RESTORATION OF ARBITRARILY WARPED DOCUMENT IMAGES BASED ON TEXT LINE AND WORD DETECTION

B. Gatos[1], K. Ntirogiannis[2]

[1]Computational Intelligence Laboratory
Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
GR-153 10 Agia Paraskevi, Athens, Greece
http://www.iit.demokritos.gr/cil/
bgat@iit.demokritos.gr

[2]Department of Informatics and Telecommunications
University of Athens
Athens, Greece
std01081@di.uoa.gr

ABSTRACT
This paper presents a novel technique for efficient restoration of arbitrarily warped document images. Our aim is to recover document images that are mainly bounded volumes captured by a digital camera and suffer from non-linear warp. The proposed technique is applied on gray scale document images and is based on several distinct steps: an adaptive document image binarization, a text line and word detection, a first draft binary image dewarping based on word rotation and shifting and, finally, a complete restoration of the original grayscale warped image guided by the binary dewarping result. In this paper, we present a detailed description of the proposed technique as well as the implementation results for each step of our methodology. The experimental results on several arbitrarily warped documents indicate the effectiveness of the proposed technique.

KEY WORDS
Document image dewarping, document image restoration, degraded document images, document pre-processing, text line detection, document image analysis

1. Introduction

Document images often suffer from non-linear warping when captured by a digital camera or a scanner, especially when these documents are digitized bounded volumes (see Fig. 1). Warping not only reduces the document readability but also affects the accuracy of an OCR application.

Several techniques have been proposed for correcting the document image warping and they can be classified in two main categories: (i) 2D image processing techniques ([1], [2], [3], [4], [5]) and (ii) techniques where the 3D document shape is discovered ([6], [7], [8]). The state-of-the-art background of our work is mainly focused on the first category of techniques since the second category usually involves image capture with special camera setup as well as document surface representation by using a 3D shape model. The latter is very difficult to accomplish when processing arbitrarily warped documents with several slope changes along the text lines as well as along the words of the same text line (see Fig. 1). In order to recover such documents, we propose a novel technique that is applied on grayscale document images and is based on (i) an adaptive document image binarization, (ii) a text line and word detection, (iii) a first draft binary image dewarping based on word rotation and shifting and, (iv) a complete restoration of the original grayscale warped image guided by the binary dewarping result.

In the following sections, we present the dewarping techniques based on 2D Image Processing, a detailed description of the proposed technique as well as experimental results on several arbitrarily warped documents that demonstrate the efficiency of the proposed technique.

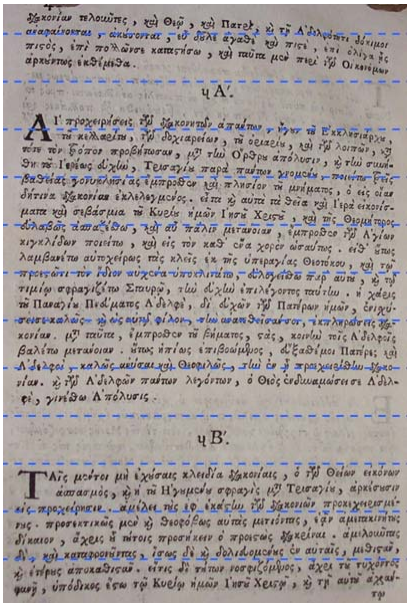


Figure 1. Example of an arbitrary warped document image.

2. Dewarping based on 2D Image Processing

Approaches in this category use 2D image processing in order to restore the warped image. In [1], a deformable system to straighten curved text image is presented. Restoration is accomplished by using an active contour network based on an analytical model with cubic B-splines which have been proved more accurate than Bezier curves. A model fitting technique has also been proposed using cubic splines to define the warping model of the document image [2]. For more accurate dewarping, a vertical division of a document image into some partial document images is also suggested. Another model fitting technique [3] divides the document image into shaded and non-shaded region and then uses polynomial regression to model the warped text lines with quadratic reference curves. In [4], the texture of a document image is calculated so as to infer the document structure distortion. A mesh of the warped image is built using a non-linear curve for each text line. The curves are fitted to text lines by tracking the character boxes on the text lines. The erroneously fitted curves are detected and excluded by a post processing based on several heuristics. The approach of [5] relies on a priori layout information and is based on a line-by-line dewarping of the observed paper surface. Each letter in the input image is enclosed within a quadrilateral cell, which is then mapped to a rectangle of correct size and position in the result image.

3. The proposed approach

In our approach, we face the problem of restoring an arbitrarily warped grayscale document image by following several distinct steps explained in detail at this section. The main novelties of our approach are the following: (i) we first proceed to a draft binary image dewarping and then to a complete restoration of the original grayscale warped image guided by the binary dewarping result; (ii) binary image dewarping is accomplished by word rotation and shifting based on the lower and upper word baselines; (iii) we propose a novel modified "box hands" method for efficient text line and word detection in warped documents.

3.1 Document image binarization

The original warped grayscale image I_g is defined as follows:

$$I_g(x,y) = (0,1,\ \ldots.\ 255) \tag{1}$$

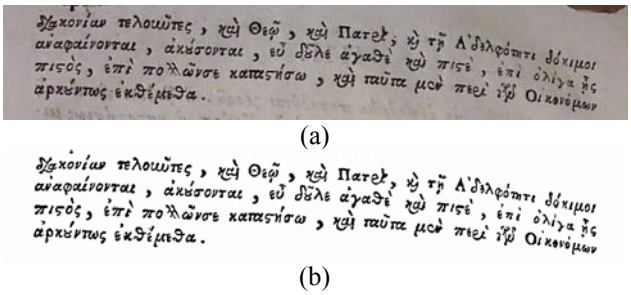
where $x \in [1, x_{\max}], y \in [1, y_{\max}]$.

Binarization is the starting step of most document image analysis systems and refers to the conversion of the gray-scale image to a binary image. The proposed scheme for image binarization and enhancement is based on the work of [9] and consists of five distinct steps: a pre-processing procedure using a low-pass Wiener filter, a rough estimation of foreground regions, a background surface calculation by interpolating neighboring background intensities, a thresholding by combining the calculated background surface with the original image and finally a post-processing step that improves the quality of text regions and preserves stroke connectivity. Fig. 2 illustrates the document image binarization step.

After following this step, we extract the warped Binary image I_b:

$$I_b(x,y) = (0,1) \tag{2}$$

where $x \in [1, x_{\max}], y \in [1, y_{\max}]$



(a)

(b)

Figure 2. (a) The original warped grayscale image and (b) the resulting binary image.

3.2 Text line and word detection

Our method for text line and word detection is based on connected component labelling followed by a modified "box hands" method.

For the connected component labeling method we follow the two pass approach of [10]. During this process, we compute the bounding boxes of the connected components (see Fig. 3) and, furthermore, calculate the dominant letter height H_{max} which corresponds to the maximum in the bounding boxes heights histogram.

Figure 3. Applying connected component labeling at a warped binary image.

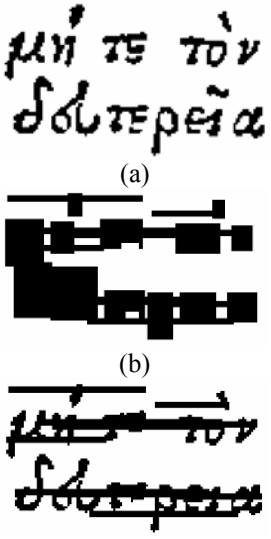
Both in text line and word detection algorithm, only connected components of height h_i satisfying the following condition are participating:

$$\frac{H_{\max}}{2} \le h_i \le 2H_{\max} \tag{3}$$

According to the "box-hand" text line detection approach ([11],[12]) the bounding boxes of all connected components are extended to give chains of connected boxes. As shown in Fig. 4b, bounding boxes are extended by adding to its left and right sides two equal parallel-

quadrilateral extensions, called the "box hands". This approach can not efficiently detect text lines in a warped document since it assumes that the text line is a straight horizontal line and the document is not warped. Our proposed modified "box hand" method for detecting text lines and words is based (i) on adding "box hands" not to the bounding boxes of the connected components but to the original connected components and (ii) on defining suitable "box hand" parameters in order to achieve either text line or word detection. Modification (i) has been added in order to manipulate historical or handwritten documents which have connected components of great width. In these cases, the bounding boxes of these connected components are likely to overlap with connected components of adjacent text lines or words (see Fig. 4b). According to our approach, we add "box hands" to the original connected components (see Fig. 4c) and the parameter h_1 of the "box hands" given in Fig. 5 is calculated as follows:

$$h_1 = \begin{cases} 2H_{\max}, \text{for text line detection} \\ \dfrac{H_{\max}}{2}, \text{for word detection} \end{cases} \quad (4)$$



(a)

(b)

(c)

Figure 4. (a) The binary warped image; (b) the result after applying the "box hand" method and (c) the result after applying the proposed modified "box hand" method.

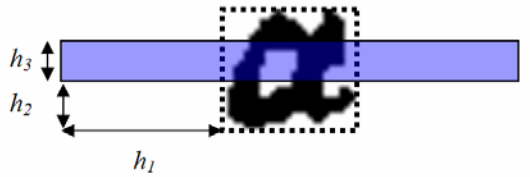


Figure 5. The modified "box hands" defined by the parameters h_1 and h_2 and h_3.

After the application of the modified "box hand" the text lines images are defined as follows:

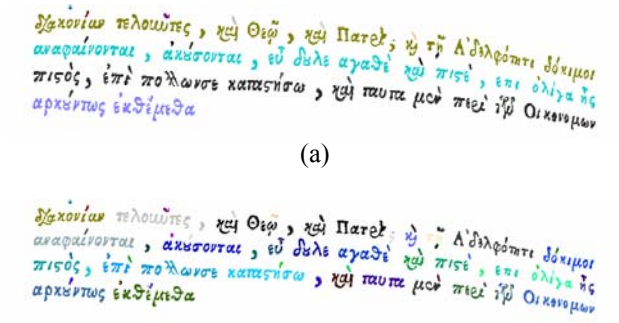
$$L_i(x,y) = \begin{cases} 1, \text{if } I_b(x,y) = 1 \text{ AND } (x,y) \in \text{Text Line } i \\ 0, \text{otherwise} \end{cases} \quad (5)$$

where $i \in [0, l_{\max} - 1]$, $l_{\max}$ is the total number of text lines found. All text lines are top-to-down sorted.

Similarly, words images are defined according:

$$W_{ij}(x,y) = \begin{cases} 1, \text{if } I_b(x,y) = 1 \text{ AND } (x,y) \in \\ \qquad \text{Word j of Text Line } i \\ 0, \text{otherwise} \end{cases} \quad (6)$$

where $i \in [0, l_{\max} - 1]$, $j \in [0, w^i_{\max} - 1]$, $w^i_{\max}$ is the total number of words found in text line *i*. All words are left-to-right sorted in every text line. Text line and word detection results are demonstrated in Fig. 6.



(a)

(b)

Figure 6. (a) Text line detection and (b) word detection results using the proposed modified "box hand" method.

3.3 Word lower and upper baseline estimation

At this step, we detect the lower and upper baselines which delimit the main body of the words (see Fig. 7).

For lower and upper baseline detection we follow the methodology given in [13] which is used for lower baseline detection. According to this approach, a linear regression is applied on the set of points that are the lowest black pixels for each text line column. In our approach, we also use a similar procedure to calculate the upper baseline.

After this procedure, upper baseline of word W_{ij} is defined from the equation:

$$y = a_{ij}x + b_{ij} \quad (7)$$

Similarly, lower baseline of word W_{ij} is defined as:

$$y = a'_{ij}x + b'_{ij} \quad (8)$$

Figure 7: Upper and lower baselines example.

3.4 Draft estimation of the binary dewarped image

At this stage, all detected words are rotated and shifted in order to obtain a first draft estimation of the binary dewarped image.

The slope of each word is derived from the corresponding baseline slopes. Upper and lower baseline slopes θ_{ij}^{u} and θ_{ij}^{l} of word W_{ij} can be calculated by the formulas:

$$\theta_{ij}^{u} = \arctan(a_{ij}) \tag{9}$$

and

$$\theta_{ij}^{l} = \arctan(a_{ij}^{'}) \tag{10}$$

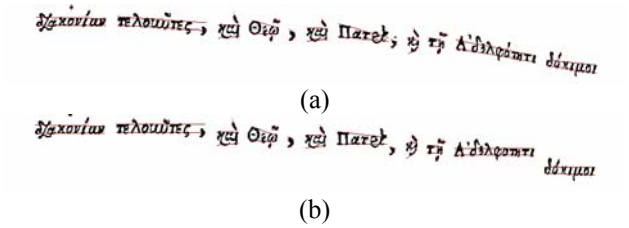
Since the smaller slope is usually the most representative, the word's slope can be defined as:

$$\theta_{ij} = \begin{cases} \theta_{ij}^{u}, \text{if } \left|\theta_{ij}^{u}\right| < \left|\theta_{ij}^{l}\right| \\ \theta_{ij}^{l}, otherwise \end{cases} \tag{11}$$

The rotation of the word W_{ij}(x,y) is done as follows:

$$\left.\begin{array}{c} y^{r} = (x - x_{\min}) * \sin(-\theta_{ij}) + y * \cos(\theta_{ij}) \\ x^{r} = x \end{array}\right\} \tag{12}$$

where $W^{r}_{ij}(x^{r}, y^{r})$ is the rotated word and x_{min} is the left side of the bounding box of the word W_{ij}. An example of correcting the skew of the words is given in Fig. 8.



(a)

(b)

Figure 8. Example of the word skew correction: (a) original image having the word upper and lower baselines marked; (b) resulting image after word skew correction.

After word rotation, all the words of every text line, except from the leftmost, must be vertically shifted in order to restore horizontal alignment. The rotation and shifting of the word W_{ij}(x,y) is done as follows:

$$\left.\begin{array}{c} y^{rs} = (x - x_{\min}) * \sin(-\theta_{ij}) + y * \cos(\theta_{ij}) + d_{ij} \\ x^{rs} = x \end{array}\right\} \tag{13}$$

where $W^{rs}_{ij}(x^{r}, y^{r})$ is the rotated and shifted word. d_{ij} corresponds to the vertical word shifting and is given by the following formula:

$$d_{ij} = \begin{cases} y_{i0}^{rl} - y_{ij}^{rl}, \text{if } \left|\theta_{ij}^{l}\right| < \left|\theta_{ij}^{u}\right| \\ y_{i0}^{ru} - y_{ij}^{ru}, \text{otherwise} \end{cases} \tag{14}$$

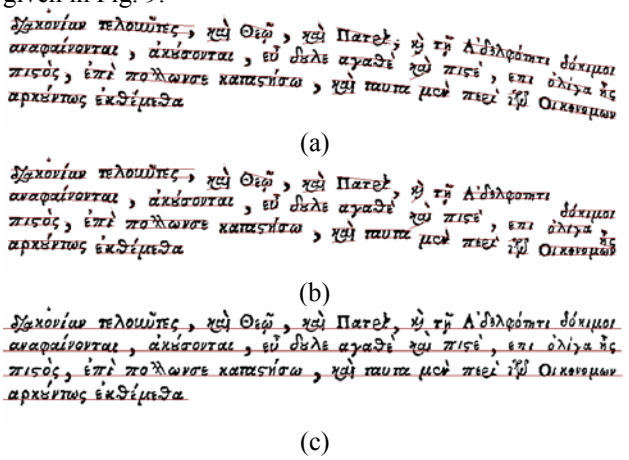
where:

$$y_{ij}^{rl} = (a_{ij} x_{\min} + b_{ij}) * \cos(\theta_{ij}) \tag{15}$$

and

$$y_{ij}^{ru} = (a_{ij}^{'} x_{\min} + b_{ij}^{'}) * \cos(\theta_{ij}) \tag{16}$$

The reason for defining two shifting levels for each text line, is that each word may be rotated either according to its lower baseline or upper baseline slope. Hence, it has to be shifted so that its lower or upper baseline is aligned with the lower or upper baseline of the leftmost word of the text line. An example of binary image dewarping is given in Fig. 9.



(a)

(b)

(c)

Figure 9. Binary image dewarping example: (a) original image having the word upper and lower baselines marked; (b) resulting image after word rotation and (c) resulting image after word rotation and shifting.

In order to construct the binary dewarped image, we follow the procedure below:

Step 1: Dewarped binary image initialization:

$$I_{b\_dew}(x, y) = 0,\ x \in [1, x_{\max}], y \in [1, y_{\max}] \tag{17}$$

Step 2: Transformation factors initialization:

$$T_{xy} = \text{NULL}, \Theta_{xy} = \text{NULL}, x \in [1, x_{\max}], y \in [1, y_{\max}] \tag{18}$$

Step 3: Dewarped binary image calculation:

$$\begin{array}{l} \forall (x, y) : W_{ij}(x, y) = 1 \Rightarrow \\ I_{b\_dew}(x, y^{'}) = 1, \\ \text{where } y^{'} = (x - x_{\min}) * \sin(-\theta_{ij}) + y * \cos(\theta_{ij}) + d_{ij} \\ \text{AND } T_{xy^{'}} = d_{ij} \text{ AND } \Theta_{xy^{'}} = \theta_{ij} \end{array} \tag{19}$$

where d_{ij} is defined in eq. 14. Transformation factors T_{xy}, Θ_{xy} are used for the final restoring procedure which is described in the section below.

3.5 Restoration of the grayscale warped image

At this stage, we proceed to a complete restoration of the original grayscale warped image guided by the binary

dewarping result of the previous stage. Since the transformation factors for every pixel in the final binary dewarped image have been already stored, the reverse procedure is applied on the grayscale pixels in order to retrieve the final grayscale dewarped image. For all pixels that do not have transformation factors allocated, the transformation factors of the nearest pixel are used.

The following steps are used:

Step 1: Dewarped grayscale image initialization:

$$I_{g\_dew}(x,y)=0 \tag{20}$$

for $x\in[1,x_{\max}], y\in[1,y_{\max}]$

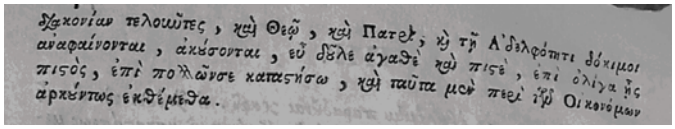
Step 2: Replace NULL values of the transformation factors:

$$\begin{aligned} &if\ T_{xy}=NULL \text{ AND } \Theta_{xy}=NULL \Rightarrow \\ &T_{xy}=T_{x'y'} \text{ AND } \Theta_{xy}=\Theta_{x'y'} \\ &\text{where } (x-x')^2+(y-y')^2=\min \\ &\text{for } x',y': \\ &T_{x'y'} <> NULL \text{ AND } \Theta_{x'y'} <> NULL \end{aligned} \tag{21}$$

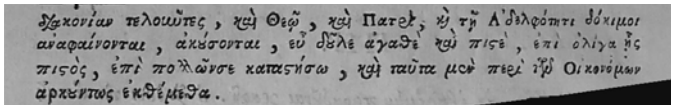
Step 3: Final dewarped grayscale image calculation:

$$\begin{aligned} &I_{g\_dew}(x,y)=I_g(x,y'), \\ &\text{where } y'=\frac{y-T_{xy}-(x-x_{\min})sin(-\Theta_{xy})}{\cos(\Theta_{xy})} \end{aligned} \tag{22}$$

An example of the grayscale warped image restoration procedure is given in Fig. 10.



(a)

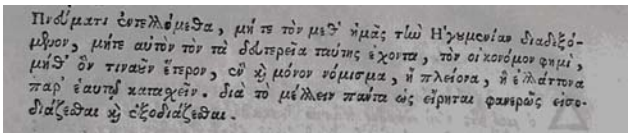


(b)

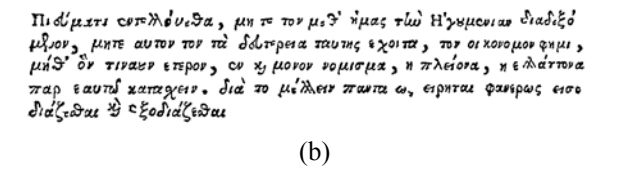
Figure 10. Grayscale image dewarping example: (a) original image and (b) resulting grayscale image dewarping.

4. Experimental results

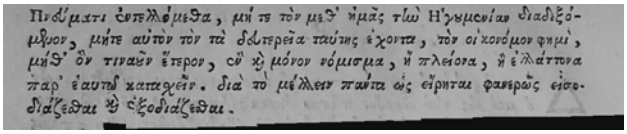
The proposed algorithm was tested using several arbitrarily warped documents. We mainly focused in handwritten historical arbitrarily warped documents. Three representative results are shown in Fig.11-13 .



(a)

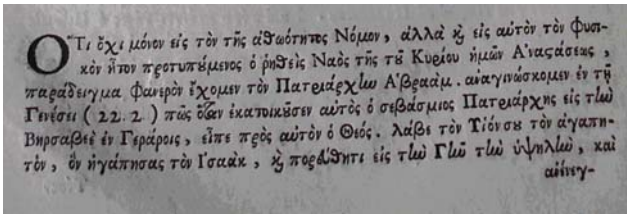


(b)

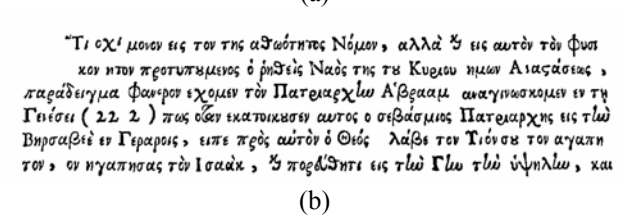


(c)

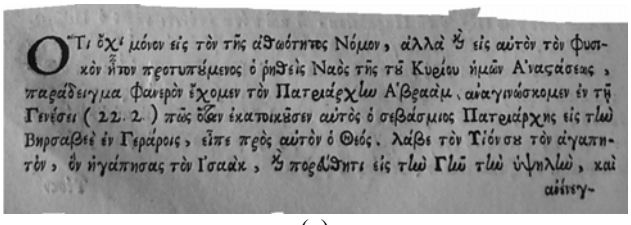
Figure 11. Example 1: (a) Original image; (b) dewarped binary image and (c) final grayscale dewarped image.



(a)

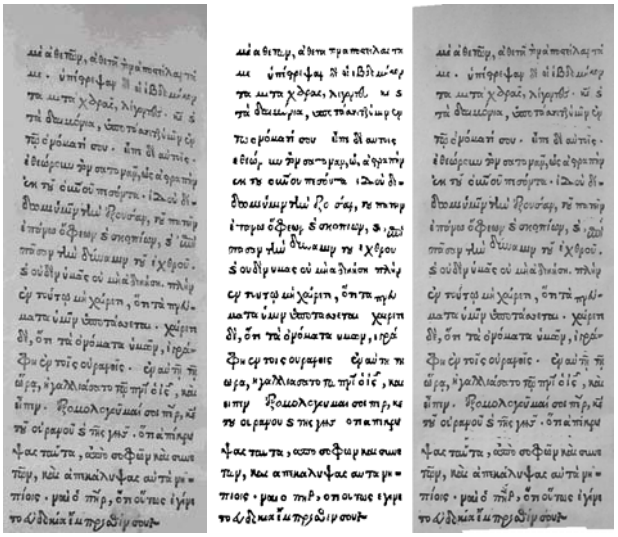


(b)



(c)

Figure 12. Example 2: (a) Original image; (b) dewarped binary image and (c) final grayscale dewarped image.



(a) (b) (c)

Figure 13. Example 3: (a) Original image; (b) dewarped binary image and (c) final grayscale dewarped image.

As it is demonstrated, the experimental results indicate the effectiveness of the proposed technique. We have observed that some problems appear in our experiments mainly due to erroneous baseline detection.

5. Conclusion

In this paper we present a novel methodology for efficient restoration of arbitrarily warped document images. The proposed scheme is applied on grayscale document images and is based on (i) an adaptive document image binarization, (ii) a text line and word detection, (iii) a first draft binary image dewarping based on word rotation and shifting and, (iv) a complete restoration of the original gray scale warped image guided by the binary dewarping result. The experimental results on several arbitrarily warped documents, mainly handwritten historical documents, indicate the effectiveness of the proposed technique. After observing some problems that occurred, we focus our future work on developing a more efficient baseline detection algorithm.

Acknowledgements

This research is carried out within the framework of the Greek Ministry of Research funded R&D project POLYTIMO [14] which aims to process and provide access to the content of valuable historical books and handwritten manuscripts.

References

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

[2] H. Ezaki, S. Uchida, A. Asano & H. Sakoe, Dewarping of document image by global optimization. *Proc. ICDAR'05*, 2005, 500-506.

[3] Z. Zhang & C. L. Tan, Correcting document image warping based on regression of curved text lines. *Proc. ICDAR'03*, 2003, 589–593.

[4] C. Wu & G. Agam, Document image de-warping for text/graphics recognition. *SSPR&SPR 2002*, LNCS 2396, 2002, 348-357

[5] A. Ulges, C.H. Lampert & T.M. Breuel, Document image dewarping using robust estimation of curled text lines. *Proc. ICDAR'05*, 2005, 1001- 1005.

[6] C.L. Tan, L. Zhang, Z. Zhang & T. Xia, Restoring Warped Document Images through 3D Shape Modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence 28*(2), 2006, 195-208.

[7] M.S. Brown & W.B. Seales, Image Restoration of Arbitrarily Warped Documents. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *26*(10), 2004, 1295-1306.

[8] H. Cao, X. Ding & C. Liu, Rectifying the Bound Document Image Captured by the Camera: A Model Based Approach. *Proc. Int'l Conf. Computer Vision*, 2003, 71-74.

[9] B. Gatos, I. Pratikakis & S.J. Perantonis, Adaptive Degraded Document Image Binarization. *Pattern Recognition*, *39*, 2006, 317-327.

[10] L. Shapiro & G. Stockman, *Computer Vision* (Prentice Hall, 2001)

[11] Z. Zhang & C.L. Tan, Recovery of Distorted Document Images from Bound Volumes. *Proc. ICDAR'01,* 2001, 429-433.

[12] C. Strouthopoulos, N.Papamarkos & C.Chamzas, Identification of Text-Only Areas in Mixed-Type Documents. *Eng. Applic. Artif Intell.*, *10*(4), 1997, 387-401.

[13] U.V. Marti & H. Bunke. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *Int. Journal of Pattern Recognition and Artifical Intelligence*, *15*(1), 2001, 65–90.

[14] POLYTIMO project, http://iit.demokritos.gr/cil/Polytimo, 2006.