Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 38 (2014) 108 – 115

Procedia Computer Science

10th Italian Research Conference on Digital Libraries, IRCDL 2014

A novel model-based dewarping technique for advanced Digital Library systems

Alessandro Pugliese[a]*, Silvestro Pomes[a], Stefano Ferilli[a], Domenico Redavid[b]

[a]*Università degli Studi di Bari "Aldo Moro", Via E. Orabona 4, 70125 Bari, Italy*
[b]*Artificial Brain S.r.l., Via Piave 63, 70125 Bari, Italy*

Abstract

Digitization often introduces distortions in the form of odd perspective and curved text lines, especially toward the spine region of bound documents, that tamper both the creation of an acceptable digital reproduction of the document and the successful extraction of its textual content using OCR techniques. While already known for traditional acquisition means such as flatbed or planetary scanners, the problem gets even worse with the use of cameras, whose current widespread availability may open new opportunities for librarians and archivists. *Dewarping* is in charge of handling this kind of problems. This paper proposes a novel model-based dewarping method, aimed at solving some of the shortcomings of existing approaches through the use of a mixture of image processing and numerical analysis tools. The method is based on the construction of a curvilinear grid on each page of the document by means of piecewise polynomial fit and appropriate equipartition of a selection of its curved text lines. We show the method on sample documents and evaluate its impact on successful rate of OCR on a dataset.

Peer-review under responsibility of the Scientific Committee of IRCDL 2014

Keywords: Dewarping; Layout Analysis; Document Processing;

1. Introduction

When digitizing bound documents, often the pages cannot be completely flattened, causing an odd curvature in the spine region. While for documents made up of a few pages the distortion is limited to the blank margins, for documents having a medium or large number of pages it soon affects the page content. In particular, text lines look

* Corresponding author. Tel.: +39 080 544 2689 ; fax: +39 080 544 3610 .
E-mail address: alessandro.pugliese@uniba.it

Peer-review under responsibility of the Scientific Committee of IRCDL 2014
doi:10.1016/j.procs.2014.10.018

curved rather than straight, which might be a problem both for the final user (the digitized document is not aesthetically pleasing) and for automatic document processing (e.g., segmentation techniques based on rectangular white background pieces or OCR fail). The problem is worse if documents are not acquired using flatbed scanners. For instance, when digitizing ancient documents these devices would stress the pages and might damage unique and precious copies. In these cases, the use of (very costly) *planetary* (or "orbital") scanners is advised, where the document is open with its face up and is acquired by a camera placed above it. Still worse is the case in which the documents are photographed, because additional perspective distortions are introduced. In fact, this latter option is being given more and more attention, due to both the availability of cheap and compact high-resolution cameras, and it causing a reduced stress on the document under digitization (an extremely important requirement in the case of historical and fragile items). Fig. 1 shows a sample picture of an open book, from which one would like to restore the original page image and automatically recognize the corresponding content. In fact, the current availability of handy and low-cost but high-performance devices for image acquisition opens new landscapes for librarians and archivists, that may easily and quickly acquire documents, or parts thereof, to be included in their repositories or to be used for the recording activity. The LIBR@RIAN application, by Artificial Brain S.r.l., provides these professionals with a feature that allows them to take pictures of document pages using their smartphone, and directly send them to the Digital Library system for processing. For instance, pictures of document pages may be used by the system to rebuild the entire document and store it in the repository; or, photos of a book's front matter pages may be used to enrich the book's record with a picture of its title page and/or to automatically extract metadata and other kinds of information useful to help the librarian in his cataloging activity. In the latter case, OCR techniques must be applied to obtain machine-readable text from the picture. *Dewarping* is the term usually used to denote the process of compensating for the distortion introduced by the flat (2D) representation of spatially spread (3D) documents[2]. In this paper, we propose a model-based dewarping technique, that is integrated in LIBR@RIAN, aimed at overcoming some of the limitations of the existing literature through the use of a mixture of image processing and standard numerical analysis tools. Essential step of the method, and also its main point of novelty, is the construction of a curvilinear grid on each page of the document based on piecewise polynomial fitting and appropriate equipartition of a selection of its curved text lines. After recalling background notions and related work in next section, we describe and evaluate the proposed technique in Sections 3 and 4, respectively. Lastly, Section 5 concludes the paper and outlines anticipated directions for future work.

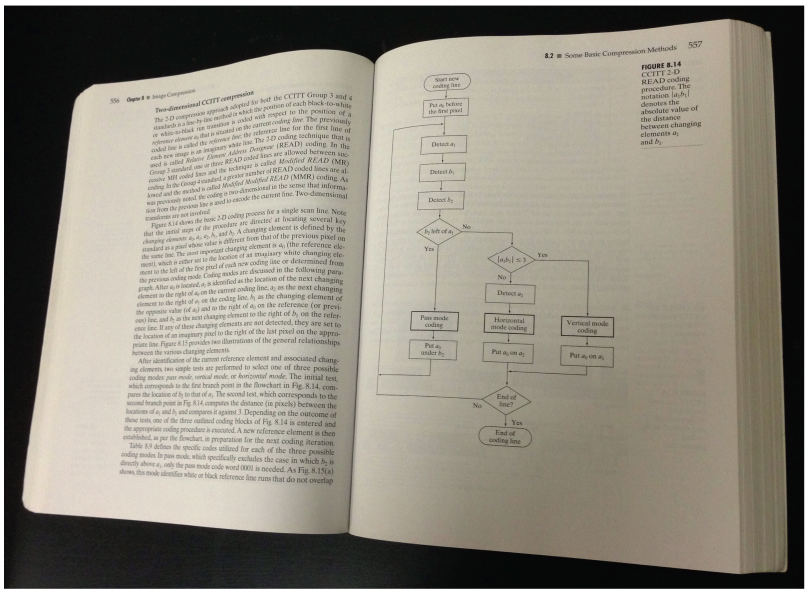

Fig. 1. Image of a book showing deformation by perspective and binding

2. Background and Related Work

The approaches proposed so far as potential solutions to restore the original flat version of a warped document have focused either on the hardware side (by exploiting dedicated acquisition devices that can better capture the 3D nature of the digitized object) or on the software one (by working on a standard picture of the item obtained using conventional means). The latter option is cheaper and ensures wider applicability, for which reason it gradually replaced the former. As already mentioned, one of the typical aims of dewarping is improving the outcome of OCR systems on distorted document pages. While OCR systems might be developed that can directly deal with the problem and fix it, a more straightforward and general approach consists in embedding a dewarping step in the normal pre-processing of the page to remove this kind of distortion before applying standard OCR.

Two kinds of approaches can be distinguished for identifying the deformations to be corrected, one based on the geometrical features of the overall document image, and one based on the distortions detected on specific elements of the document, such as text lines (a quite significant and straightforward indicator of the problem). Techniques and tools that attempted to carry out the image dewarping task include polynomial fit[12], splines[14], snakes[1], regression[15], grid modeling[11] and even fuzzy sets[10]. The technique proposed by Gatos et al.[6] exploits segmentation of the document image and linear regression, leveraging the slope of single words. Liang et al.[9] attempt to recover 3D information from 2D images in order to rectify the document images. Yang et al.[13] propose a grid-based technique which is not based on curve fitting.

3. Dewarping Technique

Before describing in details the approach we propose, we briefly elucidate, and motivate, its main points of novelty. As already mentioned, we propose a model-based approach. The decision of using a model to characterize the shape of curved text lines is driven by specific aims:

- to be able to construct a curvilinear grid with meshes of *arbitrary* size via interpolation of the model's parameters (which will be coefficients of polynomials);
- to extend the grid to zones of the image that did not give any contribution to the determination of the shape of the text lines (due to the absence of text for reasons such as presence of figures or formulas, short text lines or indentation);
- to be able to "smooth" the parameters characterizing the model to cope with local errors.

The model we choose is *piecewise* polynomial, and based on least squares spline fit. Indeed, using one single polynomial (e.g. a cubic one[12]) throughout an entire text line did not seem to be an adequate choice for the specific case of interest. When dealing with bound documents, text lines seem to follow different patterns depending on the distance from the binding, and the use of piecewise polynomials allows to correctly follow each pattern. In order to resolve the document distortion along the direction perpendicular to the text lines, we perform an equipartition of the piecewise polynomial curves by arc-length, which proves to be quite effective. Methods based on the detection of the vertical texture flow field[9] do not always produce acceptable results, and relying on detection of the character's orientation can be misleading (we implemented it but observed erroneous behavior in presence of italic characters). The dewarping technique we propose only relies on 2D information of the document. It works in several steps on the input image. A graphical summarization of the two main steps applied to a sample document (left page in Figure 1) is shown in Figure 5. For images, we adopt the standard coordinate system (x, y), with the origin being on the upper-left point, x-coordinate pointing to the right and y-coordinate pointing down. The coordinates will be normalized, so that $x, y \in [0,1]$. First, the portion that contains text is extracted from the image using the RLSO automatic segmentation technique[4,3] provided by the DOMINUS^{plus} platform[5] underlying LIBR@RIAN. Throughout the rest of the paper, we assume that most of the image contains text, and that the text lines have horizontal orientation. We split the description of the workflow into two stages, separating pre-processing steps from the actual dewarping.

3.1. Pre-processing

The pre-processing work consists of the following steps:

1. Binarization;
2. Connected Components Labeling (CCL);
3. Tracing of the text lines.

Binarization. For our purpose, a successful binarization should retain all the characters, not introduce noise, and remove the shade typically present along the spine of thick bound documents. We have used a modified version of Niblack's binarization method presented by Graham Leedham et al.[7], specifically designed for bound documents.

CCL. All the connected components are identified through the CCL algorithm outlined by Haralick and Shapiro[8]. Most of them will be characters (or part thereof), but some will be apostrophes, accents, punctuation marks, or even "noise pixels" introduced by the binarization algorithm. It is important to filter those last components out, as they would compromise the success of the *tracing* process (see below). The filtering technique we have adopted is based on the idea that the connected components that need to be filtered out are generally convex. In practice, we excluded from the tracing all components whose convex hull exceeds the number of pixels by more than 10%, and it proved to be effective (see Fig. 2).

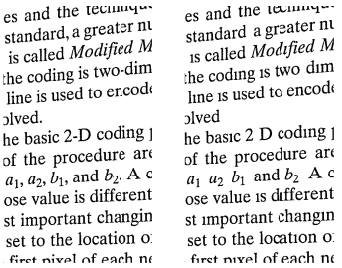

Fig. 2. A zoom-in on the binarized sample page showing the effect of filtering out the connected components nearly unchanged under "convexification" (right).

Tracing of the text lines. After all connected components are labeled and examined, we extract from each of them (which, at this point, is presumed to be a character) the upper pixels (where the value of y is minimum) and lower ones. We collect those pixels in two sets B and X, that contain, respectively, the coordinates of all points that belong to the top and bottom portion of characters. Now the curved text lines can be traced: for each text line, the goal is to identify points in B (respectively, X) that belong to the same *base line* (respectively, *x line*), see Fig. 3.a.

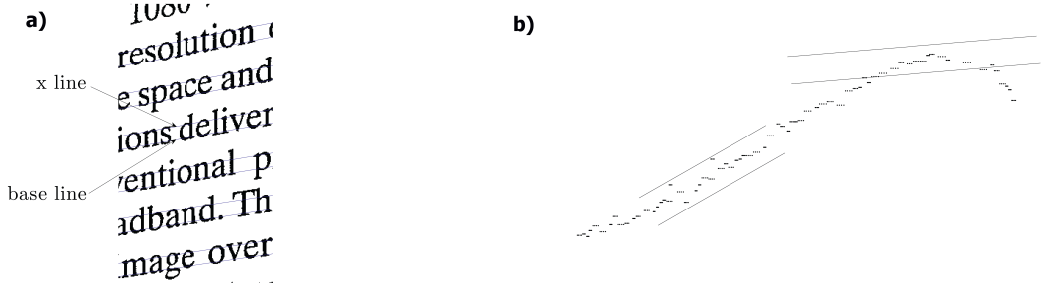

Fig. 3. (a) Approximate x line and baseline. (b) Tracing of the text lines.

This is done through an adaptation of the technique by Lu et al.[10] to our context. Let us briefly sketch this process for the case of base lines. Given $k+1$ points $p_{0,l}, \dots, p_{k,l}$ on the base line l, the next point $p_{k+1,l}$ is added to the same base line according to the following criterion: $p_{k+1,l}$ is chosen as the closest point to $p_{k,l}$ among those in B that lie inside a strip that is adaptively constructed to (locally) follow the direction of the curved text line l (see Fig. 3.b) while excluding points that belong to other base lines. For more details, we refer the reader to Lu et al.[10], where the case of straight text lines is discussed.

3.2. Dewarping

The actual dewarping consists of the following steps:

1. x lines and base lines fit;
2. construction of the curvilinear grid;
3. homography.

Fitting x lines and base lines. As already mentioned, to fit the text lines we have chosen a least squares cubic splines model. To exemplify, consider a base line l, and let $\{p_{0,l}, \dots, p_{n,l}\}$ be the points that belong to l, with $p_{i,l}$ having coordinates $(x_{i,l}, y_{i,l})$, for $i = 0, \dots, n$. Then, we seek a twice continuously differentiable function $f : [0,1] \to \mathbb{R}$ that minimizes

$$\sum_{i=0}^{n} \left(f\left(x_{i,l}\right) - y_{i,l} \right)^2,$$

and such that

$$\begin{cases} a_{1,l}x^3 + b_{1,l}x^2 + c_{1,l}x + d_{1,l} & \text{if} \quad x \in [0, e_1] \\ a_{2,l}x^3 + b_{2,l}x^2 + c_{2,l}x + d_{2,l} & \text{if} \quad x \in [e_1, e_2] \\ a_{3,l}x^3 + b_{3,l}x^2 + c_{3,l}x + d_{3,l} & \text{if} \quad x \in [e_2, 1] \end{cases}$$

Note that each function f is characterized by 12 coefficients, but only has 6 degrees of freedom due to the conditions of being C^2 at the breakpoints e_1, e_2. The line fit is performed only for text lines that are sufficiently long. In our experience, fitting a short text line would most likely result in a polynomial curve that soon departs from the actual text line we seek to approximate. Once all the relevant base lines and x lines have been fitted (see left picture in Fig. 4.a), we have 24 vectors made of polynomial coefficients (12 for the base lines and 12 for the x lines). Each vector is further "smoothed" via a simple centered moving average, in order to smooth out the influence of erroneous fits. For instance, fitting a text line that exhibits indentation often results in aberrations towards the left margin of the page. The smoothing has to be performed with care in order not to introduce aberrations in case of uneven spacing of text lines. Finally, for every base line and x line we have traced, we detect the x coordinates where the text begins and ends. Also these data are "smoothed" to void the influence of indentation.

Curvilinear grid. The coordinates of the leftmost and rightmost point on the upper x line and lower base line are detected. Now, given an *arbitrary* user supplied integer M, we create $M-1$ "equidistant" curves between the topmost x line and the bottommost base line. We do this by appropriately interpolating the coefficients of all the x lines and base lines detected so far. Now the x lines and base lines are dropped, and simply $M+1$ piecewise polynomial curves are retained. For each curve, the points where the text actually begins and ends are approximated by interpolation of the already possessed data. Then the relevant portion of each curve is partitioned into an

arbitrary user supplied number N of arcs, each having approximately the same arc-lenght. This gives us $(M+1)(N+1)$ points. Appropriately joining those points, we obtain a curvilinear grid (see right picture in Fig. 4.a).

Homography. Finally, we apply a standard homography to transform each quadrangle of the grid into a horizontal rectangle. The size of the output rectangles is carefully chosen so to approximately preserve proportions and density of pixels as of the original image. Fig. 4.b shows the outcome of the whole process on the sample document.

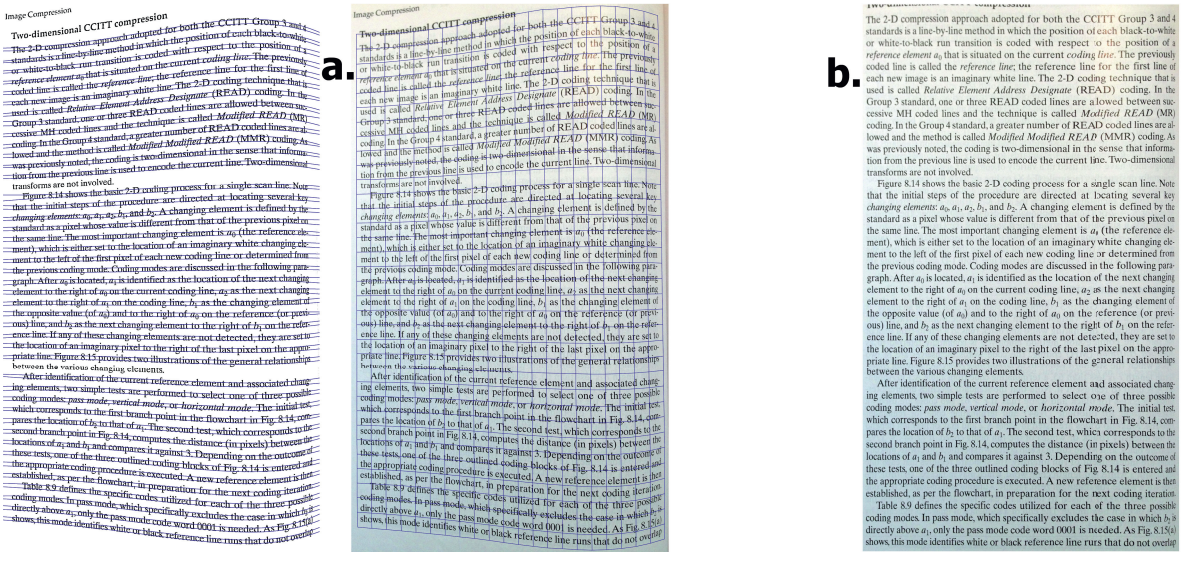

Fig. 4. (a) Main processing steps on a sample image: fit of x lines and base lines (left), construction of a curvilinear grid (right). (b) Sample page dewarped.

4. Evaluation

Our technique was implemented in Matlab. Typically, and in particular for the experiments shown on this paper, we chose $M \approx 35$, $N \approx 20$, $e_1 = 0.35$, $e_2 = 0.65$. A first qualitative evaluation was carried out on several documents, each having approximately a resolution of 5Mpx. Fig. 5 shows the outcome of the dewarping process on two rather difficult sample images. Even though significant portions of the documents do not contain text lines, the method is extremely successful at dewarping the whole the pages. Given the promising results, we run our technique on the 11 documents in the dewarping benchmark dataset used at the ICDAR 2011 Page Dewarping Contest (`http://didcontest2011.blogspot.com/`. The contest was not held, and it was impossible to find any result obtained by other approaches on it. So, to the best of our knowledge our results are the only available outcome on this dataset). Specifically, our objective consisted in analyzing the improvement in performance that could be reached by applying a standard OCR system (Tesseract by Google) after dewarping the pages, compared to the baseline represented by the performance of the same OCR without dewarping.

The results are reported in Table 1, where some documents are associated to two columns due to the segmentation step having split the page content into different portions. Looking at the table, we see that in one case (043/1) neither the original nor the dewarped version of the page could be read satisfactorily. This was due to the spine shadow in the picture being too close to the text, so that the segmentation technique was not able to strip it off. Clearly, proper pre-processing might solve the problem. In 8 out of the remaining 14 cases the dewarping outcome allowed to read documents that were otherwise unreadable, with error rates always less than 10%, and in one case (070/1) reaching even 0%. Only in one case (044/1) the dewarped version was unusable, whereas the original one gave very good results; this was due to a strange dewarping behavior that dramatically distorted the document instead of straightening it up. In case 009/1 the dewarped version allowed to read significantly more text than the

original one. In the remaining 4 cases, the dewarped version provided slightly worse results (especially on character performance) than the original one. This can be explained by looking at the last 4 rows in Table 1, that reveal that most errors are squeezed in the recognition of the first and/or last line of the document. Considering only the rest of the document, performance is much better, both compared to the non-dewarped version and as absolute error rate. This behavior can be easily associated to the specificity of the technique, that when building and processing the grid sometimes cuts away part of the first and/or last line. Future work will take care of fixing this problem.Overall, the proposed dewarping technique was effective in significantly improving the performance of OCR reading on warped documents, showing marginal problems that are well-defined and can be likely tackled in next versions.

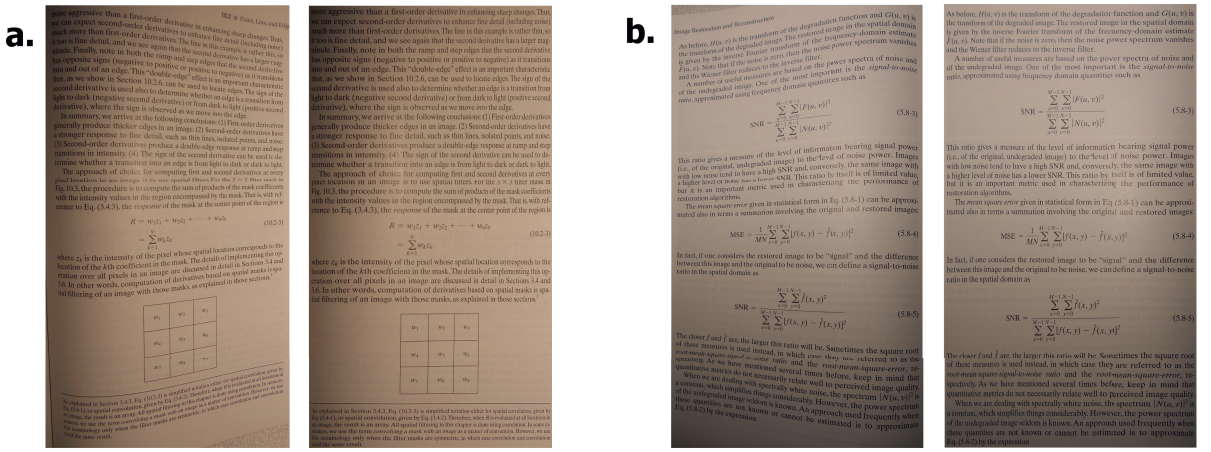

Fig. 5. Two sample documents, original (left) and dewarped (right).

Table 1. Results of quantitative evaluation of the proposed dewarping technique: doc (document id in the dataset); gt(w/c) = ground truth (words/characters); e(w/c)b = errors (word/character) baseline with no dewarping; e(w/c)d = errors (word/character) after dewarping; (w/c)m(f/l) = (words/characters) missed (first/last) row. '?' denotes cases with more than 50% error.

Document	**001**	**006**	**009**	**009**	**011**	**043**	**043**	**044**	**044**	**067**	**070**	**070**	**088**	**096**
part			**1**	**2**		**1**	**2**	**1**	**2**		**1**	**2**		
gtw	132	104	293	174*	121	160	177	398	334	416	156	46	271	293
gtc	748	427	1379	902*	513	866	885	2189	1846	2147	882	208	1439	1526
ewb	?	?	18	1	15	?	1	10	?	?	?	4	?	?
ecb	?	?	34	1	49	?	2	28	?	?	?	5	?	?
ewd	11	10	3	1	17	?	3	?	7	15	0	4	8	13
ecd	59	40	4	5	71	?	6	?	11	44	0	14	66	44
wmfr	9	0	0	8	5	?	10	?	5	12	0	2	1	7
wmlr	0	9	0	0	12	?	0	?	10	5	0	1	12	3
cmfr	56	0	0	46	30	?	37	?	33	58	0	7	1	37
cmlr	0	39	0	0	41	?	0	?	74	22	0	6	46	4

*: the pre-processing steps stripped away some matter (32 words, 154 characters) at the bottom of the page.

5. Conclusion

We have proposed an original technique for dewarping digital documents, and showed its performance through several examples and a benchmark dataset. Key idea is a simple recipe for the construction of a curvilinear grid on

the document, that captures both its vertical and horizontal geometrical distortions, especially those caused by the presence of thick bindings.

Throughout the experiments we performed, the technique proved to be very effective. Yet, we have identified some points that need improvement. For instance, the pre-processing stage is time-consuming (in the present implementation, it accounts for nearly 2/3 of the total computation time), and fairly sensitive to the choice of several parameters and thresholds. For this reason, we anticipate working on a new, greatly simplified, line tracing strategy that is expected to be more robust and not require most of the present pre-processing steps.

Acknowledgements

This work was partially funded by the Italian PON 2007-2013 project PON02 00563 3489339 “Puglia@Service”.

References

1. Bukhari SS, Shafait F, Breuel T. Dewarping of document images using coupled-snakes. In Proceedings of the Third International Workshop on Camera-Based Document Analysis and Recognition, 2009.
2. Ferilli S. Automatic Digital Document Processing and Management – Problems, Algorithms and Techniques. Advances in Pattern Recognition. Springer, 2011.
3. Ferilli S, Basile TMA, Esposito F. A histogram-based technique for auto matic threshold assessment in a run length smoothing-based algorithm. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, DAS ’10, pages 349–356, New York, NY, USA, 2010. ACM.
4. Ferilli S, Biba M, Esposito F, Basile TMA. A distance-based technique for non-manhattan layout analysis. In Document Analysis and Recognition, 2009. ICDAR ’09. 10th International Conference on, pages 231–235, 2009.
5. Ferilli S, Esposito F, Basile TMA, Redavid D, Villani I. Dominus plus - document management intelligent universal system (plus). In Maristella Agosti, Floriana Esposito, Carlo Meghini, and Nicola Orio, editors, IRCDL, volume 249 of Communications in Computer and Information Science, pages 123–126. Springer, 2011.
6. Gatos B, Pratikakis I, Ntirogiannis K. Segmentation based recovery of arbitrarily warped document images. In Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR), pages 989–993, 2007.
7. Graham Leedham CY, Takru K, Tan JHN, Mian L. Comparison of some thresholding algorithms for text/background segmentation in difficult document images. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, volume 2, pages 859–864. Citeseer, 2003.
8. Haralick RM, Shapiro LG. Computer and robot visi´on. Number v. 1 in Computer and Robot Vision. Addison-Wesley Pub. Co., 1992.
9. Liang J, DeMenthon D, Doermann D. Geometric rectification of camera-captured document images. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 30(4):591–605, 2008.
10. Lu S, Chen BM, Ko CC. Perspective rectification of document images using fuzzy set and morphological operations. Image Vision Comput., 23(5):541– 553, 2005.
11. Luand S, Tan CL. Document flattening through grid modeling and regularization. In 18th International Conference on Pattern Recognition, ICPR 2006, volume 1, pages 971–974, 2006.
12. Stamatopoulos N, Gatos B, Pratikakis I. A methodology for document image dewarping techniques performance evaluation. In Document Analysis and Recognition, 2009. ICDAR’09. 10th International Conference on, pages 956–960. IEEE, 2009.
13. Yang P, Antonacopoulos A, Clausner C, Pletschacher S. Grid-based modelling and correction of arbitrarily warped historical document images for large-scale digitisation. In Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, pages 106–111. ACM, 2011.
14. Zhang Y, Liu C, Ding X, Zou Y. Arbitrary warped document image restoration based on segmentation and thin-plate splines. In 19th International Conference on Pattern Recognition, ICPR 2008, pages 1–4, 2008.
15. Zhang Z, Tan CL. Correcting document image warping based on regression of curved text lines. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, volume 1, pages 589–593, 2003.