Indexing of Historical Document Images: Ad Hoc Dewarping Technique for Handwritten Text

Federico Bolelli

Dipartimento di Ingegneria "Enzo Ferrari"
Università degli Studi di Modena e Reggio Emilia
Via Vivarelli 10, Modena MO 41125, Italy
federico.bolelli@unimore.it

Abstract. This work presents a research project, named XDOCS, aimed at extending to a much wider audience the possibility to access a variety of historical documents published on the web. The paper presents an overview of the indexing process that will be used to achieve the goal, focusing on the adopted dewarping technique. The proposed dewarping approach performs its task with the help of a transformation model which maps the projection of a curved surface to a 2D rectangular area. The novelty introduced with this work regards the possibility of applying dewarping to document images which contain both handwritten and typewritten text.

Keywords: document indexing, page rectification, dewarping

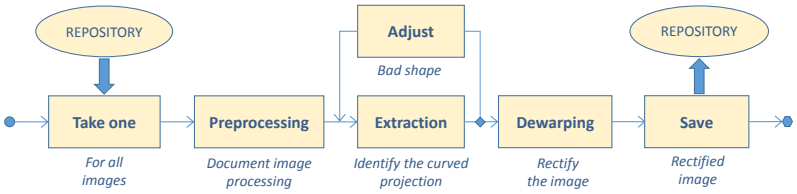
1 Introduction

XDOCS is designed with the intention of extending to a much wider audience of scholars, or even simply curious people, the possibility to access a variety of historical documents published on the web [1].

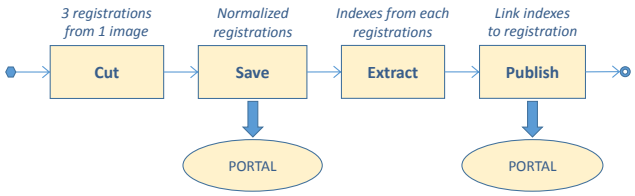
To that purpose, the project is developing an innovative data capturing technique able to extract document indexes in quasi-automatic mode from their handwritten contents. The devised solution intervenes after the dematerialisation action of scanning the historic documents and obtaining one image per couple of adjacent pages, and it is intended to be especially applied to a long series of documents such as the large number of civil registries that are available since the constitution of the Italian state.

Since warping affects, as well as documents readability, most of the high level text processing such as OCR, word spotting, and handwritten recognition, dewarping digital text is one of the fundamental requirements to perform a correct extraction of indexes. This process starts from a curled page, usually captured by a flatbed scanner or by a digital camera, and aims to obtain an output image constituted only of horizontal straight text lines, without suffering from any distortion due to perspective or page warping.

[1] http://www.antenati.san.beniculturali.it/



(a) Image rectification.



(b) Indexing & publication.

Fig. 1. Phases of the XDOCS indexing process.

Over the last two decades many methods for document dewarping have been proposed. These approaches are usually classified in two categories according to the surface model adopted: restoration approaches based on 2D document image processing [13, 8] and restoration approaches based on 3D document shape reconstruction [5, 7]. Most of the dewarping techniques proposed in the past, of both categories, are specifically designed for typewritten text. These methodologies produce bad results when applied to handwritten text or, worse, to documents containing a mix of handwritten and typewritten text. In order to improve the XDOCS indexes extraction, the coarse dewarping technique originally proposed by Stamatopoulos *et. al* [12] was adjusted to address also handwritten documents.

The remainder of the paper is organized as follows. Section 2 presents an overall description of the indexing process. In Section 3, the dewarping adopted technique is detailed. Section 4 reports some visual experimental results. Finally, in Section 5, are drawn the conclusions.

2 Indexing Process

The XDOCS indexing process is split into two main phases, namely "image rectification" and "indexing & publication". The former is depicted in Figure 1(a), showing the steps that move from a scanned image up to its final squaring. More specifically:

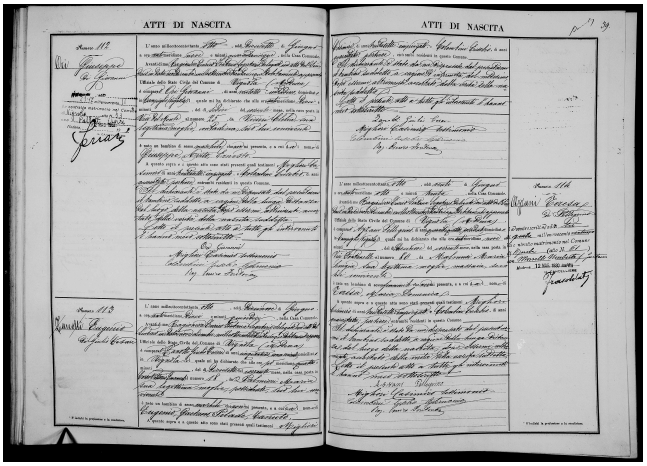
- *Repository* is the place where the original scanned images are found.
- *Preprocessing* is the document image processing step, filtering out noise due to the intrinsic features of the original image and to the digitization process.
- The *Extraction* step aims to find the projection of the curved surface represented by two almost vertical straight lines and by two third degree polynomial curves surrounding the document page (see Section 3.2 for details). This is required by the proposed dewarping method.
- *Adjust* is the manual operation required whenever the *Extraction* operation fails.
- *Dewarping* is the core step of the "image rectification" phase; its purpose is to compute the dewarping, which transforms the original image into a rectified and normalized one.
- *Save* is the final step associating the rectified image to the original one in the *Repository* for further processing.

The latter phase of the indexing process is depicted in turn in Figure 1(b), showing the steps that cut individual registrations from the rectified image and lead to indexing each of those registrations. More specifically:

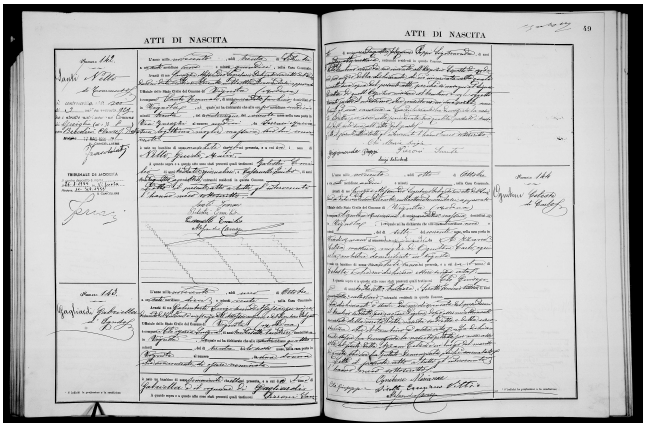
- *Portal* is the place where the registrations and their indexes are made available for consultation.
- *Cut* is the image processing step separating and normalizing the three registrations that are present in every rectified image.
- *Save* is the step storing the normalized registrations into the Portal.
- *Extract* is the image processing function which extracts and compares parts of images, corresponding to names, places and numbers (mainly dates of birth) [4].
- *Publish* is the final step associating the extracted indexes to the corresponding registration in the Portal.

Of course, the degree of completeness and confidence of the extracted indexes are strongly affected by the quality of the handwritten text and the state of preservation of the original registry. Those factors can however be increased by driving the *Extraction* action by templates representing the limited areas to be examined for the purpose of finding out each of the desired indexes. The templates are manually defined on the normalized registrations and in principle can depend on the single registry (year, municipality, handwriting style of the registrar).

Figure 2 reports two examples of birth registry, the first one dated June 1888 and the second one September 1900. Each image shows two pages containing three birth registrations: one on the left-upper side, one on the right-bottom side and one split between the two pages. The three registrations share the same structure and present the intended indexes in equivalent and well identified positions. Moreover, the most critical indexes, namely family name and given name, appear twice in each registration and this redundancy can increase the level of confidence in the indexing.



(a) Historical birth act dated June 1888.



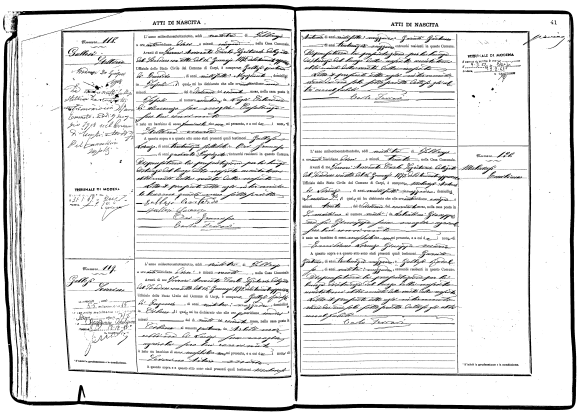
(b) Historical birth act dated September 1900.

Fig. 2. Examples of warped digital document images.

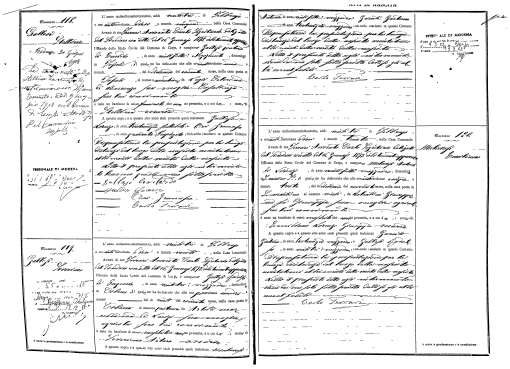
The metadata identifying the double page image are: registry type, year and volume, place of registration (typically, a municipality). The intended indexes are in turn: birth month and day, name and family name, sex, father's name, mother's name and family name, possibly grandparents names.

3 The Proposed Dewarping Approach

This section describes in detail "Image Rectification" phase focusing on *Preprocesing*, *Extraction* and *Dewarping* steps.



(a) Output binary image.



(b) Output of preprocessing step.

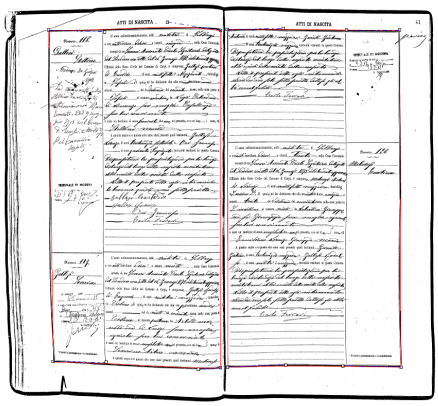
Fig. 3. Example of preprocessing results.

3.1 Preprocessing

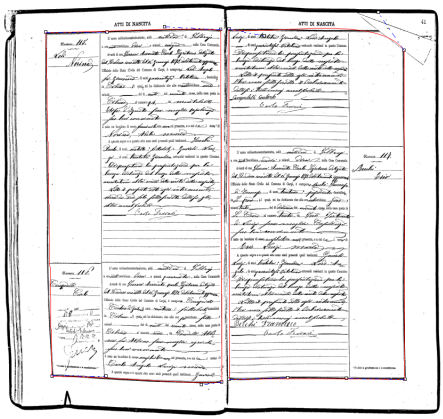
Before proceeding with the dewarping step, which is detailed in the following, the gray level images are mapped into black-white ones using the adaptive threshold described in [11] (see Figure 3(a) as example). Then, noise is filtered out principally using information related to statistics of connected components calculated using [9]. An example of preprocessing output is reported in Figure 3(b).

3.2 Extraction

The extraction step aims at identifying the 2D projection of the curved surface defined by the four polynomial curves which surround the document text on every single page (see Figure 5).



(a) 2D projection correctly extracted.



(b) 2D projection badly extracted.

Fig. 4. Example of curved 2D projection extraction during the *Adjust* phase. Best viewed in color.

According to the warping model characterizing historical documents, the right and left polynomial curves are supposed to be lines and generically defined as:

$$y = ax + b \tag{1}$$

To identify these lines the approach combines the information obtained by the Hough transform [6] and the position of A, B, C and D vertexes of the curved surface projection retrieved using the Harris algorithm and starting from a thinned image [10].

Top and bottom curves, instead, are supposed to be third degree polynomial lines and their coefficients are fitted with the Least Square Estimation algorithm

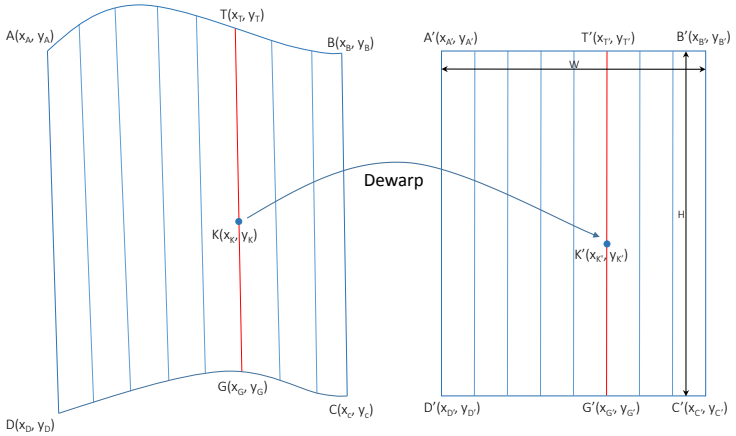


Fig. 5. Dewarping transformation model: projection of the curved surface on the left side, 2D rectangular destination area on the right side.

and have the following general expression:

$$y = ax^3 + bx^2 + cx + d \tag{2}$$

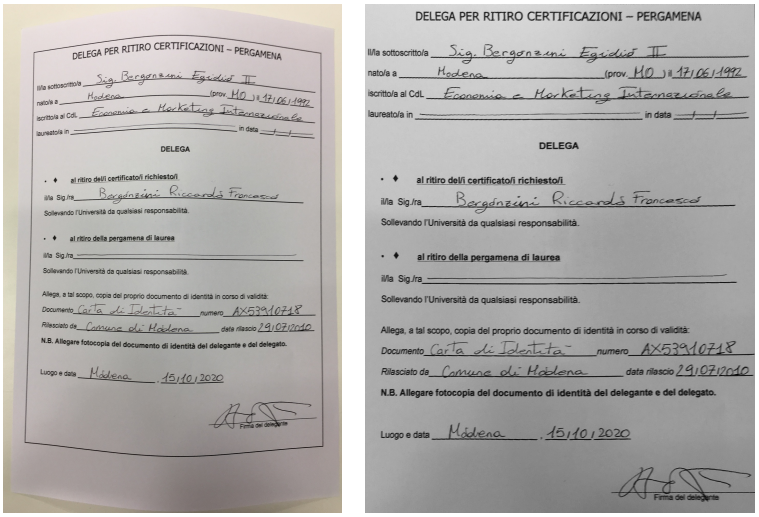
More accurate results could be achieved modeling these curves as higher polynomial functions. This change increases dewarping computation time slightly improving accuracy, so it is not recommended. Boundary extraction significantly influences the quality of the dewarping process, and then the indexes extraction: if it fails the *Adjust* step leaves the user the possibility to correct curves via a GUI: examples of the automatic extracted 2D projections are reported in Figure 4.

3.3 Dewarping

This is the core step of the dewarping approach which aim to map the projection of the curved surface to a 2D rectangular area with fixed dimensions H and W (see Figure 5 for details). Stages of the mapping process are detailed in this section using the following notation: $A(x_A, y_A)$, $B(x_B, y_B)$, $C(x_C, y_C)$, and $D(x_D, y_D)$ are the vertexes of the projection surface whereas $A'(x'_A, y'_A)$, $B'(x'_B, y'_B)$, $C'(x'_C, y'_C)$, and $D'(x'_D, y'_D)$ are the ones of the rectangular destination area. Moreover, the euclidean distances between points A and D, B and C are respectively called $|AD|$ and $|BC|$, and the lengths of the polynomial curves $|\widehat{AB}|$ and $|\widehat{CD}|$ are defined as:

$$|\widehat{AB}| = \int_{x_A}^{x_B} \sqrt{1 + [f'(x)]^2} dx \tag{3}$$

$$|\widehat{DC}| = \int_{x_D}^{x_C} \sqrt{1 + [g'(x)]^2} dx \tag{4}$$



(a) Original warped document. (b) Dewarping Result.

Fig. 6. Example of dewarping applied on generic digital document image.

where f(x) and g(x) are the functions describing the polynomial lines.

Therefore, given a generic point $K(x_K, y_K)$ on the warped image, the corresponding one on the 2D rectangular area $K'(x'_K, y'_K)$ can be found preserving proportions between dimensions of projected curves and 2D destination area. First of all it is necessary to find the two points $T(x_T, y_T) \in |\widehat{AB}|$ and $G(x_G, y_G) \in |\widehat{DC}|$ such that $K \in TG$ and $|\widehat{AT}| : |\widehat{AB}| = |\widehat{DG}| : |\widehat{DC}|$. The transformation equations are then defined as follows:

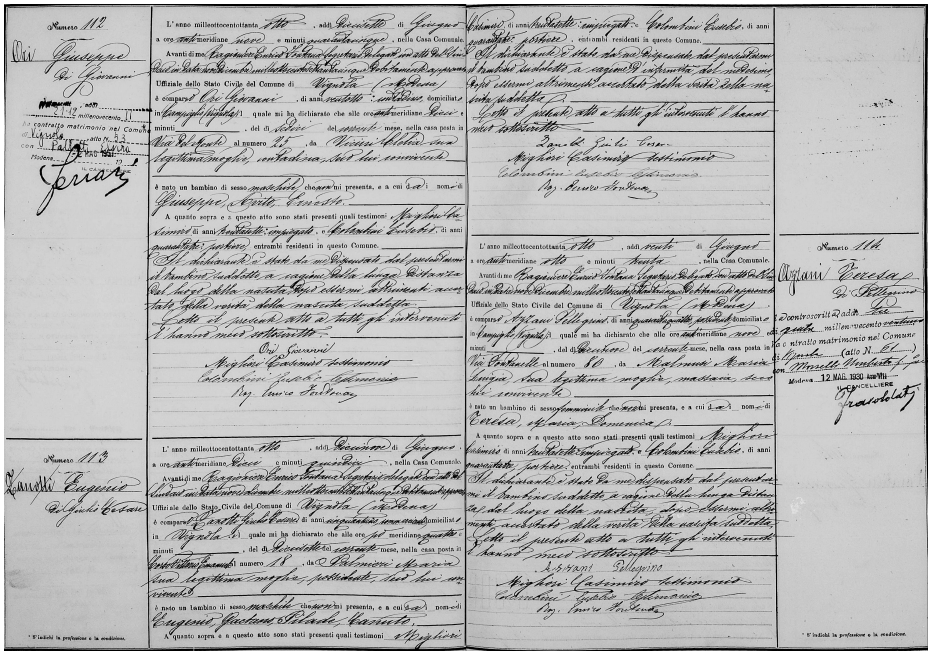
$$x'_K = x'_A + W * \frac{|\widehat{AT}|}{|\widehat{AB}|} \tag{5}$$

$$y'_K = y'_A + H * \frac{|TK|}{|TG|} \tag{6}$$

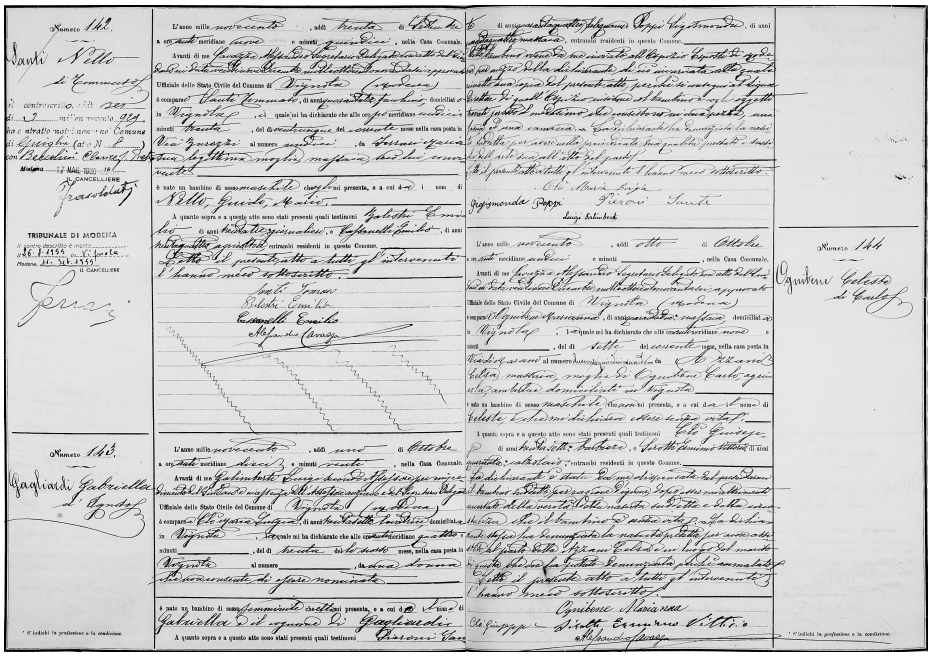
To compute the final page dewarping every pixel in the dewarped image is mapped to a floating-point coordinate in the warped image, therefore the process is concluded using a simple interpolation.

4 Experimental Results

Common practices in the evaluation of dewarping techniques consist of comparing the error rate of OCR software applied on the original and dewarped images or are simply based on visual pleasing impressions. Unfortunately, the first strategy is unfeasible for documents which contain handwritten text, so the second one is adopted in this paper.



(a) Historical birth act dated June 1888: original image is depicted in Figure 2(a).



(b) Historical birth act dated September 1900: original image is depicted in Figure 2(b).

Fig. 7. Examples of dewarping applied on historical digital document images.

Figure 6 reports an example result of the proposed dewarping technique applied to a regular modern form image, which contains a bounding box and both handwritten and typewritten text. It is possible to see that the horizontal lines in the dewarped image are perfectly aligned with the horizontal boundaries of the box.

Figure 7 instead, shows another result applied on the typical historical documents treated in this work. Also here both handwritten and printed text is present and the detection phase is complicated by the presence of many distracting elements.

The method has been tested on more than 4.000 birth acts and on almost 200 generic digital documents similar to the one reported in Figure 6. Experimental results reveal that more than 85% of curved 2D projections are correctly extracted and do not require the manual *Adjust* step before performing the dewarping procedure.

5 Conclusions

This paper describes the rationale and objectives of a research project presently underway at SATA s.r.l. in collaboration with the University of Modena and Reggio-Emilia, and co-funded by the Emilia-Romagna regional administration. In particular, a relatively novel approach for performing dewarping on digital document images containing both handwritten and typewritten text was detailed. The proposed method assumes that original text is surrounded by a bounding box from which the projection of the curved surface is extracted. This is a strong assumption, but it is not uncommon to find such documents: most of the "precompiled" modules present this kind of structure and the historical documents tested confirm the assumption. Moreover, experimental results demonstrate the quality of the proposed approach. Future work will require the exploration of Convolutional Neural Networks architectures, in order to improve the image *Extract* stage [2, 3, 1].

References

1. Balducci, F., Grana, C., Cucchiara, R.: Classification of affective data to evaluate the level design in a role-playing videogame. In: Games and Virtual Worlds for Serious Applications (VS-Games), 2015 7th International Conference on. pp. 1–8. IEEE (2015)
2. Baraldi, L., Grana, C., Cucchiara, R.: Hierarchical boundary-aware neural encoder for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
3. Baraldi, L., Grana, C., Cucchiara, R.: Recognizing and presenting the storytelling video structure with deep multimodal networks. IEEE Transactions on Multimedia 19(5), 955–968 (2017)
4. Bolelli, F., Borghi, G., Grana, C.: Historical handwritten text images word spotting through sliding window hog features. In: 19th International Conference on Image Analysis and Processing (2017)

5. Cao, H., Ding, X., Liu, C.: Rectifying the bound document image captured by the camera: A model based approach. In: Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on. pp. 71–75. IEEE (2003)
6. Duda, R.O., Hart, P.E.: Use of the hough transformation to detect lines and curves in pictures. Communications of the ACM 15(1), 11–15 (1972)
7. Fu, B., Wu, M., Li, R., Li, W., Xu, Z., Yang, C.: A model-based book dewarping method using text line detection. In: Proc. 2nd Int. Workshop on Camera Based Document Analysis and Recognition, Curitiba, Barazil. pp. 63–70 (2007)
8. Gatos, B., Pratikakis, I., Ntirogiannis, K.: Segmentation based recovery of arbitrarily warped document images. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). vol. 2, pp. 989–993. IEEE (2007)
9. Grana, C., Baraldi, L., Bolelli, F.: Optimized connected components labeling with pixel prediction. In: International Conference on Advanced Concepts for Intelligent Vision Systems. pp. 431–440. Springer (2016)
10. Grana, C., Borghesani, D., Cucchiara, R.: Decision trees for fast thinning algorithms. In: Pattern Recognition (ICPR), 2010 20th International Conference on. pp. 2836–2839. IEEE (2010)
11. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern recognition 33(2), 225–236 (2000)
12. Stamatopoulos, N., Gatos, B., Pratikakis, I., Perantonis, S.J.: A two-step dewarping of camera document images. In: Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on. pp. 209–216. IEEE (2008)
13. Ulges, A., Lampert, C.H., Breuel, T.M.: Document image dewarping using robust estimation of curled text lines. In: Eighth International Conference on Document Analysis and Recognition (ICDAR'05). pp. 1001–1005. IEEE (2005)