[9] Brown, M., and Seales, W., "Image restoration of arbitrarily warped documents", *IEEE Trans. PAMI*, 2004, pp. 1295-1306.

[10] Yamashita, A., Kawarago, A., Kaneko, T., and Miura, K., "Shape reconstruction and image restoration for non-flat surfaces of documents with a stereo vision system", *17th Int. Conf. on Pattern Recognition*, 2004, pp. 482–485.

[11] Brown, M., and Tsoi, Y., "Geometric and shading correction for images of printed materials using boundary", *IEEE Trans. on Image Processing*, 15, (6), 2006, pp. 1544-1554.

[12] Gatos, B., Pratikakis, I., and Ntirogiannis, K., "Segmentation based recovery of arbitrarily warped document images", *Int. Conf. on Document Analysis and Recognition*, 2007, pp. 989-993.

[13] Stamatopoulos, N., Gatos, B., Pratikakis, I., and Perantonis, S., "A two-step dewarping of camera document images", *8th IAPR Workshop on Document Analysis Systems,* 2008, pp. 209–216.

[14] IMPACT: Improving Access to Text, EU FP7 project http://www.impact-project.eu

[15] Book Restorer, image restoration software, http://www.i2s-bookscanner.com