Grid-Based Modelling and Correction of Arbitrarily Warped Historical Document Images for Large-Scale Digitisation†

Po Yang, Apostolos Antonacopoulos, Christian Clausner and Stefan Pletschacher

Pattern Recognition and Image Analysis (PRImA) Research Lab
School of Computing, Science and Engineering, University of Salford, Greater Manchester, United Kingdom

http://www.primaresearch.org

ABSTRACT

Historical document images frequently show evidence of geometric distortions mostly due to storage conditions (arbitrary warping) but also due to the original printing process (non-straight text lines), the use of the document (folds) and scanning method (page curl). Correcting such distortions improves both recognition rate and visual appearance (e.g. for easier human reading or on-demand printing). However, the nature of the documents with layout irregularities and broken/touching characters of archaic fonts poses significant challenges. In addition, for large-scale digitisation of books and newspapers, methods need to be robust, efficient, reversible and must be able to be applied unsupervised on (possibly multi-columned) documents that may or may not be warped (no distortion should be introduced on unwarped images). No such method exists in the literature. In this paper, an effective grid-based method is presented to geometrically model and correct arbitrarily warped historical documents with relatively complex layout (multi column with graphics). A global grid with sub-grids for differing parts of a page is constructed by accurately determining text baselines. The warped image is corrected by transforming each quadrilateral sub-grid of the global grid into its intended rectangular form. Preliminary experimental results show that this method efficiently corrects arbitrarily warped historical documents, with an improved performance over a leading geometric correction method and the industry standard commercial system.

Categories and Subject Descriptors

H.3.7 [INFORMATION STORAGE AND RETRIEVAL]: Digital Libraries, I.7.1 [DOCUMENT AND TEXT PROCESSING]: Document and Text Editing --- Document management, I.7.5 [DOCUMENT AND TEXT PROCESSING]: Document Capture,I.5.4 [PATTERN RECOGNITION] Applications --- Text processing, Computer vision.

HIP '11, September 16 - September 17 2011, Beijing, China

Keywords

Geometric distortion, historical documents, text line detection, dewarping, geometrical model

1. INTRODUCTION

With the growing requirement for digital information, the demand for digitising historical paper documents has become essential for most libraries and museums [1]. However, the performance of OCR systems depends heavily on the quality of the document images, which are frequently influenced by various geometrical distortions. In historical documents such distortions are mainly due to adverse storage conditions like humidity (arbitrary warping), but also due to the original printing process (non-straight text lines), the use of the document (folds) and the scanning method (page curl). Accordingly, the restoration techniques can be mainly classified into geometrical correction approaches for page curl [2-5] and for arbitrary warping [6-12].

Correction approaches for page curl usually construct a geometrical model by either analysing image features (script-dependent) or through specialised scanning hardware (correction process independent from the image content). Cao *et al.* [2] proposed an analytically accurate cylindrical surface model to represent and rectify the bound image warping. This approach can successfully rectify a single image, without any specialised scanning hardware. Stamatopoulos *et al.* [5] also presented a coarse-to-fine strategy to compensate for undesirable document image distortions aiming to improve the OCR results. These two methods can both efficiently handle page curl, but not arbitrary warping of images. Additionally, there is a number of methods based on curve fitting (using text line representations or baselines) [3][4]. These approaches can efficiently handle page curl in single-column and purely textual documents, but are sensitive to the characteristics and layout of document images, such as font size, resolution and blurring.

Compared to the page curl approaches, most geometrical correction approaches for arbitrary warping [6-12] are based on the accurate acquisition of the 3D document geometry. This way they can directly handle arbitrary warping of documents, but they require complicated hardware setups for scanning the three-dimensional document surface (e.g. laser projector [11], structured light 3D acquisition [9][10] or two-camera stereo vision [12]). The high costs attached to such scanning solutions, however, prevent them from being implemented in large scale

† This work was funded by the EU 7th Framework Programme grant IMPACT (Ref: 215064).

digitisation projects with typically tight budgets. Moreover, it has to be noted that such approaches would require already existing images collections to be rescanned in order to obtain 3D surface models.

Image analysis based approaches for arbitrary warping, on the other hand, have to face specific challenges arising from the characteristics of historical documents. Text regions typically contain complex layouts with noisy characters, various font sizes and narrow text line spacing.

This paper presents an effective grid-based method to geometrically model and correct arbitrarily warped historical documents with relatively complex layouts (multiple columns, graphics, drop capitals etc.). A global grid with sub-grids for differing parts of a page is constructed by accurately determining text baselines. The warped image is corrected by transforming each quadrilateral sub-grid of the global grid into its intended rectangular form. An advantage of the grid-based method is that the transformation (correction) is reversible – a major requirement of the libraries (to be able to go back to the original master scans).

2. GEOMETRICAL CORRECTION METHOD

The proposed geometrical correction method for distorted historical documents can be described as follows: Given a binary document image:

1. Construct a global grid structure with multiple initial sub-grids (one for each identified text region in the document image).
2. Detect text lines for each text region (each sub-grid), and assign each connected component in the region to a corresponding text line.
3. Generate a mesh for each sub-grid with a precisely extracted baseline of the components in each text line.
4. Correct the distorted document image by processing each sub-grid with an affine transformation model.

3. PROCESSING STEPS

3.1 Construction of a Global Grid by Region Segmentation

A region segmentation algorithm based on geometric features has been developed and used to identify text regions in a given document image. Initially, the noise components within the input binary document image are filtered out by a threshold related to the average area of components. Then the image is divided into small squares called 'fragments'; fragments belonging to the document background (areas with a low foreground pixel count) are marked as processed. Then a feature called 'Erosion Count' is calculated and labelled for each fragment. The 'Erosion Count' of a fragment represents the number of erosions needed to erase the corresponding components belonging to this fragment. Fragments with similar erosion counts are grouped into a cluster by adopting several smearing, levelling and merging steps. Then, region polygons are calculated by merging the components within fragment clusters and tracing their outlines. Finally, the regions are classified as text or image using features of their corresponding fragment cluster.

Using this region segmentation method, the image and text regions in a document can be efficiently identified and segmented, since they show different 'Erosion Count' values. This approach is capable of detecting regions in historical books with simple layouts, but it tends to produce overlapping and inaccurately segmented regions when processing historical newspapers and magazines with complex layouts. Hence, the approximate detection of vertical and horizontal separating lines is compulsory for the identification of qualified text regions, which refer to the text regions without a clear vertical white space in text region and feasible size. In order to achieve this, global projection profiles in horizontal and vertical direction are calculated for the document image. Then, the peaks and valleys are detected to get the approximate vertical and horizontal separating lines. Finally, qualified text regions are chosen to construct a global grid for the document, marking them as areas to process, as shown in Fig.1.

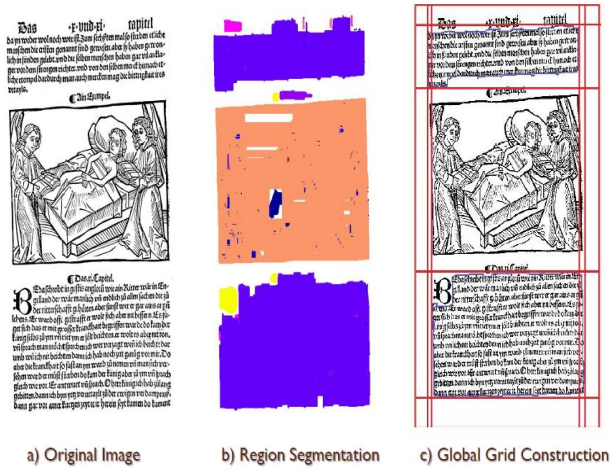


Figure 1. Construction of the global grid.

3.2 Text Line Segmentation

In this step, a dedicated text line segmentation method, which was developed by the authors in the scope of the IMPACT project [13], is used to detect text lines in the identified text regions. This method is based on a combination of connected components grouping and local projection profile analysis and is especially suitable for historical documents. The algorithm uses a data structure called Flex-Text-Line, which consists of horizontally aligned line segments with both top and bottom points, as shown in Fig.2. The segmentation process comprises three major steps, a) to group the extracted components of the regions to line candidates, b) to detect and split under-segmented line candidates using local projection profiles, c) to merge line candidates that are too small to their nearest neighbour. The result can be seen in Fig.3.

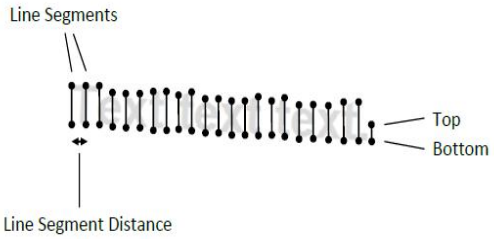


Figure 2. Flex-Text-Line with line segments.

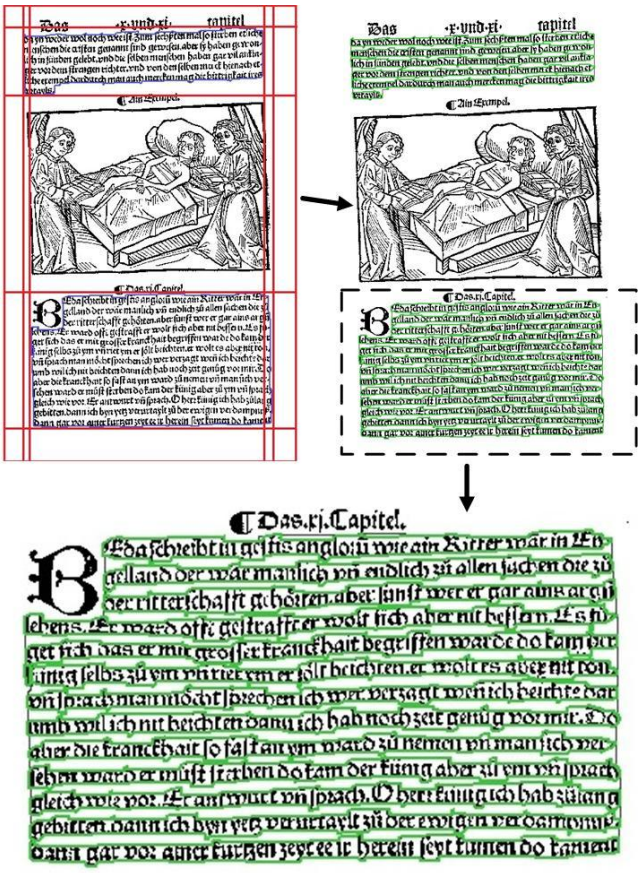


Figure 3. Text Line Segmentation Process.

3.3 Mesh Generation and Adjustment

The definition and identification of components in the presented method is based on connected component labelling (grouping of pixels of a binary image based on pixel connectivity). A noticeable issue is that the identified components in historical documents often do not correspond to individual characters, due to the occurrence of many broken/merged characters and noise. For each individual component i, a bounding box with four border parameters $x_i^1, x_i^2, y_i^1, y_i^2$ is built to determine its position in the document image. Area, height and width of each component are measured (parameters s_i, h_i, w_i) and used for region and text line segmentation. The average area, height and width of all components in this region can be observed as S_c, H_c, W_c.

The generation of a mesh for each qualified text region involves three steps. The first step is to create an initial rectangle-based mesh. The area of this rectangle is determined by the bounding box of the region. The number of rows in the mesh is equivalent to the number of text lines in the corresponding text region. Initially, all rows are evenly spaced within the mesh. The number of columns in the mesh is adjusted according to the average width of components in the regions and the number of components being covered by the sub-grid (see equation (1)).

$$C_N = \mathrm{Mod}\left(\frac{W_r}{\beta \times W_c}\right) + 1 \qquad (1)$$

Where:

C_N : Number of columns in the initial rectangle sub-grid.

W_r : Width of the bounding box of this region.

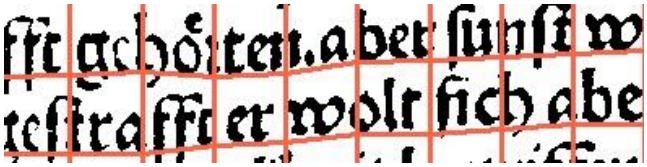
W_c : Average width of components in this region.

β : Number of components being covered by a sub-grid.

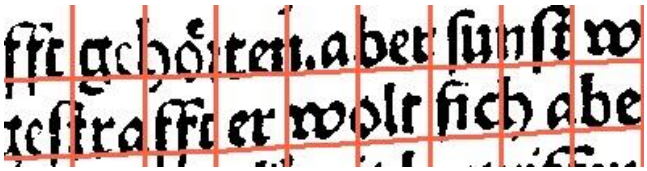
The second step is to detect the nearest connected components of each point in the mesh and adjust its position accordingly. The nearest neighbour components of a point can be detected by measuring the horizontal distance between point and components in a line cluster. The parameter y_i^2 of detected neighbour component i is used to adjust the vertical position of the corresponding point. The calculated row lines in the mesh reflect approximately the baselines of text in this region. However, during this process, it is possible to generate some incorrect point positions, due to broken components or descenders of letters. So a third process is employed to correct the ambiguous points in the mesh. The detection of ambiguous points in each row line is based on regression analysis. A linear polynomial equation is estimated for each row line by using the least-square algorithm. The ambiguous points are the points with a larger distance to this equation. If an ambiguous point is detected, its vertical position is replaced by the value produced by the linear polynomial equation.

There are two important issues in the mesh generation and correction. The first one is how to determine C_N in a qualified region. If C_N is too large, the observed points of each row line in the mesh would be dense. In this case, the probability of detecting the correct baseline is likely to increase, but the probability of detecting the incorrect baseline (due to descenders, noise, broken components) will likely increase too. Conversely, if C_N is too small, the observed points of each row line in the mesh would be sparse. Therefore, the probability of detecting the incorrect baseline is likely to decrease; but the probability of detecting the correct baseline would decrease too. So C_N should be within a feasible range. After testing several samples, we suggest the parameter β in equation 3 should be within range 2 to 4, which means each two adjacent columns in the mesh are on average covering two to four components.

The second issue is whether to use a curve fitting technique to improve the accuracy of baseline detection, as a number of researchers have done. Dealing with a large volume of observed data with a low correlation, the curve fitting technique can efficiently filter the outliers and smooth them. However, in our case, experiments show that the correlation coefficient of observed points in each row line is sufficiently high, at least 90%. Consequently, the utilization of curve fitting technique is not necessary in this case. Fig.4. illustrates the different mesh adjustment with curve fitting technique and with real baseline of components.



a) Mesh adjustment by bottom line of components



b) Mesh adjustment by curve fitting estimation.

Figure 4. Mesh adjustment.

3.4 Dewarping

The dewarping process with global grid in the presented method uses a transformation model to correct individual quadrilateral sub-grids (local meshes) to a rectangular shape. While there are several available transformation models for shape restoration algorithms, such as affine transformation, perspective transformation, bilinear transformation etc., the affine transformation is suitable to be applied in our case, since it can keep the straight lines and the ratios of distances during the transformation process. The vertical position of the target rectangles top and bottom line are determined by the average vertical position of corresponding rows in the mesh. The horizontal position of the target rectangles left and right would be kept at the horizontal position of the quadrilateral sub-grid. Experimental results can be seen in Fig. 5 and 6.

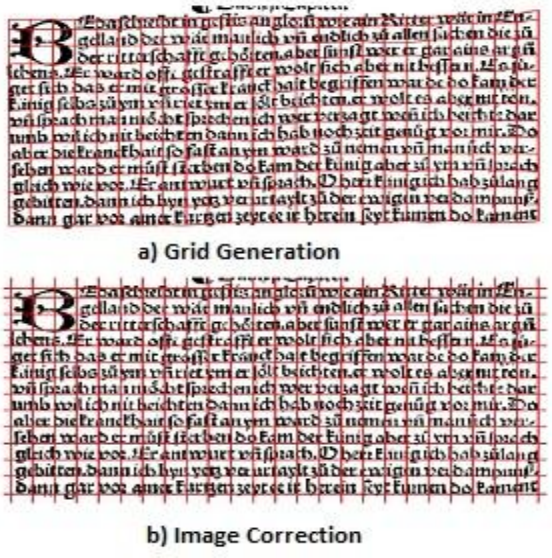
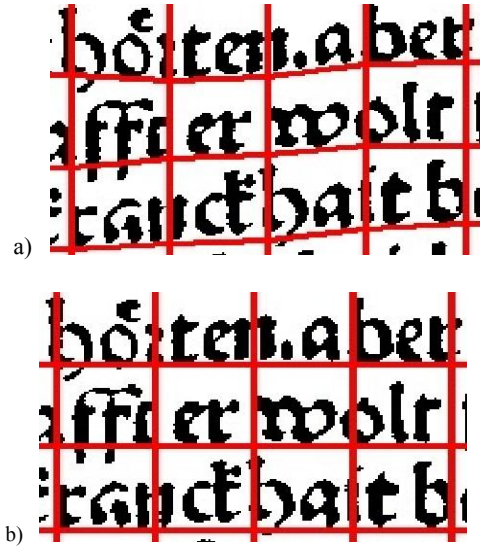


Figure 5. Image transformation using a local grid.



a)

b)

Figure 6. Detail of image in Fig.4. a) Original Image with text line based grid, b) Corrected Image with rectangle based grid

4. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed geometrical correction method has been evaluated using a diverse and representative sample of 10 arbitrarily warped historical document images with complex layout from the IMPACT project dataset [13]. The evaluation methodology used in this paper is based on supervised evaluation with (manually created) ground-truth data. Baselines on both the original warped document image and result document image are marked manually. Then, the average baseline straightness of the original and the corrected image is calculated according to equation (2). The results of two additional geometric correction methods are compared: one is a state-of-the-art page-curl correction method designed for IMPACT by NCSR [5] and the leading commercial product Book Restorer™ [14].

In order to evaluate dewarping by measuring the "straightness" of baselines, the average percentage of the sum of sub cross-area by rectangle-area in each baseline is measured. As shown in Fig.7, the sub cross-area refers to the area of the sub-region shaped by the baseline and average Y line.

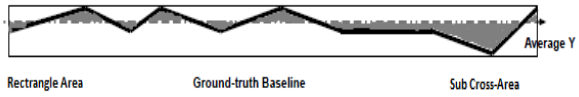


Figure 7.Rectangle area and sub cross-area of a baseline

Typically, for perfectly straight text lines, the ratio of sub cross-area to bounding box area is low (high accuracy), whereas, for heavily warped text lines, this ratio is expected to be significantly higher (low accuracy). So the accuracy of processing arbitrary warping in a document image with N baselines can be expressed by the following equation (2):

$$\text{Accuracy} = 1 - \frac{\sum_{1}^{N} \frac{\sum_{1}^{M} Sub_{ji}}{\text{Re}\, c_i}}{N} \qquad (2)$$

Where:

Sub_{ji} : Area of one sub cross-area in a marked baseline.

$\text{Re}\, c_i$: Area of bounding box of a marked baseline.

N : Number of baselines marked in a document image.

M : Number of sub-cross areas in one marked baseline.

The experimental results for the test set are shown in Fig.8.

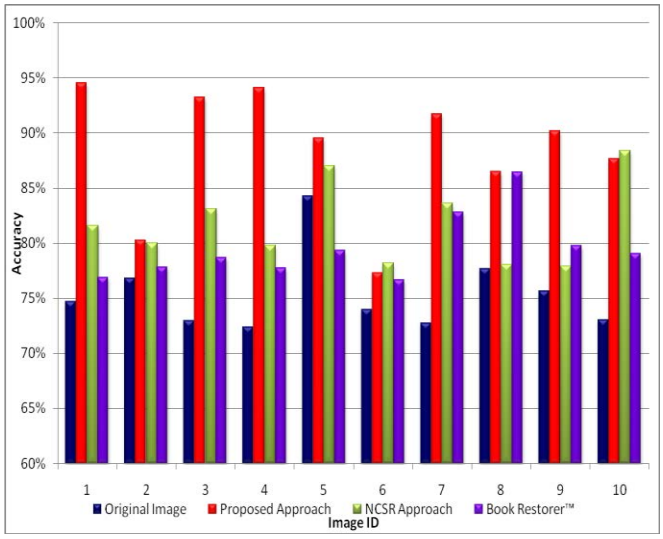
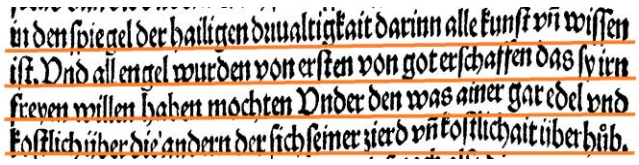
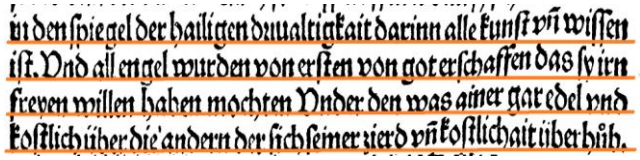


Figure 8. Evaluation Results of NCSR, Book Restorer™ and proposed method on arbitrarily warped documents

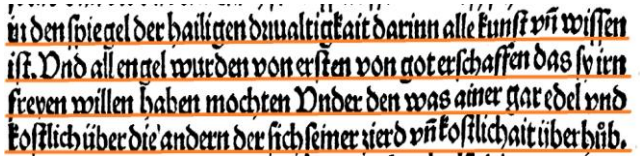
The results show that the proposed method can improve these 10 historical document images with arbitrary warping by increasing the accuracy from an average of 70% to 90%. The NCSR method can approximately improve the accuracy from an average of 70% to 80%. The Book Restorer™ software can achieve a maximum accuracy of just over 85%, but mostly the accuracy is between 75% and 85%. Overall, the presented method performs better than both the NCSR method and Book Restorer™ on correcting document images with arbitrary warping. Fig. 9 shows some regions from sample image 1 to illustrate the performance of different dewarping methods. From this figure, it can be seen that the proposed dewarping method is more advanced in processing arbitrary warping in images.



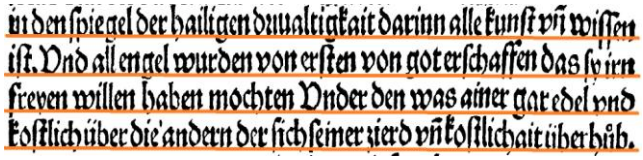
a) Arbitrarily warped region in original image.



b) Correction by Book Restorer™.



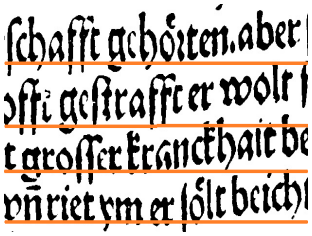
c) Correction by NCSR.



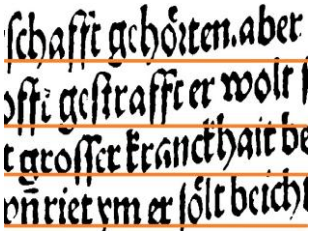
d) Correction by Proposed correction approach.

Figure 9. Performance of different correction approaches processing image 1.

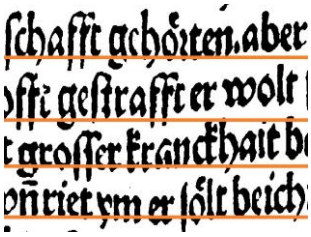
However, in some images with only slight arbitrary warping (images 2,6,10), the NCSR method shows a roughly equivalent improvement to our method. Those images suffer mostly from page curl and not so much arbitrary warping. Fig.10 shows some regions in the sample image 6 to illustrate the performance of different dewarping methods. It can be seen that the arbitrary warping has been significantly improved. However, considering that the area affected by arbitrary warping is rather small in the image, the overall accuracies of these three dewarping methods are not significantly different.



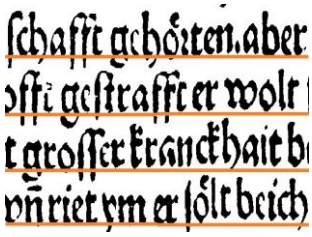
a) Arbitrary warping region



b) Correction by Book Restorer™.



c) Correction by NCSR.

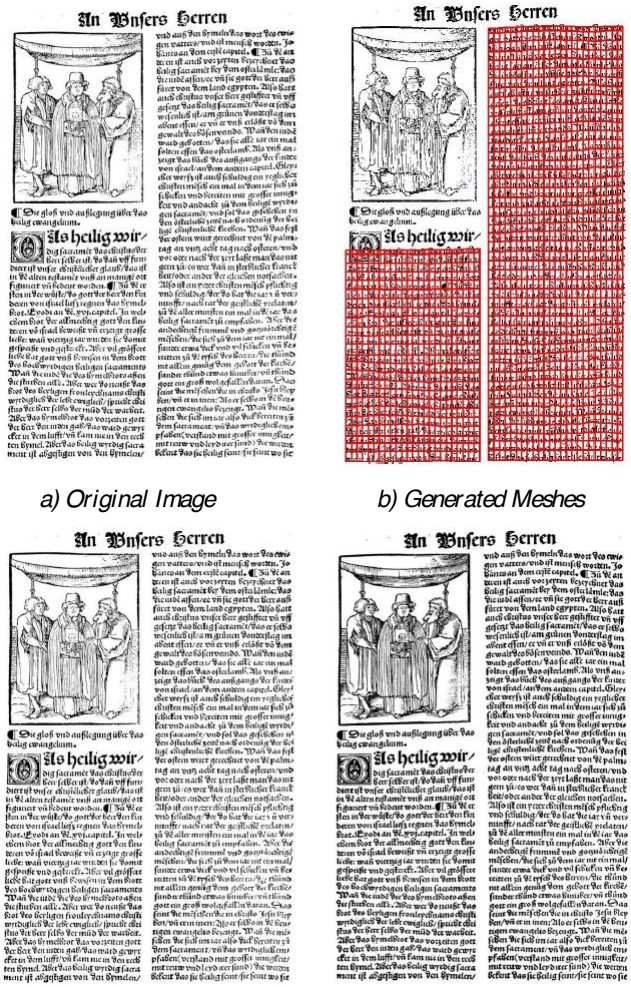


d) Correction by Proposed approach.

Figure 10. Performance of different correction approaches processing image 6.

Additionally, the document images used for evaluation mostly have a complex layout containing but not limited to: graphics, various font sizes and multiple columns. While the NCSR method can also process multiple column documents, it mainly solves the

problem of page curl and does not perform as effectively on arbitrarily warped documents. Fig.11 illustrates the proposed method and shows the result of the NCSR method for comparison. It can be seen that the proposed method can achieve a better performance than the page curl removal method of NCSR on arbitrarily dewarped document images with complex layout.



a) Original Image *b) Generated Meshes*

Dewarped Image by c) Proposed Method *d) NCSR method*

Figure 11. Dewarping of a sample image with complex layout.

5. CONCLUSION

In this paper, a novel grid-based geometrical correction method is presented which aims at arbitrarily warped historical documents. Our experimental results show that the proposed method can perform better than the leading start-of-the-art geometric correction methods on dealing with arbitrary warping effects. Another advantage is that this method can process document images with complex contents such as graphics, multiple font sizes and multiple column layouts. Furthermore, the geometric correction is reversible due to the use of a dewarping model (grid) which can be saved for this purpose.

The method in its current state is designed for book images (as most other methods in the literature), albeit with complex layouts; it has yet to be evaluated on more challenging historical documents with highly dense and intricate layouts (for instance newspapers). Future work will focus on this problem.

6. REFERENCES

[1] A. Antonacopoulos, D. Karatzas, "*Document Image analysis for World War II personal records*", First Int. Workshop on Document Image Analysis for Libraries, DIAL'04, Palo Alto, pp. 336-341, 2004.

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

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

[4] Z. Zhang., and C. Tan., "*Document flattening through grid modeling and regularization*", Int'l Conf. on Pattern Recognition, 2002, pp. 977-980.

[5] N. Stamatopoulos, B. Gatos, I. Pratikakis., and S.J. Perantonis, "Goal-*Oriented rectification of camera based document images*", IEEE Transactions on Image Processing, pp910-920, 2011.

[6] C.L.Tan, L. Zhang, Z. Zhang, T. Xia. "*Restoring warped document images through 3D shape modelling*" IEEE Transactions on Patten Analysis and Machine Intelligence, Vol 29, issue 3, pp195-210, 2006.

[7] Y. Y. Tang and C.Y. Suen, "*Image Transformation Approach to Nonlinear Shape Restoration*," IEEE Trans. Systems, Man, and Cybernetics, vol. 23, no. 1, pp. 155-171, Jan./Feb. 1993.

[8] O. Lavaille, X. Molines, F. Angella, and P. Baylou, "*Active Contours Network to Straighten Distorted Text Lines*," Proc. Int'l Conf. Image Processing, pp. 1074-1077, Oct. 2001.

[9] M.S. Brown and W.B. Seales, "*Image Restoration of Arbitrarily Warped Documents*," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, no. 10, pp. 1295-1306, 2004.

[10] M.S. Brown and Y.-C. Tsoi, "Geometric and shading correction for images of printed materials using boundary," Image Processing, IEEE Transactions on , vol.15, no.6, pp.1544-1554, June 2006

[11] A. Doncescu, A. Bouju, and V. Quillet, "*Former Books Digital Processing: Image Warping*," Proc. Int'l Workshop Document Image Analysis, pp. 5-9, 1997.

[12] A. Yamashita, A. Kawarago, T. Kaneko, and K. Miura, "Shape reconstruction and image restoration for non-flat surfaces of documents with a stereo vision system", 17th Int'l Conf. on Pattern Recognition, 2004. Pp. 482-485.

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

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