Effective Geometric Restoration of Distorted Historical Document for Large-Scale Digitization

Po Yang [1*], Apostolos Antonacopoulos[2], Christian Clausner[2], Stefan Pletschacher[2], Jun Qi[1]

[1] Department of Computer Science, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK
[2] School of Computing, Science and Engineering, Salford University, 43 Crescent, Salford, M5 4WT, UK
[*] p.yang@ljmu.ac.uk

Abstract: Due to storage conditions and material's non-planar shape, geometric distortion of the 2-D content is widely present in scanned document images. Effective geometric restoration of these distorted document images considerably increases character recognition rate in large-scale digitisation. For large-scale digitisation of historical books, geometric restoration solutions expect to be accurate, generic, robust, unsupervised and reversible. However, most methods in the literature concentrate on improving restoration accuracy for specific distortion effect, but not their applicability in large-scale digitisation. This paper proposes an effective mesh based geometric restoration system, (GRLSD), for large-scale distorted historical document digitisation. In this system, an automatic mesh generation based dewarping tool is proposed to geometrically model and correct arbitrary warping historical documents. An XML based mesh recorder is proposed to record the mesh of distortion information for reversible use. A graphic user interface toolkit is designed to visually display and manually manipulate the mesh for improving geometric restoration accuracy. Experimental results show that the proposed automatic dewarping approach efficiently corrects arbitrarily warped historical documents, with an improved performance over several state-of-the-art geometric restoration methods. By using XML mesh recorder and GUI toolkit, the GRLSD system greatly aids users to flexibly monitor and correct ambiguous points of mesh for the prevention of damaging historical document images without distortions in large-scale digitalisation.

Keywords: Geometric restoration, document processing, historical documents, large-scale digitalisation

1. Introduction

With the emergence of cheap digital storage, large-scale historical document digitisation has received significant attentions by libraries and museums in the past decade [1]. Owning to storage conditions and materials' non-planar shape, a common problem affecting the character recognition rate in digitising historical documents is geometrical distortion. Typically, geometrical distortions in historical documents are mainly results from adverse storage conditions like moisture, the original printing or scanning procedure and the use of the document [2]. For large-scale digitisation, effective geometric restoration of these distorted document images considerably increases character recognition rate in large-scale digitisation. Given these problems, the motivation of this research is to seek out an accurate, generic, robust, unsupervised and reversible geometrical restoration approach to restore the large number of historical document that may or may not be distorted.

Over the last few years, there has been related automatic geometrical restoration techniques reported in

the literature. Regarding major types of distortions, the restoration techniques are categorised into two classes: page curl correction methods [3-12] and arbitrary warping correction methods [13-19]. Page curl correction approaches normally build a geometrical model by analysing image features [3-8] or using scanning specialised scanning hardware [9-12]. Page curl usually causes consistent global distortions on document images. The dewarping results of page curl correction methods are mostly promising with given an accurate estimation of geometrical model. Arbitrary warping correction methods rely on a precise acquisition of 3D geometry of document. These procedures need complicated hardware setups [13-16] for scanning the three-dimensional document surface. In large-scale digitisation for historical documents, neither of the above geometric restoration technique can be widely recognised as an efficient solution. The primary issue is that large-scale digitisation of historical documents practically needs a geometrical restoration handling both page curl and arbitrary warping. Considering the large volume of historical books, it is impractical for libraries or museum to manually identify and classify the page curl image and arbitrary warped image in all documents. Existing geometrical restoration methods rarely work efficiently on both page curl and arbitrary warping cases. Secondly, the complex layouts of historical documents introduce significant difficulties to segment and identify their text regions. Text regions in historical documents contain multiple columns, small graphs, various character font sizes and indistinguishable text line spacing. Such issues make difficulties precisely segment text lines as the representation of geometrical distortion. Finally, geometrical correction solution for large-scale digitisation needs to be robust and revisable, which means that it should not damage the historical document images without distortions, and also be able to correct it if there are some damages.

In this paper, we present an effective mesh based geometric restoration system (GRLSD), for large-scale distorted historical document digitisation. The idea of this system is inspired from our early work [2] in IMPACT project [26], which proposes an effective grid-based method to unsupervised correct warped historical documents with complex layouts. But the findings in [2] shows that individual unsupervised geometrical restoration approach is hard to accurately tackle all types of distortions in warped documents and inevitable to introduce unexpected distortions to normal documents in large-scale digitalisation. Therefore, GRLSD system aims to integrate both automatic and manual modes into one effective geometrical restoration way. In this system, an automatic mesh generation based dewarping tool is proposed to geometrically model and correct arbitrary warping historical documents. An XML based mesh recorder is proposed to record the mesh of distortion information for reversible use. A graphic user interface toolkit is designed to visually display and manually manipulate the mesh for improving geometric restoration accuracy. Experimental results show that the proposed automatic dewarping approach efficiently corrects arbitrarily warped historical documents, with an improved performance over several state-of-the-art geometric restoration methods. By using XML mesh recorder and GUI toolkit, the GRLSD system greatly

aids users to flexibly monitor and correct ambiguous points of mesh for the prevention of damaging historical document images without distortions in large-scale digitalisation. The major advantages of the proposed new geometrical restoration system are:

- **Ability of processing document with arbitrary warping and complex layout:** The global grid construction and the simultaneous transformation of sub-grid can efficiently handle historical document with complex layout. An unsupervised mesh generation approach is introduced to enhance the accuracy of baseline extraction, with a capability of robustly correcting arbitrary warping effect. In comparison to the current leading geometric restoration method [12] and industry standard commercial system [27], the proposed approach has an improved performance.

- **Suitable to large-scale digitalisation**: The whole process of proposed geometrical correction approach is unsupervised. The parameters of each process step are automatically generated and depended on the characteristics of documents. Also, the plausibility check and outlier correction process is introduced to detect and correct the ambiguous points or lines in the mesh, for the prevention of damaging historical document images without distortions.

- **Reversible transformation process**: An advantage of the proposed system is that the transformation (correction) is reversible – a major requirement of the libraries (to be able to go back to the original master scans). It greatly aids users to flexibly monitor and correct ambiguous points of mesh for the prevention of damaging historical document images without distortions in large-scale digitalisation

The rest of the paper is organized as follows. Section 2 reviews the related literature of geometric restoration techniques. Section 3 provides a system overview of GRLSD system. Section 4 represents the detailed optimization and improvement in proposed system. And the experimental validation results in Section 5. Section 6 includes conclusions to be drawn from the work and suggested areas for future investigation.

2. Related Work

As mentioned previously, geometrical correction approaches for page curl model the geometric warping by 2D-based approaches [3-8] (document image processing and analysis) or 3D-based approaches [9-12] (physical capture surface shape). Cao *et al.* [3] represented an analytically accurate cylindrical surface model to correct the bound image warping. Cylindrical surface model is capable of successfully rectifying the page curl of a single bound document image without any specialised scanning hardware. But it requires the document image scanned from a particular angle to guarantee the generatrix of the cylinder paralleling the image plane. Additionally, Zhang and Tan [4-8] presented several curve fitting based geometrical restoration

methods for document images from bound volumes. The basis of their approaches is to divide the image into shade and non-shade regions; model the alignment of text by straight reference lines in non-shade regions; and model the distorted text lines in shade field using polynomial regression. Other similar curve fitting techniques are also used to model and improve the smooth variations of the representation of text lines [9]. While they are sufficiently efficient to deal with page curl in simple document layout (single-column and purely textual documents), it is hardly to restore the documents with complex layout and severe arbitrary warping or folds. Later on, some attempts on using energy minimisation algorithms (active contours) have been investigated [10-11] to optimize the accuracy of text lines detection. However, the success of active contours based methods need the high quality initialisation and accurate text line segmentation; also the slowness of these methods is not suitable for handling massive digitisation.

Unlike page curl geometric correction methods, geometrical correction approaches for arbitrary warping [12-18] are mostly 3D-based approaches, which use the surface shape of the document to model the 3D geometric warping and then rectify using physical flattening process. The primary benefit of these approaches is high restoration accuracy, since they can observe highly accurate presentations of the physical warping. Also, they are independent to the diverse content of document images. However, the utilization of these solutions normally needs complicated hardware configurations for scanning the 3D document surface (e.g. laser projector [17], structured light 3D acquisition [15] [16] or two-camera stereo vision [18]). Also, the high costs attached to such scanning solutions are not suitable for large-scale digitisation. Stamatopoulos *et al.* [12] [21] presented a goal-oriented coarse-to-fine rectification strategy to compensate for moderate arbitrary warping in historical documents, aiming to improve the OCR results without auxiliary hardware. This method can significantly improve the OCR results by rectifying historical document image, but it is limited to use for single column document and not suitable to large-scale digitalisation.

In terms of above reviews, there are no existing efficient geometric restoration solutions for large-scale digitisation of historical documents. Meanwhile, most approaches in the literature concentrate on improving restoration accuracy for specific distortion effect, but not their practical applicability. This paper considers proposing an accurate, generic, robust, unsupervised and reversible geometric restoration solution.

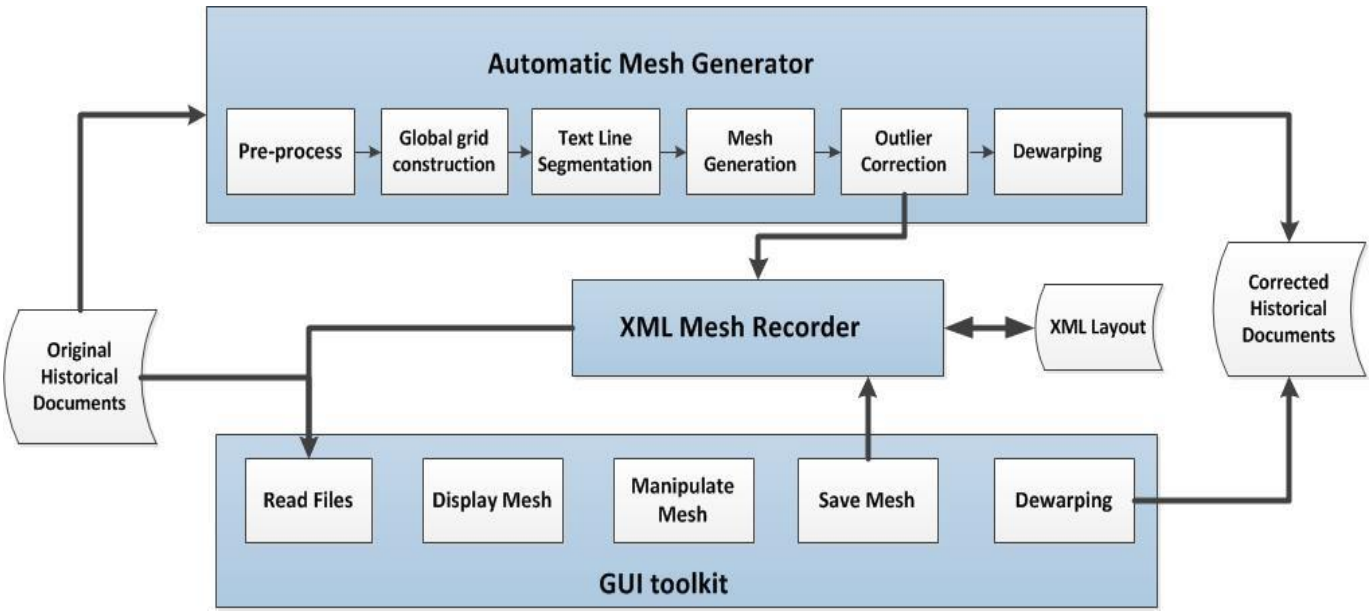


Fig. 1. *Work flow of GRLSD system.*

3. System Overview

GRLSD system aims to accurately, efficiently, effectively and robustly restore the distorted historical documents for massive digitization. Fig.1 schematically represents the workflow in GRLSD system.

GRLSD system initially focuses on unsupervised correcting the geometric distortion of historical documents by implementing a method from our early work [2]. This method automatically constructs a global grid with sub-grids by successively carrying out text region segmentation and text baselines extraction. In GRLSD system, this grid-based dewarping method is extended as an automatic mesh generator by adding an automatic outlier correction scheme. As shown in Fig.1, the mesh generator geometrically model arbitrary warping historical documents though pre-process, global grid construction, text line segmentation and mesh generation. Compared to previous work [2], a new automatic outlier correction scheme is introduced to improve the quality of mesh by using plausibility check and outlier correction. The dewarping process uses affine transformation model correct individual quadrilateral sub-grids of meshes to a rectangular shape. The unsupervised correction method is implemented as a command line based library.

Reversible transformation process is another important feature of GRLSD system. In order to achieve this goal, recording and storing the content of automatic generated mesh are critical. The meshes are represented according to a sophisticated XML schema which is component of the PAGE (Page Analysis and Ground truth Elements) Format Framework [22]. An XML based mesh recorder is designed and implemented to record the mesh reflecting geometric distortion for reversible use.

GRLSD system also enables manual correction of generated meshes using human perception. A graphic user interface (GUI) toolkit is designed to visually display and manually manipulate the mesh for

improving geometric restoration accuracy. After executing automatic mesh generator, an XML file restoring meshes are produced to each document image. Using GUI toolkit, users can read the XML file and view the meshes with relevant document images. If any incorrect outliers of meshes are founded though human perception, users can manipulate the positions of outliers and save the revised meshes.

3.1. Automatic Mesh Generator

The automatic mesh generator in GRLSD system adapts and implements our unsupervised geometric dewarping approach [2], which is an effective grid-based method to geometrically model and correct arbitrarily warped historical document with relatively complex layout. The detailed steps in this method are below:

- **Pre-process:** In preparation for the succeeding steps, connected components within the bitonal (distorted) input image are identified, based on a standard labeling approach using pixel connectivity. In addition, specific noise (small black components) is filtered out.

- **Global Grid Construction:** First, the page is segmented into regions (zones) using a bottom-up approach based on local features. The results are then labeled as text or non-text using a classification method that exploits the same features. Lastly, vertical and horizontal separating lines are identified and used to refine the recognized page layout. This step is especially useful for documents with complex layouts, such as newspapers and magazines. Fig.2 shows outcome of the page analysis step for an example document. The global grid is finally constructed based on the identified text regions.

- **Text Line Segmentation**: As a prerequisite for finding geometric distortions, text lines need to be detected. To this end, a dedicated hybrid text line segmentation algorithm [29] is applied to the text regions of the global grid. 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. Fig.3 shows an example of detected text lines.

The segmentation method is suitable for mass digitization and has proven its applicability to historical documents in two major EU-funded projects [27] [30]. While already optimized for a broad range of historical documents, it was specifically designed to be adaptable to different document classes. Furthermore, the method is comparatively robust against geometric distortions (e.g. skew) – a crucial feature in the context of this work.

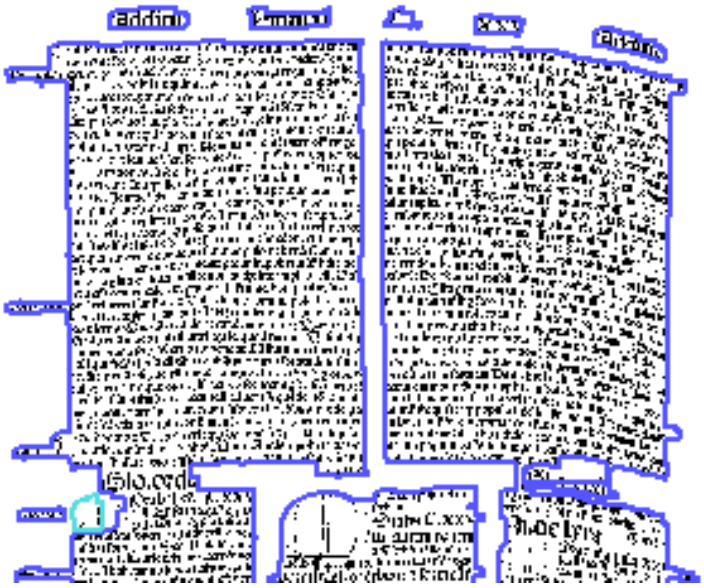


Fig. 2. *Segmentation into regions*

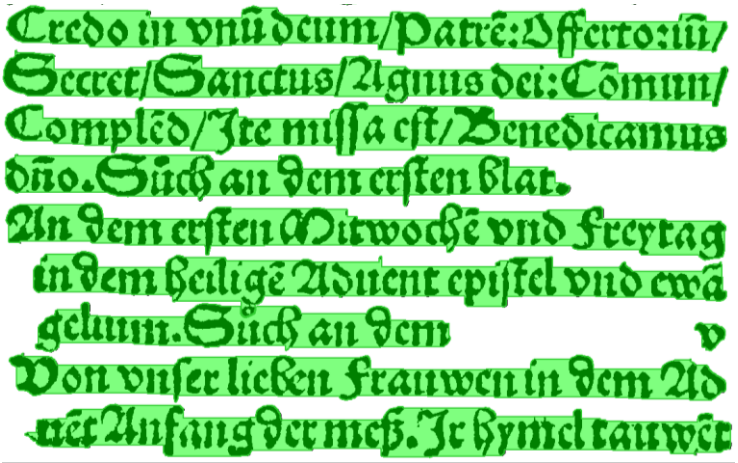


Fig. 3. *Segmentation into text lines*

- **Mesh Generation:** it firstly creates an initial rectangle- based mesh in terms of the average width of connected components; and then detects the nearest connected components of each point in the mesh and adjust its position accordingly.

- **Dewarping:** it uses a transformation model to correct individual quadrilateral sub-grids (local meshes) to a rectangular shape.

3.2. XML Mesh Recorder

The XML meshes recorder is implemented by PAGE (Page Analysis and Ground truth Elements) [22] XML schema. The PAGE format framework consists of a *root* instance to reflect the structure of automatic mesh generator work flow in GRLSD system, and a *gts* (ground truth and storage) instance recording the actual data of a mesh in document images. The actual data of a mesh normally includes the point position of

a mesh grid, the horizontal and vertical reference line of a mesh grid, and the number and location of sub-grid in a global grid. Also, some attributions of images and metadata are recorded in the XML schema. Fig.4 shows a schematic view of a PAGE dewparing instance of the GRLSD system.

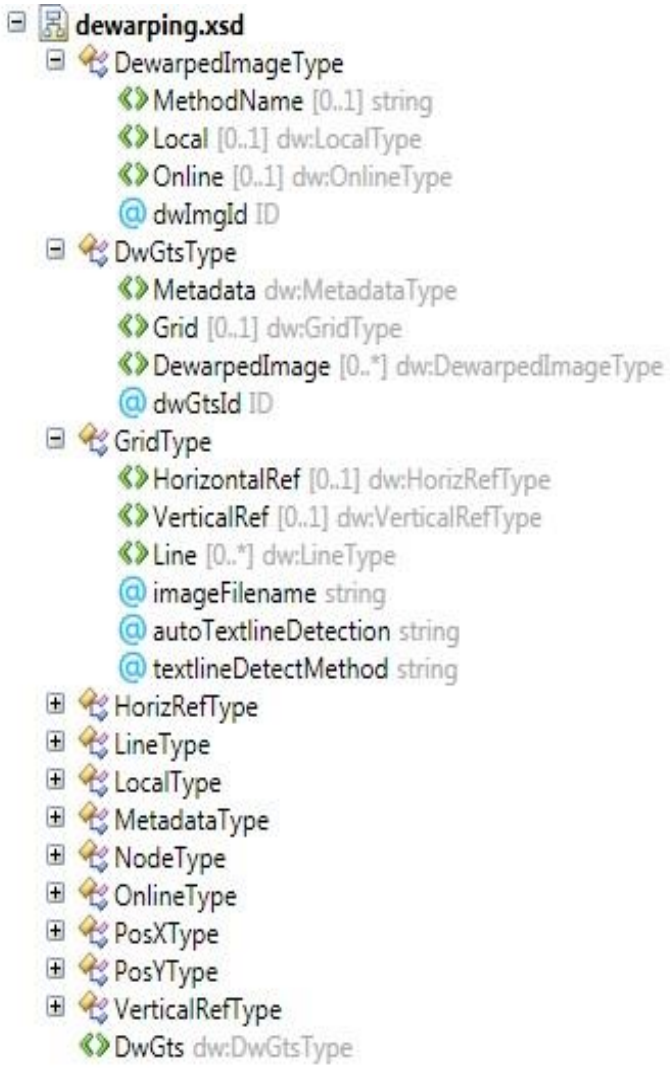


Fig. 4. *Schematic view of a PAGE dewarping instance*

3.3. GUI Toolkit

The GUI toolkit is designed for manually or automatically correct warped document images with a global grid based mesh. Regarding Fig.1, the main functionalities of GUI toolkit overs the aspects below:

- **Read File:** To read a XML mesh and document image can either using "New" or "Open" in GUI toolkit. The "New" button will pop up to select a colour document image and the corresponding black-and-white image (only images in TIFF format are supported). Then users can manually create a grid with expected mesh shape. The "Open" button enables opening an existing document with XML based mesh file. A dialogue will be presented allowing user to select a PAGE XML file. The XML file contains

the file name of the colour image and GUI toolkit will load it automatically.

- **Display Mesh:** Three main different display modes in the GUI toolkit are defined, which are "Display both grid and image", "Display image only" and "Display grid only". Also, to hide image or grid, zoom in/out are also implemented here.

- **Manipulate Mesh:** The manipulation of mesh occurs on both lines and point of grid. The "add" and "delete" buttons can add or delete a vertical or horizontal line in the grid. The points of mesh can be moved by "Single Point" individually or "Multiple Points" together. It also can move "Reference Lines" regarding the initialized grid.

- **Save Mesh and Dewarping:** The **"**Save" function enables recording the manually created grid or revised grid in the XML files. The "Dewarping" button provides two ways to straighten each horizontal line with affine transformation, which are using "Average" points or "Reference" line of each horizontal line of grid to dewarp the image. Fig.5 shows a user interface of proposed GUI Toolkit of the GRLSD system.

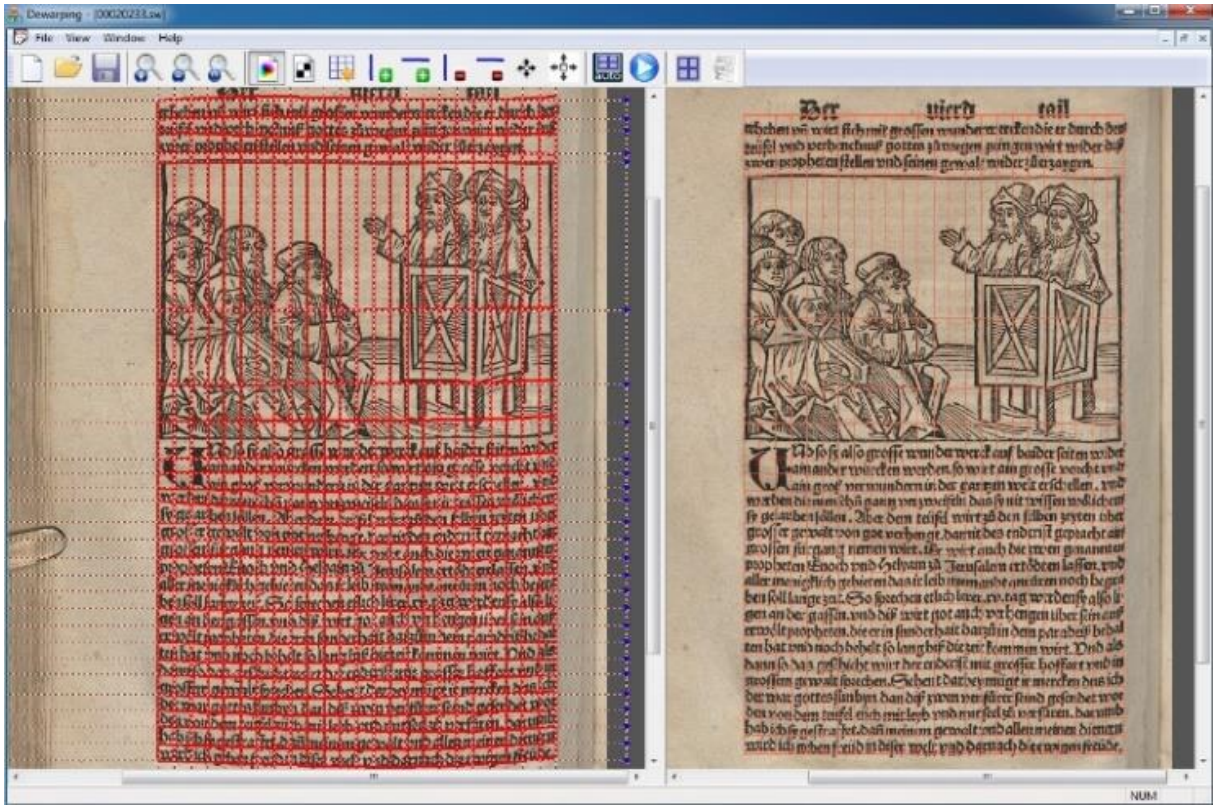
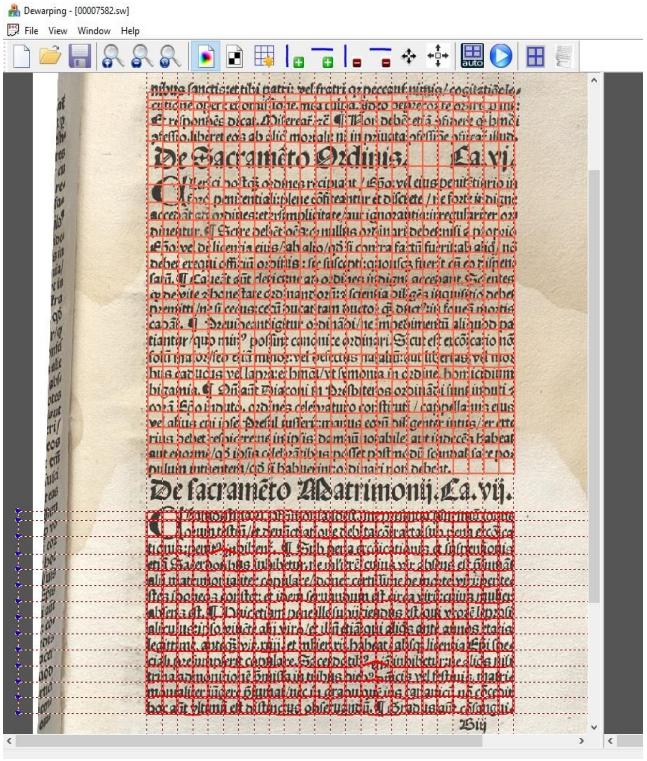


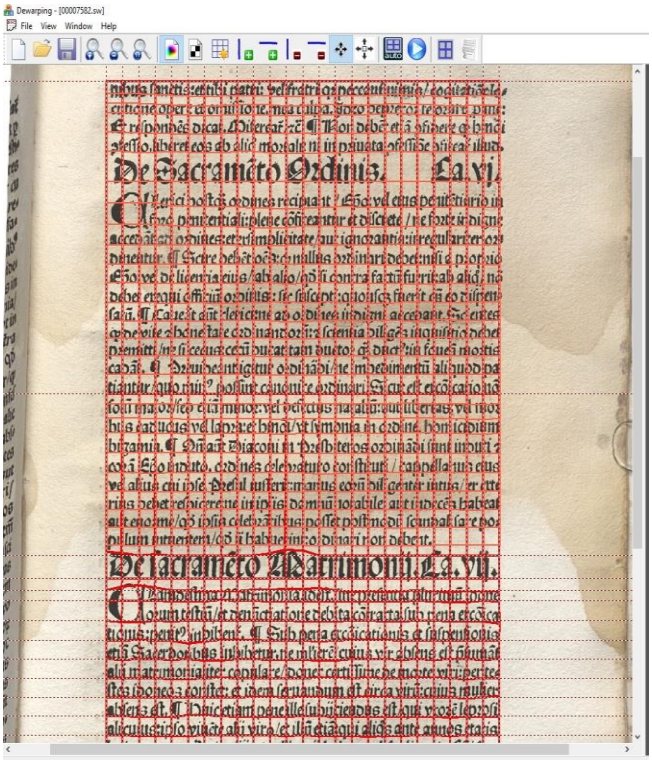
Fig. 5 *GUI Toolkit for displaying and dewarping historical document images.*

- **Reversible use of XML meshes:** The GRLSD system has also provided a command line tool to automatically dewarp a batch of images, with outputting auto-generated meshes saving in the XML file and a dewarped image. However, these meshes usually contain ambiguous points resulting in local distortions in the final image; thus, the GUI toolkit of the GRLSD has to support displaying and manipulating the

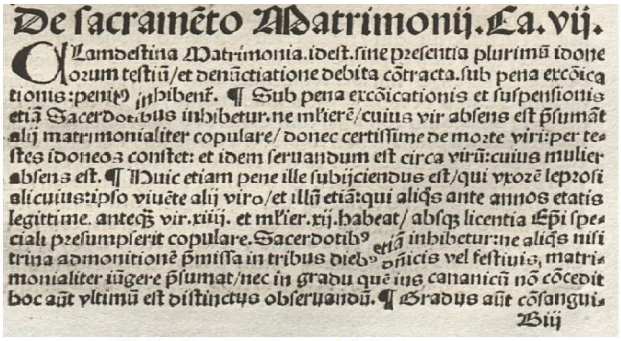
a) Auto-generated meshes XML file



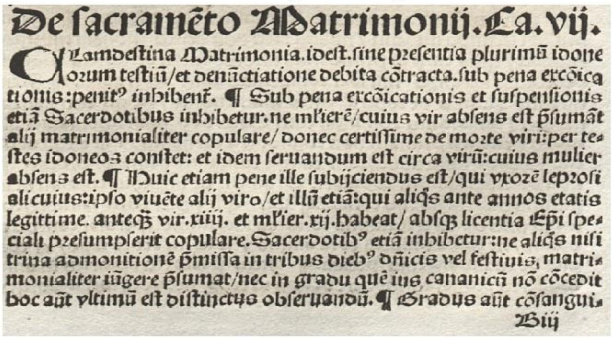
b) Auto-generated meshes displaying in GUI



c) Reversible corrected meshes displaying in GUI



d) Auto-dewarped image



f) Reversible mesh-correcting dewarped image

d) & f) Comparision of dewarped images with b) and c)

Fig. 6. *GUI Toolkit for reversibly correcting dewarped images.*

Auto-generated meshes and generating the better quality of dewarped images. As shown in Fig.6 a)-f), after running a cmd file to auto-process a batch of original wapred images, the auto-generated meshes is produced and saved in a XML file in fig.6.a; we could also use GUI toolkit to display this mesh in fig.6.b, which misses some text lines and contains some ambiguous points in the bottom grid; then we correct the meshes in fig.6.c by adding some extra text lines and adjusting the points of baselines; the final dewarped images in fig.6.d and f indicate that the reversible mesh-correcting dewarped image has a better quality than auto-dewarped image.

4. Optimization of Automatic Mesh Generator

The mesh generated through automatic generator by method [2] in GRLSD is capable of reflecting an approximate geometrical distortion effect of historical documents, but it also generates some incorrect point positions. We optimize the automatic mesh generator by introducing a process of Plausibility check and outlier correction. It aims to detect the outliers of mesh and correct them. Regarding the former processes, the outliers of mesh are possibly caused by three reasons.

The first reason is the incorrect region segmentation. In this case, the row lines of mesh could possibly be non- smooth, non-regular; and the outliers are distributed globally. The second reason is the incorrect text line segmentation. This case usually occurs on documents with serious page curl, so the row lines of mesh could possibly be roughly smooth and regular; the outliers are distributed locally on each row lines. The third reason is due to broken components or descenders of letters. In this case, the row lines of mesh could possibly be quite smooth and regular; the outliers are distributed occasionally and locally on some row lines.

Regarding above classifications, we separately use regression analysis and RMSE measurement methods to identify outliers and correct mesh.

4.1. Correction of Local Ambiguous Point

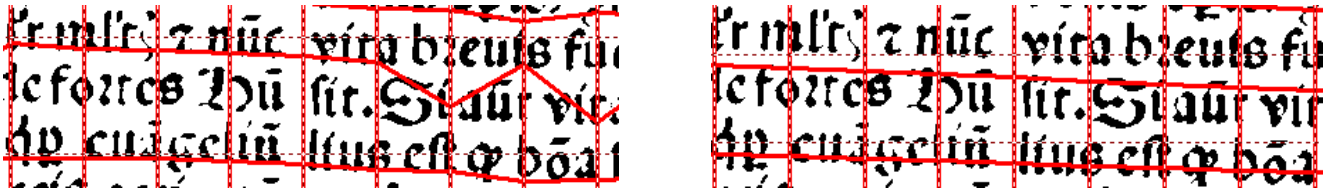
The first step is to correct some local ambiguous point in the mesh by using RMSE measurement, as shown in Fig. 6(a). Given a raw mesh $M = \{a_{ij} : i \in (1,2,...R_g), j \in \{1,2,...C_g\}\}$, the position of each point a_{ij} are denoted as (x_{ij}, y_{ij}). For each row line i, a linear polynomial equation can be fitted by equation 1.

$$y = a_i * x + b_i \tag{1}$$

The goodness of fit of equation 1 is calculated by using RMSE measurement. If RMSE is large, the row line i has a bad fit to linear polynomial equation, so this row line may have many outliers. For each point (x_{ij}, y_{ij}) of this row line i, to do the action by equation 2.

$$y_{ij} = \begin{cases} a_i * x_{ij} + b_i, & \text{if RMSE} < \partial \\ y_{ij}, & \text{if RMSE} \geq \partial \end{cases} \tag{2}$$

In equation 2, the parameter ∂ is used as a benchmark of RMSE to correct the local outliners in mesh. In our case, the value of ∂ are justified to set as 25 for removing the majority local ambiguous points. Fig.7 a) and (b) shows a sample document image being corrected through first step.



a *b*

Fig.7. *Mesh adjustment by RMSE measurement. a)* Mesh with local ambiguous points. b).Outlier correction by RMSE measurement.

4.2. Correction of Global Outliers

The second step is to correct some global outliners in the mesh by using regression analysis and linear polynomial fitting methods. Firstly, it uses average slope of each row line for identifying the outliers from incorrect text line segmentation. For each row line i, the sum of slope M_i can be measured by equation 3.

$$M_i = \sum_{2}^{C_g} \left(\left| y_{i,j} - y_{i,j-1} \right| / \left| x_{i,j} - x_{i,j-1} \right| \right) \tag{3}$$

The average mean of slope M_i of all row lines in document region can be measured as:

$$M_A = \left(\sum_{2}^{R_g} M_i \right) / (R_g - 1) \tag{4}$$

Considering the global effect of page curl, document with page curl is supposed to have a close M_i for each row line. The range of these row lines with close M_i is denoted as [a, b]; the row lines with distinguishable M_i are effected by outliers from text line segmentation. Additionally, document with arbitrary warp is supposed to have a M_A near to zero, which is normally smaller than the M_A of document with page curl. In order to correct these outliers, for each row line *i*, a 3 degree polynomial equation can be used to fit these row lines with close M_i .

$$y = A_i * x^3 + B_i * x^2 + C_i * x^1 + D_i \tag{5}$$

So for each point (x_{ij}, y_{ij}) in this row line *i*, the correction procedure is carried out by equation 6.

$$y_{ij} = \begin{cases} A_i * x_{ij}^3 + B_i * x_{ij}^2 + C_i * x_{ij}^1 + D_i, & \text{if } \mathrm{M}_A \geq \beta \\ y_{ij}, & \text{if } \mathrm{M}_A < \beta \end{cases} \tag{6}$$

In equation 6, the parameter β is used as a benchmark of M_i to correct the global outliners in mesh.

4.3. Recheck of Ambiguous Points

It has to check if there are some new possible ambiguous points generating by equation 6. A linear polynomial equation is estimated for each row line by using the least-square algorithm. If an ambiguous point is detected, its vertical position is replaced by the value produced by the linear polynomial equation. Given a linear polynomial equation by a row line of mesh:

$$y_{ij} = a * x_{ij} + b \tag{7}$$

For each point (x_i, y_i) of this row line, to check:

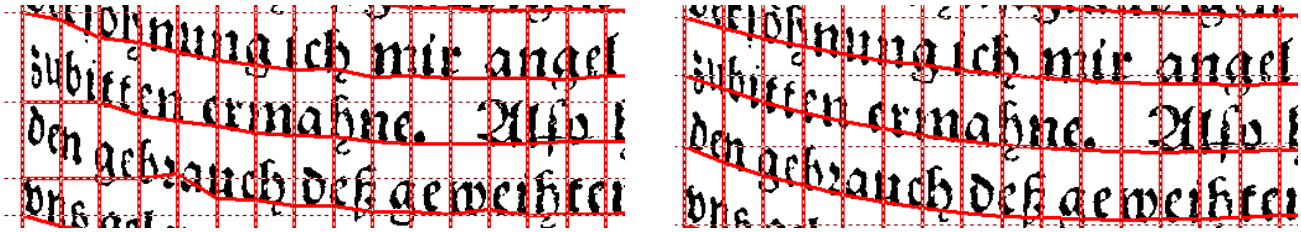
$$y_i = \begin{cases} a * x_i + b, & \text{if } |y_i - a * x_i + b| \geq \delta * H_c \\ y_i, & \text{if } |y_i - a * x_i + b| < \delta * H_c \end{cases} \tag{8}$$

Where:

H_c: Average height of components in this region.

α: Parameter to correct ambiguous point, from 0 to 1.

Fig.8 a) and b) shows a sample document image being corrected through linear polynomial curve fitting. Dealing with a large volume of observed data with a low correlation, the curve fitting technique can efficiently filter the ambiguous points and smooth them. However, in some cases, experiments show that the correlation coefficient of observed points in each row line is sufficiently high, at least 90%. Here, the utilization of curve fitting technique to smooth all points of each row is not necessary.



a *b*

Fig. 8. *Mesh adjustment by linear polynomial curve fitting*. a) Mesh with global ambiguous points. b) Outlier correction by linear polynomial fitting

5. Performance Evaluation

Some experiments have been carried out to evaluate the effectiveness of the proposed GRLSD system. The evaluation methodology used in this paper is based on supervised evaluation with (manually created) ground-truth data. The experiment aims to compare the performance of proposed approach with the current state of the art of geometrical correction approaches. The experiment is performed with a diverse and representative sample of 24 arbitrarily warped historical documents image with complex layout from the IMPACT project dataset [27]. Baselines on both the original warped document image and result document image are marked manually. The accuracy of each document is estimated by the average baseline straightness of the original and the corrected image is calculated according to equation (2). The results of three additional geometric correction methods are compared: our initial arbitrary dewarping approach by Prima group [2], a

state-of-the-art page-curl correction method designed for IMPACT by NCSR [12] and the leading commercial product Book Restorer™ [28].

In order to evaluate dewarping by measuring the "straightness" of baselines, we use the same evaluation methodology from our previous work [2]. The average percentage of sub cross-area over rectangle-area in each baseline is measured. As shown in Fig.9, the sub cross-area refers to the area of the sub-region shaped by the baseline and average Y line.

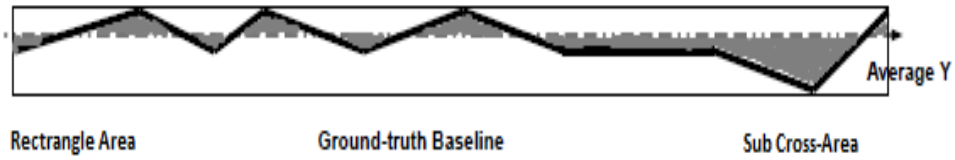


Fig. 9. *Rectangle area and sub cross-area of a baseline*

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 9.

$$\text{Accuracy} = 1 - \frac{\sum_{1}^{N} \frac{\sum_{1}^{M} Sub_{ji}}{\text{Re}\, c_i}}{N} \qquad (9)$$

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.

5.1. Overall accuracy improvement

The accuracy improvement aims to demonstrate the globally geometrical correction performance of proposed approach on distorted historical documents. The experimental results for the test set are shown in Fig.10.

The results show that the both GRLSD system and Prima algorithm improve these 24 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%. Compared to our previous Prima algorithm, GRLSD system with optimization of generated mesh has improved moderately the overall accuracy between 2%-5%. Particularly on some images (ID: 7, 12, 14, 15), GRLSD system has apparent enhancement than Prima algorithm. Only on image 21, GRLS system performs slightly lower accuracy than Prima solution. Overall, the presented method performs better than both the NCSR method and Book Restorer™ on correcting document images with arbitrary warping.

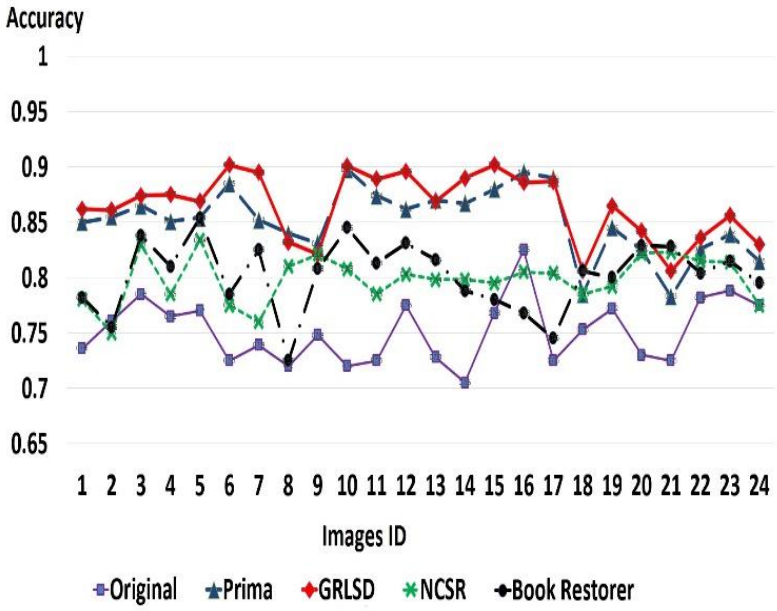


Fig.10. *Evaluation results of NCSR, Book Restorer™ and GRLSD system on arbitrarily warped documents*

Fig. 11 and 12 show some regions from sample image 1 and 11 to illustrate the performance of different dewarping methods on both simple and complex layouts. However, in some images with only slight arbitrary warping (images 2, 6, 10), the NCSR method shows a roughly equivalent improvement to the GRLSD system. Those images suffer mostly from page curl and not so much arbitrary warping. 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 also processes multiple column documents, it mainly solves the problem of page curl and does not perform as effectively on arbitrarily warped documents. Fig.10 and 11 illustrate the GRLSD system and shows the result of the NCSR method for comparison. It can be seen that the proposed system can achieve a better performance than the page curl removal method of NCSR on arbitrarily dewarped document images with both layouts.

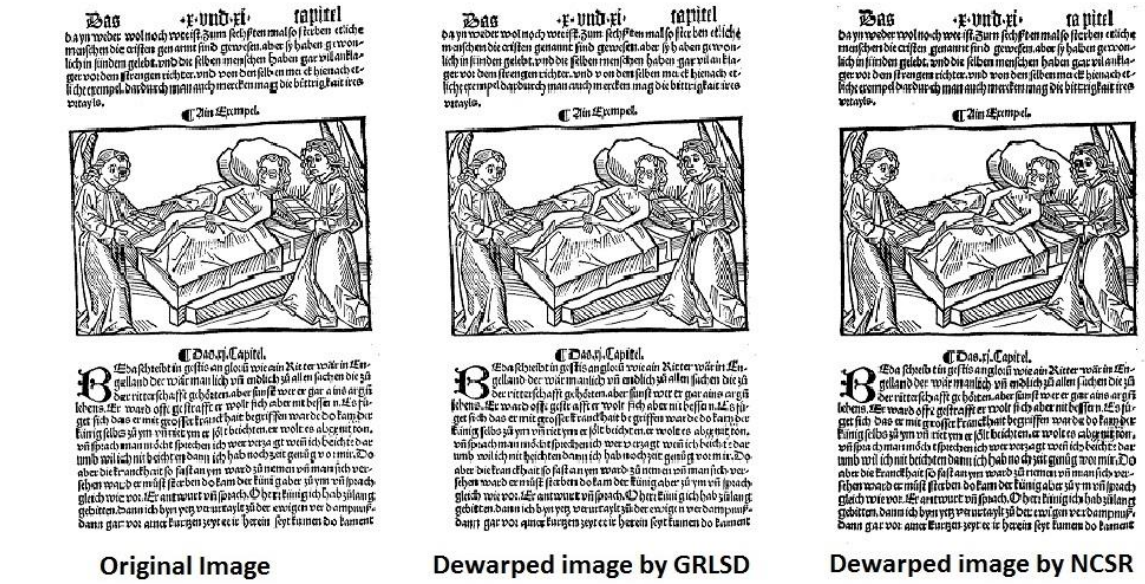


Fig. 11. *Dewarping of a sample image with simple layout*

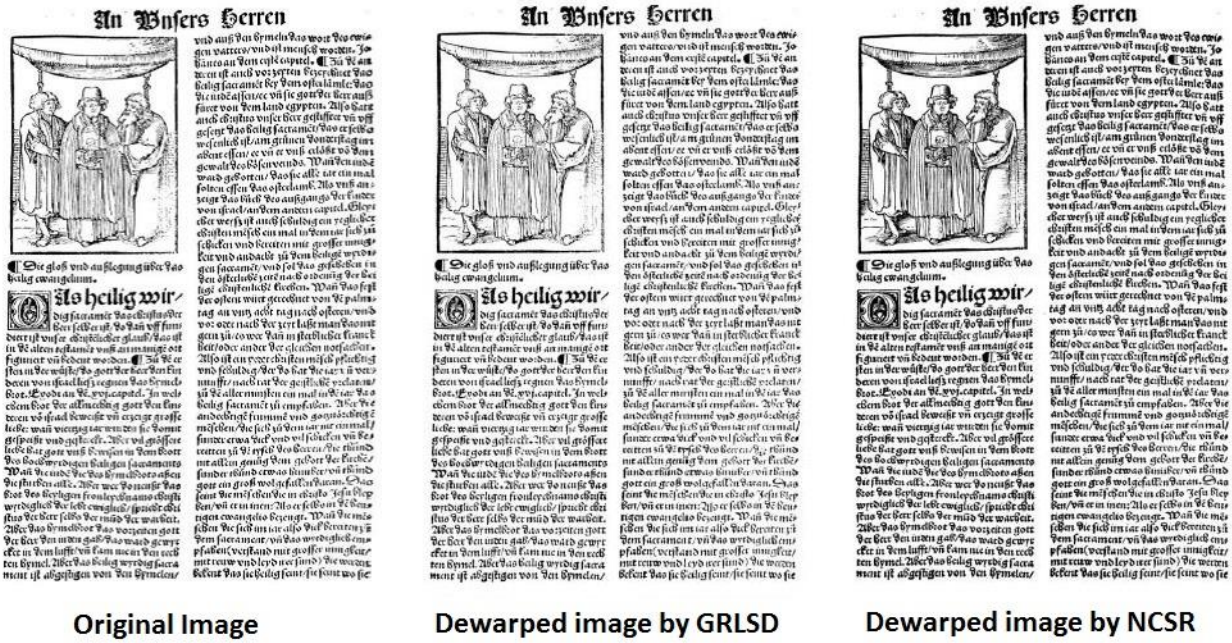


Fig. 12. *Dewarping of a sample image with complex layout*

5.2. Pixel error comparison

The second experiment uses the allocations of points of the marked baselines in Fig.6 to provide a qualitative pixel error comparison among different geometrical correction approaches. This experiment expects to evaluate the proposed approach global correction performance. The experiment is performed with the same sample as last section. Baselines on both the original warped document image and result document image are marked manually. The error of each document image is calculated respectively by standard mean of pixel error (SME), maximum pixel error (MPE), and standard derivation of pixel errors

(STD). Given that there are N baselines being marked in a document image, and each baseline is marked with M points; then there are totally N *M points being allocated in a document image.

For each baseline i, it can calculate the average straight line Y , then the pixel error of each point can be defined in equation 12.

$$PixError_i^j = \left| Y_i^j - Aver\_Y_i \right| \quad (10)$$

Where:

$PixError_i^j$: Pixel error of a marked point j in the baseline i of the mesh.

Y_i^j : Actual value of Y axis of point (i, j).

$Aver\_Y_i$: Average value of Y axis of all points in the baseline i.

Then, SME, MPE and STD of pixels error can be respectively measured and showed in Fig.13. The results show that the GRLSD system in red line can significantly improve these 24 historical document images with arbitrary warping by decreasing the pixel errors, such as SME from an average of 10 pixels to 2 pixels, MPE from an average of 20-30 pixels to 10-20 pixels, STD from an average of 4-6 pixels to 2 pixels.

The Book Restorer™ software in black line and the NCSR method in green line can also decrease the pixel errors, but their performances are not good as the proposed one. In the Book Restorer™ software, SME is an average 3 pixels, MPE is an average of 10-20 pixels. STD is an average of 2-4 pixels. In NCSR cases, SME is an average 4 pixels, MPE is an average of 10-30 pixels. STD is an average of 2-5 pixels. Overall, the presented method performs better than both the NCSR method and Book Restorer™ on correcting document images with arbitrary warping.

Actually, STD and SME reflect a global correction capability of dewarping methods, which is similar to the overall accuracy in last section. MPE reflects a local correction capability of dewarping methods, which demonstrates if the methods generate some individual ambiguous points with worse effect on original document image. Regarding these two aspects, the proposed methods have both better performances than current state-of-the-art approaches.

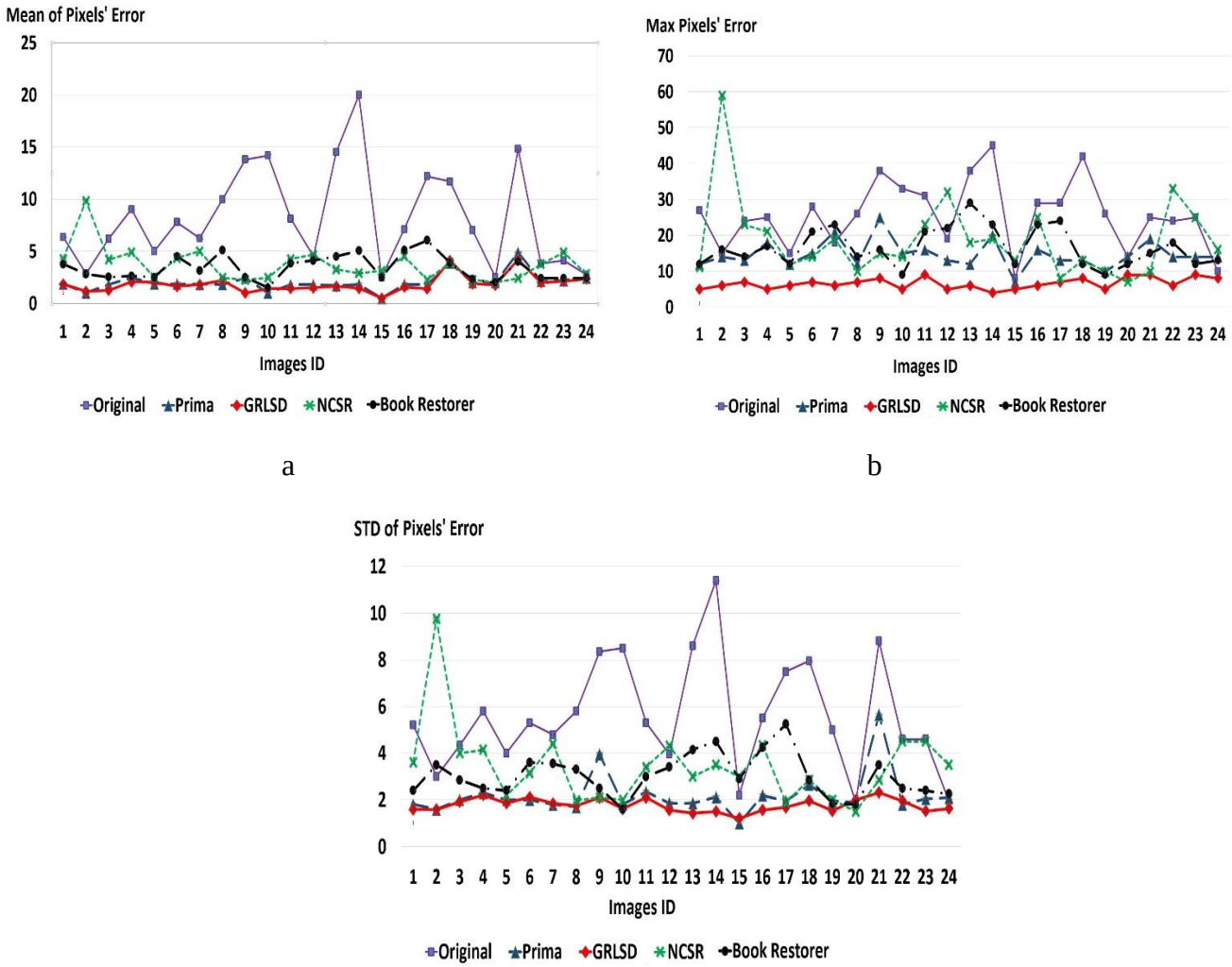
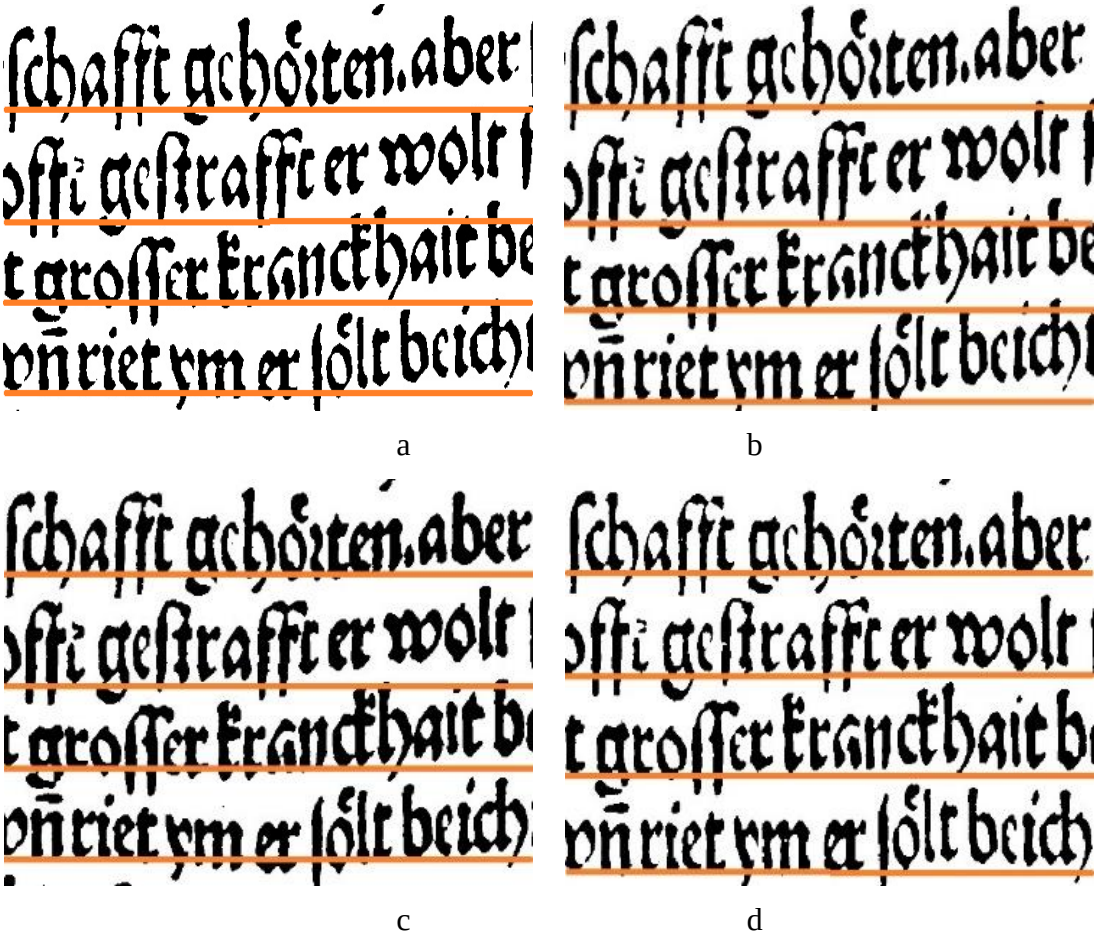


Fig. 13. *Evaluation results of NCSR, Book Restorer™ and proposed method on SME, MPE, SDE. a) Evaluation results on Mean of Pixel's Error. b) Evaluation results on Max of Pixels' Error. c) Evaluation results on STD.*

Fig.14 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 global area affected by arbitrary warping is rather small in the image, the overall accuracies of these three dewarping methods are not significantly different.



a b c d

Fig. 14. *Performance of different correction approaches processing image 6*, a) Arbitrary warping region. b) Correction by Book Restorer^{TM}. c) Correction by NCSR. d) Correction by GRLSD system.

5.3. Local dewarping correction performance

The third experiment is designed to reflect the local correction performance of proposed dewarping method. Regarding as the result of first experiment, for each document image, there are N baselines being marked in a document image, and each baseline is calculated the straightness accuracy; then the original document image and dewarped document image would have the matched straightness accuracy of each baselines. So the improvement of straightness accuracy for each baseline is defined as:

$$Gain\_S\_Line_i = Dewarp\_S\_Line_i - Ori\_S\_Line_i \qquad (11)$$

Where:

$Gain\_S\_Line_i$: Improvement of straightness accuracy for the baseline i of the mesh.

$Dewarp\_S\_Line_i$: Straightness accuracy for the baseline i of the mesh in dewarped image.

$Ori\_S\_Line_i$: Straightness accuracy for the baseline i of the mesh in original image.

Meanwhile, a criterion is defined to check if the baseline is improved. If the improvement of straightness accuracy is over 0.01, it means that the baseline is improved. If the improvement of straightness

accuracy is lower than -0.01, it means that the baseline is getting worse. If the improvement of straightness accuracy is between -0.01 and 0.01, it means that the baseline is kept the same. Then, we would account the number of baselines in a document image has been improved, or kept the same, or worse. The local correction performance of processed dewarping method can be presented by the percentage of them over the total number of baselines.

$$Percentage\_Line\_Improved = Number\_Line\_Improved / N ;$$
$$Percentage\_Line\_Worse = Number\_Line\_Worse / N ;$$
$$Percentage\_Line\_Same = Number\_Line\_Same / N ;$$

The experimental results for the test set are shown in Fig.15. The results show that the percentage of line improved by proposed method can be stably within an average level about 80%. But the percentage of line improved by NCSR and Book Restorer™ are instable within an average level about 70% or 65%. On some images, the percentage of line improved is even lower than 20%. As for the percentage of unchanged and damaged lines, the proposed method can keep it in a stable average level below 10% or 20%. But the NCSR and Book Restorer™ methods generate a higher percentage of unchanged and damaged lines, which is about over 20% and 30%, even with some significantly worse result in image 16. Overall, the presented method performs better than both the NCSR method and Book Restorer™ on correcting document images with arbitrary warping. In Fig. 16, it is seen that the GRLSD system is more advanced in processing arbitrary warping in images.

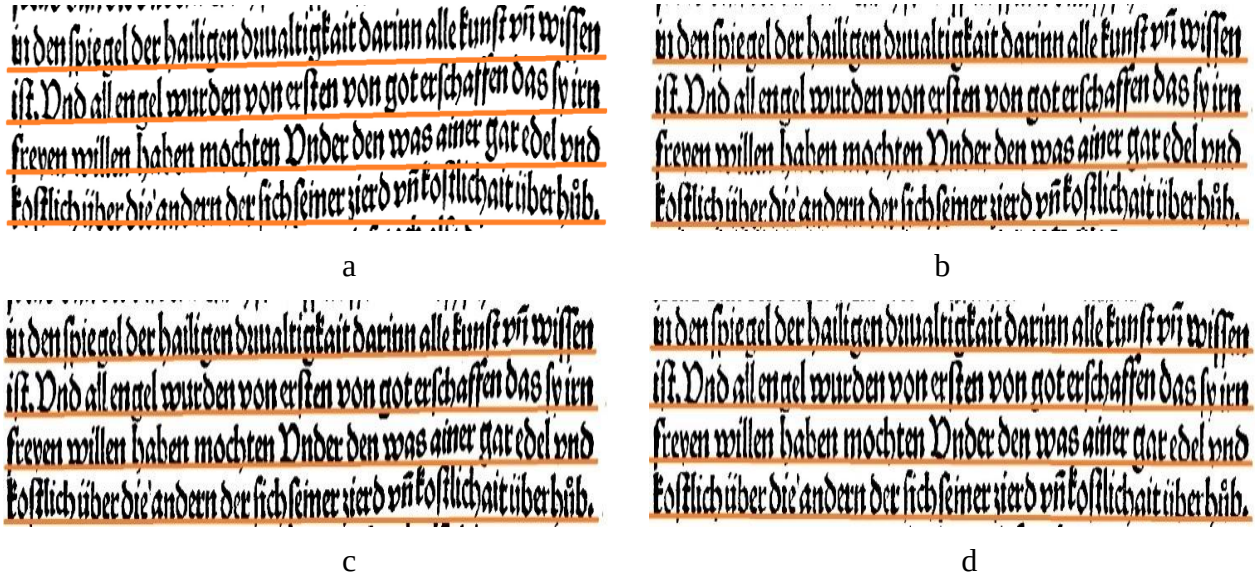


Fig. 15. *Performance of different correction approaches processing image 1, a)* Arbitrarily warped region in original image b) Correction by Book Restorer™. c) Correction by NCSR. d) Correction by GRLSD system

.

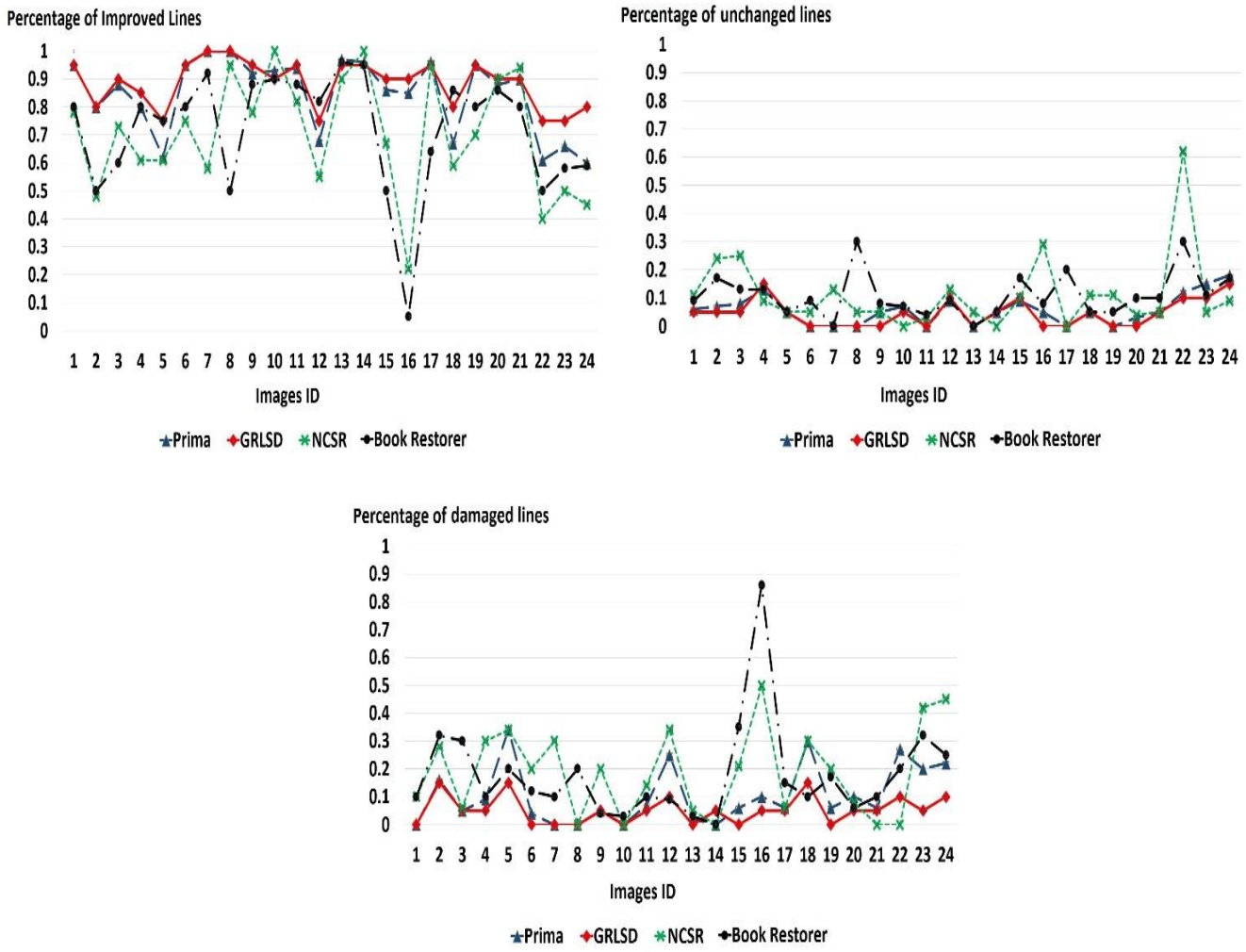


Fig. 16. *Percentage of improved lines, unchanged lines and damaged lines of NCSR, Book Restorer™ and proposed method on SME.* a) Percentage of improved lines. b) Percentage of unchanged lines. *c)* Percentage of damaged lines

6. Conclusion and Future Work

In this paper, an effective mesh based geometric restoration system (GRLSD), for large-scale distorted historical document digitisation is presented. Our experimental results show that the GRLSD system performs better than the leading start-of-the-art geometric correction methods. GRLSD system also enables to process document images with complex contents, such as multiple font size, as shown in Appendix A-C. An advantage of the proposed system is that the transformation (correction) is reversible – a major requirement of the libraries (to be able to go back to the original master scans). It greatly aids users to flexibly monitor and correct ambiguous points of mesh for the prevention of damaging historical document images without distortions in large-scale digitalisation. The method in its current state is only evaluated by book images; it has yet to be evaluated on more challenging historical documents such as newspapers, as shown in Appendix D.

Future work will firstly focus on investigating more advanced text region segmentation method to process historical document on a large-scale. As shown in Appendix D, the text region segmentation method fails to process high resolution newspaper image with extremely complex layout, further resulting in a poor dewarping performance. Secondly, this work will attempt to evaluate the auto-dewarped historical document image by accessing their OCR performance. It is challenging since most commercial OCR software may not support historical fonts so far, thus it requires specifically-designed OCR software to access its performance. Finally, we will attempt to investigate utilisation of GPU techniques for accelerating the auto-dewarping algorithms in processing large-scale document images.

7. 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] P. Yang, A. Antonacopoulos, C. Clausner, S. Pletschacher, "Grid-based modelling and correction of arbitrarily warped historical document images for large-scale digitization", in: *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, *HIP '11*, ACM, Beijing, China, pp. 106–111, Sep. 2011.

[3] 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.

[4] 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.

[5] S. J. Lu, and C. L. Tan., "Document flattening through grid modelling and regularization", *Int'l Conference on Pattern Recognition*, pp. 971-980, 2006.

[6] Z. Zhang and C.L. Tan, "Restoration of Document Images Scanned from Thick Bound Document," *Proc. Int'l Conf. Image Processing*, pp. 1074-1077, Oct. 2001

[7] Z. Zhang and C.L. Tan, "Recovery of Distorted Document Image from Bound Volumes," *Proc. Int'l Conf. Document Analysis and Recognition*, pp. 429-433, 2001.

[8] Z. Zhang and C.L. Tan, "Straightening Warped Text Lines Using Polynomial Regression," *Proc. Int'l Conf. Image Processing*, pp. 977-980, Sept, 2002.

[9] W. Wu, R. Li, B. Fu, W. Li, and Z. Xu., "A Model Based Book Dewarping Method to Handle 2D Images Captured by a Digital Camera", *Proc. Int'l Conference on Document Analysis and Recgonition*, pp. 158-162, 2007.

[10] 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.

[11] S. Bukhari, F. Shafait, and T. Breuel, "Dewarping of document image using coupled-snakes", *Int. Workshop on Camera-Based Document Analysis and Recognition*, July, 2009.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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*, pp. 482-485, 2004.

[19] L.Zhang, Y. Zhang, and C.L.Tan. “An Improved physically-based method for geometric restoration of distorted document images” *IEEE Transactions on Patten Analysis and Machine Intelligence*, Vol 30, issue 4, pp728-734, 2008.

[20] L. O’Gorman, “The Document Spectrum for Page Layout Analysis.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 15, pp1162–1173, 1993.

[21] N. Stamatopoulos, B. Gatos, I. Pratikakis., and S.J. Perantonis, “Performance evaluation methodology for document image dewarping techniques”, *IET Image Processing*, Vol 6, Issue 6, pp738-745, 2012.

[22] S. Pletschacher, A. Antonacopoulos, “The PAGE (Page Analysis and Ground-Truth Elements) Format Framework”, Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010), Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260.

[23] G.F.Meng, C.H. Pan, S. M. Xiang, J. Y. Duan and N.N.Zheng. “Metric Rectification of Curved Document Images” *IEEE Transactions on Patten Analysis and Machine Intelligence*, Vol 34, issue 4, pp707-722, 2012.

[24] V. Papavassiliou, V. Kartsouros, G. Carayannis, “A Morphological Approach for Text-Line Segmentation in Handwritten Documents”, *Int’l Conference on Frontiers in Handwriting Recognition* 2010, pp 19-24, 2010.

[25] R.P. Dos Santos, G.S. Clemente, I.R. Tsang, G.D.C, Cavalcanti, “Text Line Segmentation Based on Morphology and Histogram Projection”, *Int’l Conference on Document Analysis and Recognition 2009*, pp651-657, 2009**.**

[26] P.V.C. Hough, Methods and means for recognizing complex patterns, U.S. Patent No 3069654, 1962.

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

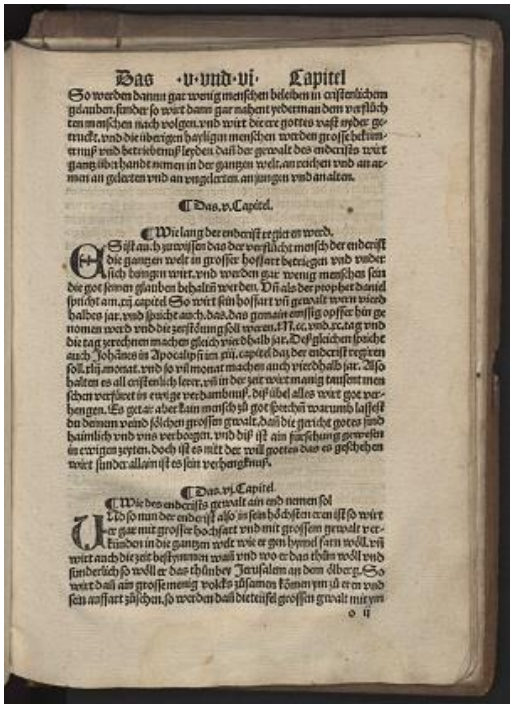
[28] Book Restorer, image restoration software, http://www.i2s-bookscanner.com

[29] C. Clausner, A. Antonacopoulos, S. Pletschacher “A robust hybrid approach for text line segmentation in historical documents”, 2012 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, Nov11-15, 2012, IEEE-CS Press, pp. 335-338.

[30] ENP: Europeana Newspapers Project. http://www.europeana-newspapers.eu/.

[31] P. Yang, B.Q.Liu, D. Williams, V. Codreanu, B. Mahdian, X. Zhao, J. Roerdink, J. Gordon, F. Dong, “GSWO: A programming model for GPU-enabled parallelization of sliding window operations in image processing”, *Signal Processing: Image Communication*, Vol 47, pp.332-345, 2016.

Appendix A: Sample 1 of dewarped historical document images



Das .v. vnd .vj. Capitel

So werden dann gar wenig menschen beleiben in cristenlichem
gläuben, sunder so wirt dann gar nahent yederman dem verflüch
ten menschen nach volgen, vnd wirt die ere gottes vast nyder ge
truckt, vnd die übrigen hayligen menschen werden grosse beküm
ernuß vnd betrübnuß leyden, daß der gewalt des endcrists wirt
gantz über handt nemen in der gantzen welt, an reichen vnd an ar
men an gelarten vnd an vngelarten, an jungen vnd an alten.

Das .v. Capitel.

Wie lang der endcrist regieren werd.

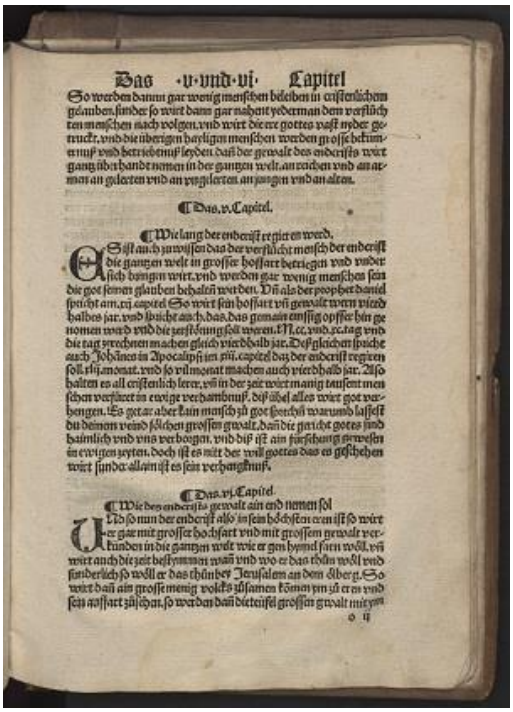
Gist auch zu wissen das der verflüchet mensch der endcrist
die gantzen welt in grosser hoffart betriegen vnd vnder
sich bringen wirt, vnd werden gar wenig menschen sein
die got seinen gläuben behalten werden. Dis als der prophet daniel
spricht am .xj. capitel So wirt sein hoffart vil gewalt wern vierd
halbes jar, vnd spricht auch, das das gemain ewssig opffer hin ge
nomen werd vnd die zerstörung sol weren .M.cc. vnd .xc. tag vnd
die tag yrechnen machen gleich vierdhalb jar. Desgleichen spricht
auch Johänes in Apocalipsi im .xij. capitel, das der endcrist regie
ren sol .xlij. monat, vnd so vil monat machen auch vierdhalb jar. Also
halten es all cristenlich leren, wie in der zeit wirt manig tausent men
schen verfüret in ewige verdambnuß. Disz übel alles wirt got ver
hengen. Es get ar aber kain mensch zu got herein warumb lassest
du deinen veind sölchen grossen gewalt, dann die gericht gottes sind
haimlich vnd vns verborgen, vnd disz ist ain forsehung gewesen
in ewigen zeyten, doch ist es mit der will gottes das es geschehen
wirt sunder allein ist es sein verhengknuß.

Das .vj. Capitel.

Wie des endcrists gewalt ain end nemen sol

Vnd so nun der endcrist also in sein höchsten ern ist so wirt
er gar mit grosser hochfart vnd mit grossem gewalt ver
künden in die gantzen welt wie er gen hymel faren wöll, vil
wirt auch die zeit besynnen wann vnd wo er das thün wöll vnd
sunderlich so wöll er das thünder Jerusalem an dem ölberg. So
wirt dann ain grosse menig volcks zusamen kömen ym zu er en vnd
sein auffart zu schen, so werden dann die teüfel grossen gewalt mit ym

o ij



Das .v. vnd .vj. Capitel

So werden dann gar wenig menschen beleiben in cristenlichem
gläuben, sunder so wirt dann gar nahent yederman dem verflüch
ten menschen nach volgen, vnd wirt die ere gottes vast nyder ge
truckt, vnd die übrigen hayligen menschen werden grosse bekum
ernuß vnd betrübnuß leyden, daß der gewalt des endcrists wirt
gantz über handt nemen in der gantzen welt, an reichen vnd an ar
men an gelarten vnd an vngelarten, an jungen vnd an alten.

Das .v. Capitel.

Wie lang der endcrist regieren werd.

Gist auch zu wissen das der verflüchet mensch der endcrist
die gantzen welt in grosser hoffart betriegen vnd vnder
sich bringen wirt, vnd werden gar wenig menschen sein
die got seinen gläuben behalten werden. Dis als der prophet daniel
spricht am .xj. capitel So wirt sein hoffart vil gewalt wern vierd
halbes jar, vnd spricht auch, das das gemain ewssig opffer hin ge
nomen werd vnd die zerstörung sol weren .M.cc. vnd .xc. tag vnd
die tag yrechnen machen gleich vierdhalb jar. Desgleichen spricht
auch Johänes in Apocalipsi im .xij. capitel, das der endcrist regie
ren sol .xlij. monat, vnd so vil monat machen auch vierdhalb jar. Also
halten es all cristenlich leren, wie in der zeit wirt manig tausent men
schen verfüret in ewige verdambnuß. Disz übel alles wirt got ver
hengen. Es get ar aber kain mensch zu got herein warumb lassest
du deinen veind sölchen grossen gewalt, dann die gericht gottes sind
haimlich vnd vns verborgen, vnd disz ist ain forsehung gewesen
in ewigen zeyten, doch ist es mit der will gottes das es geschehen
wirt sunder allein ist es sein verhengknuß.

Das .vj. Capitel.

Wie des endcrists gewalt ain end nemen sol

Vnd so nun der endcrist also in sein höchsten ern ist so wirt
er gar mit grosser hochfart vnd mit grossem gewalt ver
künden in die gantzen welt wie er gen hymel faren wöll, vil
wirt auch die zeit besynnen wann vnd wo er das thün wöll vnd
sunderlich so wöll er das thünder Jerusalem an dem ölberg. So
wirt dann ain grosse menig volcks zusamen kömen ym zu er en vnd
sein auffart zu schen, so werden dann die teüfel grossen gewalt mit ym

o ij

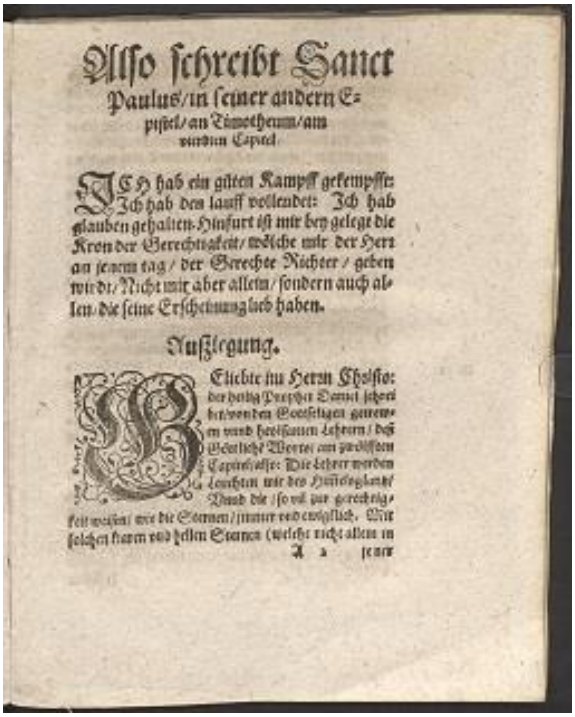
a) Original image **b) Dewarped Image**

Appendix B: Sample 2 of dewarped historical document images

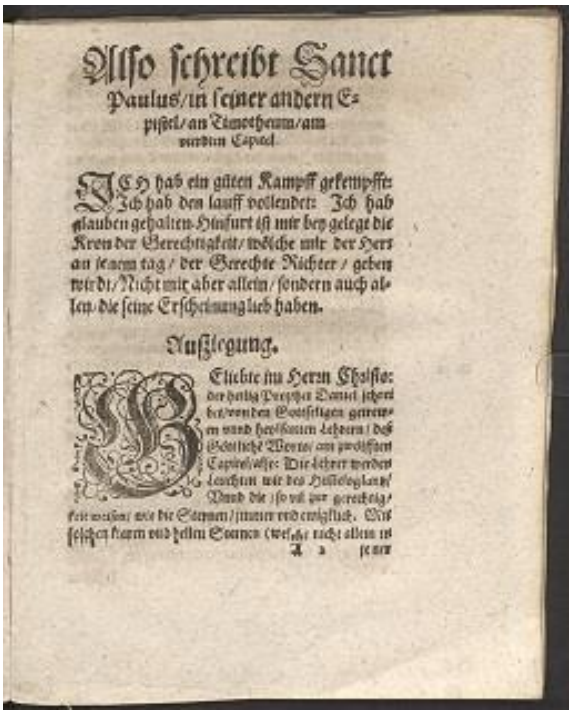
a) Original image

b) Dewarped Image

Appendix C: Sample 3 of dewarped historical document images

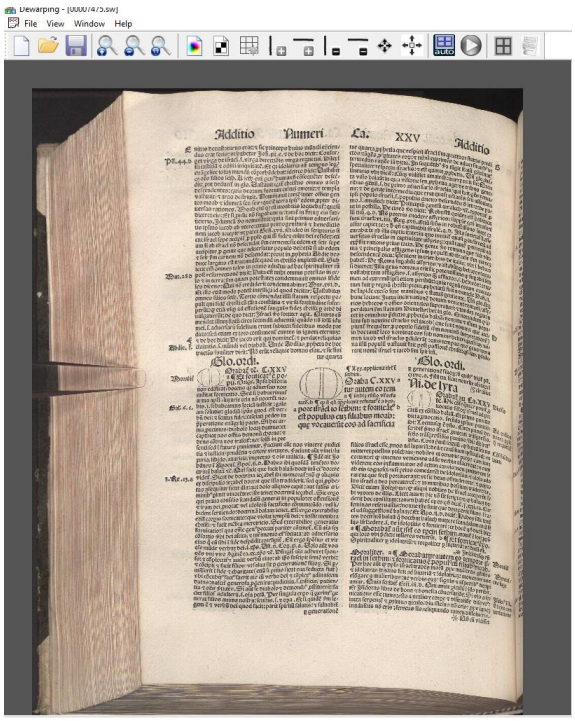


a) Original image

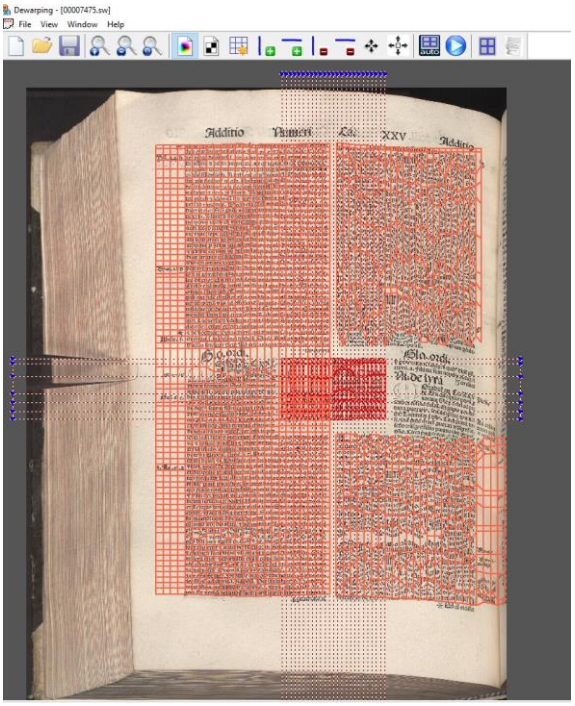


b) Dewarped Image

Appendix D: Sample 4 of challenging historical document images



a) Original image



b) Auto-generated meshes