Optical Process and Analysis of Historical Documents

Nikolaos Stamatopoulos*

[1] Department of Informatics and Telecommunications
National and Kapodistrian University of Athens

[2] Computational Intelligence Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research "Demokritos"
nstam@iit.demokritos.gr

Abstract. The collections of historical books are an important source of information, both for the history of previous periods and for the development of the cultural documentation itself. Although to date, there have been made several attempts of digitalization and electronic navigation, there is not an appropriate frame of optical process and analysis of the content of these collections, consequently a large number of historical books have not been studied yet and remain unexploited. In this thesis, we studied the preprocessing stages which are performed before the recognition process and we focused on the enhancement and segmentation of historical documents. Preprocessing stages play an important role in document image processing since they affect the performance of subsequent processing, such as optical character recognition. At the enhancement stage, we focused on the border removal as well as on the dewarping of document images, which are common problems associated with historical documents. Two methodologies that detect and remove black borders as well as noisy text regions are proposed. Furthermore, optimal page frames of double page document images are detected. The experimental results on several historical documents demonstrate the effectiveness of the proposed techniques. Concerning the warping problem, a coarse-to-fine rectification methodology to compensate for undesirable document image distortions is proposed. To verify the validity of the proposed methodology, experiments have been carried out using indirect evaluation techniques as well as a novel semi-automatic evaluation methodology. At the document image segmentation stage we proposed a novel combination method of complementary text line segmentation techniques. Furthermore, a methodology for character segmentation in historical documents is suggested. Comparative experiments using several historical documents from different languages and time periods prove the efficiency of the proposed technique. Finally, in order to ease the construction of document image segmentation ground-truth that includes text-image alignment we presented an efficient technique.

Keywords: document image enhancement, border removal, document image dewarping, document image segmentation, combined segmentation techniques

* Dissertation Advisors: [1] Sergios Theodoridis, Professor – [2] Basilis Gatos, Researcher

1 Introduction

Recognition of historical documents is essential for quick and efficient content exploitation of the valuable historical collections that are part of our culture heritage. Several factors such as low paper quality, dense and arbitrary layout, low print contrast, typesetting imperfections, lack of standard alphabets and fonts do not permit the application of conversional recognition techniques to historical documents. Due to these reasons, recognition of historical documents is one of the most challenging tasks in document image processing.

In this thesis, we studied the preprocessing stages which are performed before the recognition process and we focused on the enhancement and segmentation of historical documents. Preprocessing stages play an important role in document image processing since they affect the performance of subsequent processing, such as optical character recognition. At the enhancement stage, we focused on the border removal as well as on the dewarping of document images, which are common problems associated with historical documents. Moreover, at the document image segmentation stage we proposed a novel combination method of complementary text line segmentation technique as well as a methodology for character segmentation in historical documents. Finally, in order to ease the construction of document image segmentation ground-truth that includes text-image alignment we presented an efficient technique.

2 Document Image Enhancement

2.1 Border Removal

Document images are often framed by a noisy black border or include noisy text regions from neighbouring pages when captured by a digital camera. Approaches proposed for document segmentation and character recognition usually consider ideal images without noise. However, there are many factors that may generate imperfect document images. When a page of a book is captured by a camera, text from an adjacent page may also be captured into the current page image. These unwanted regions are called “noisy text regions”. Additionally, there will usually be black borders in the image. These unwanted regions are called “noisy black borders”. All these problems influence the performance of segmentation and recognition processes. There are only few techniques in the literature for page borders detection [1-5]. Most of them detect only noisy black borders and not noisy text regions.

We propose a new and efficient algorithm for detecting and removing noisy black borders as well as noisy text regions [6]. This algorithm uses projection profiles and a connected component labelling process to detect page borders. Additionally, signal cross-correlation is used in order to verify the detected noisy text areas. The experimental results on several historical document images indicate the effectiveness of the proposed technique.

Moreover, document images are usually produced by scanning books or periodicals. Scanning two pages at the same time is a very common practice as it helps to accelerate the scanning process. However, it may affect the performance of subsequent processing such as document analysis and optical character recognition (OCR) since the majority of approaches are able to process only single page images. Furthermore, another drawback of scanning two pages at the same time is the appearance of noisy black borders around text areas as well as of noisy black stripes between the two pages. For these reason, we propose a novel methodology that detects the optimal page frames of double page document images that is based on the vertical and horizontal white run projections [7]. Our aim is to split the image into the two pages as well as to remove noisy borders. At a first step, a pre-processing which includes binarization, noise removal and image smoothing is applied. At a next step, the vertical zones of the two pages are detected. Finally, the frame of both pages is detected after calculating the horizontal zones for each page.

2.2 Dewarping

Document image acquisition by a flatbed scanner or a digital camera often results in several unavoidable image distortions due to the form of printed material (e.g. bounded volumes), the camera setup or environmental conditions (e.g. humidity that causes page shrinking). Text distortions not only reduce document readability but also affect the performance of subsequent processing such as document layout analysis and optical character recognition (OCR).

Over the last decade, many different techniques have been proposed for document image rectification and they can be classified into two main categories based on (i) 3D document shape reconstruction [8-9] and (ii) 2D document image processing [10-15]. Techniques of the former category obtain the 3D information of the document image using special setup or reconstruct the 3D model from information existing in document images. On the other hand, techniques in the latter category do not depend on auxiliary hardware or prior information but they only rely on 2D

In this thesis, we propose a goal-oriented rectification methodology to compensate for undesirable distortions of document images captured by flatbed scanners or hand-held digital cameras (TSD.ver2) [16]. The proposed technique is directly applied to the 2D image space without any dependence to auxiliary hardware or prior information. It first detects words and text lines to rectify the document image in a coarse scale and then further normalize individual words in finer detail using baseline correction. Although the coarse rectification stage applies word and text line detection at the original distorted document image, which is a well-known hard task, potential erroneous detection results do not seriously affect it as it requires only some specific points. Experimental results on several document images with a variety of distortions show that the proposed method produces rectified images that give a significant boost in OCR performance. This work is an extension of our previous work (TSD.ver1) [17] which incorporates a new method for the curved surface projection, the word baseline fitting as well as the restoration of horizontal alignment. We also propose to rectify the distortion of individual words using baseline estimation. Finally, we propose a new semi-automatic evaluation method [18] based on matching manually marked points of the original image and corresponding points of the rectified image. A quantitative measure is calculated to evaluate the performance of our method.

3 Document Image Segmentation

3.1 Text Line Segmentation

In document analysis and recognition, several approaches have been proposed for improving OCR accuracy through combination [19]. These approaches can be categorized in two categories: (i) techniques in classifier combinations and (ii) string alignment combination methods [20]. Approaches of the second category combine several OCR outputs to produce a more accurate string estimate of the original text, but this cannot be done on character-by-character basis because of segmentations errors. Outputs strings must be aligned to extract an estimate and also errors must be uncorrelated.

Based on a similar way of thought we could combine the results of different segmentation techniques in order to achieve better segmentation results. Document segmentation into text lines is a major task in a document image analysis system. A wide variety of methods have been proposed in the literature for document segmentation which can be categorized in five major categories: (1) projection profiles methods; (2) smearing methods; (3) methods based on the Hough transform; (4) grouping methods and (5) stochastic methods. Techniques from each category can confront some specific problems such as overlapping, touching components, image degradations, variability in skew angles and directions, disturbing elements, variability in inter-word and inter-character distances and others. So, we propose a combination method of complementary segmentation techniques where each technique can solve some different difficult problems [21]. Our goal is to increase the efficiency and the accuracy of the segmentation result using (i) the results of segmentation techniques which belong to different categories and (ii) specific features of the initial document.

3.2 Character Segmentation

The most recognition errors are due to character segmentation errors. Very often, even in printed text, adjacent characters are touching, and may exist in an overlapped field. Therefore, it is essential to segment a given word correctly into its character components. Any failure or error in this segmentation step can lead to a critical loss of information from the document. Character segmentation previous work concerns mostly handwritten text but methods for machine-printed text have also been proposed [22-23].

The proposed character segmentation algorithm [24] is based on skeleton segmentation paths which are used to isolate possible connected characters. The basic idea is that we can find possible segmentation paths linking the feature points on the skeleton of the word and its background

3.3 Creation of Document Image Segmentation Ground Truth

Efficient ground truth creation is essential for training and evaluation purposes in the document image analysis and recognition pipeline. Since a large number of tools have to be trained and evaluated in realistic circumstances we need to have a quick and low cost way to create the corresponding ground truth. Moreover, the specific need for having the correct text correlated with the corresponding image area in text line and word level makes the process of ground truth creation a difficult, tedious and costly task. Transcript mapping (or text alignment) techniques are used in order to map the correct text information to a segmentation result produced automatically. Usually, these techniques are very useful in order to automatically create benchmarking data sets. They are mainly based on hidden Markov models (HMMs) [25] and dynamic time warping (DTW) [26] and mainly focus on the alignment of handwritten document images with the corresponding transcription on word level.

We introduce an efficient transcript mapping technique to ease the construction of document image segmentation ground truth that includes text-image alignment in text line, word and character level [27]. We facilitate the annotation of text line, word and character segmentation ground truth regions as well as the correlation with corresponding text making use of the correct document transcription. In the proposed framework, we assume that the transcription includes the correct text line break information. This information is used in a novel transcript mapping module in order to efficiently create the text line, word and word segmentation ground truth. The proposed text line transcript mapping technique is based on Hough transform that is guided by the number of the text lines in order to efficiently create the text line segmentation result. Concerning the word and character segmentation ground truth, a gap classification technique constrained by the number of the words and character is used. We recorded that using the proposed technique for handwritten documents, the percentage of time saved for ground truth creation and text-image alignment is more than 90%.

4 Experimental Results

4.1 Border Removal

The performance evaluation method used is based on a pixel based approach and counts the pixels at the correct page frames and the detected page frames. For this purpose, we manually mark the correct page frames in the original document image in order to create the ground truth set. Let G be the set of all pixels inside the correct page frame in ground truth, R the set of all pixels inside the result page frame and T(s) a function that counts the elements of set s. We calculate the Precision and Recall as follows:

$$Precision = \frac{T(G \cap R)}{T(R)} \ \& \ Recall = \frac{T(G \cap R)}{T(G)} \tag{1}$$

A performance metric *FM* can be extracted if we combine the values of precision and recall:

$$FM = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2}$$

To verify the validity of the proposed method [6] we used two different datasets. The first (“POLYTIMO”) was a set of Greek historical documents [28] consisted of 370 document images. The second set (“IMPACT”) [29] consisted of 22383 historical documents including newspapers, periodical etc. For comparison purposes, we applied at the same dataset the state-of-the-art method [1] as well as the commercial packages BookRestorer [30], WiseBook [31] and ScanFix [32]. Tables 1 and 2 illustrate the overall evaluation results.

Table 1. Border Removal - Evaluation Results using “POLYTIMO” dataset.

Method	Precision (%)	Recall (%)	F-Measure (%)
Proposed Method [6]	91.11	96.95	93.94
Le et al. [1]	70.90	99.33	82.74
BookRestorer [30]	74.33	95.47	83.58
WiseBook [31]	53.00	99.02	69.05
ScanFix [32]	51.49	87.96	64.95

Table 2. Border Removal - Evaluation Results using “IMAPCT” dataset.

Method	Precision (%)	Recall (%)	F-Measure (%)
Proposed Method [6]	98.62	98.46	98.54
Le et al. [1]	97.28	94.01	95.62
BookRestorer [30]	94.11	96.92	95.50
WiseBook [31]	83.32	98.94	90.46
ScanFix [32]	85.00	98.04	91.05

To verify the validity of the proposed method [7] we used 3467 double page document images from 50 different historical books. For comparison purposes, we applied at the same dataset the commercial package ABBYY FineReader Engine 10 [33]. Table 3 illustrates the overall evaluation results.

Table 3. Border Removal & Page Split - Evaluation Results

Method	Precision (%)	Recall (%)	F-Measure (%)
Proposed Method [7]	92.04	98.35	95.09
ABBYY FineReader Engine 10 [33]	62.66	91.53	74.39

4.2 Dewarping

To verify the validity of the proposed methodology we use as a performance measure the character and word accuracy metrics by carrying out OCR on original and rectified document images. Furthermore, experiments have been carried out using the proposed semi-automatic evaluation methodology [18]. The experimental results from both procedures are presented in the sequel.

OCR Evaluation: The use of OCR as a means for indirect evaluation is widely used in the evaluation of rectification techniques. Character accuracy metric is defined as the ratio of the number of correct characters (number of characters in the correct document transcription minus the number of errors) over the total number of characters in the correct document transcription:

$$Character\ Accuracy = (\#chars - \#errors) / \#chars \tag{3}$$

In order to define the errors we count the minimum number of edit operations (insertion, deletion or substitution) that are required to correct the text generated by the OCR system (string edit distance). Moreover, we carried out OCR testing on original and rectified document images using also the word accuracy metric. Word accuracy is defined as the ratio of the number of correct words (number of words in the correct document transcription minus number of misrecognized words) to the total number of word in the correct document transcription:

$$Word\ Accuracy = (\#words - \#misrecognized\_words) / \#words \tag{4}$$

We used a dataset of 100 distorted document images at 200 dpi. The document images contain different font sizes and suffer from several distortions. For comparison purposes, we applied at the same dataset the first version of the proposed method [17], the state-of-the-art method [14] as well as the commercial package BookRestorer [30]. OCR testing is performed using ABBYY FineReader Engine 8.1 [33]. Both the distorted document images and the rectified documents are fed into OCR Engine for text recognition. Table 4 illustrates the average character accuracy as well as the average word accuracy.

Table 4. Average Character and Word Accuracy on 100 Document Images

Rectification Technique	#characters	Character Accuracy	#words	Word Accuracy
Without Rectification	170726	56,54%	27012	44,78%
Gatos et. al [14]	170726	81,51%	27012	62,71%
Proposed method TSD.ver1	170726	85,56%	27012	66,06%
BookRestorer [30]	170726	90,52%	27012	78,85%
Proposed method TSD.ver2	170726	93,82%	27012	84,07%

Semi-Automatic Evaluation: The evaluation methodology proposed in [18] avoids the dependence on an OCR engine or human interference. It is based on a point-to-point matching procedure using Scale Invariant Feature Transform (SIFT) [34] as well as the use of cubic polynomial curves for the calculation of a comprehensive measure which reflects the entire performance of a rectification technique in a concise quantitative manner. First, the user manually mark specific points on the distorted document image which correspond to N appropriate text lines of the document with representative deformation. Then, using SIFT transform, the marked points of the distorted document image are matched to the corresponding points of rectified document image. Finally, the cubic polynomial curves which fit to these points are estimated and are taken into account in the evaluation measure DW:

$$\boldsymbol{DW} = \frac{\sum_{j=1}^{N} DW_j}{N} \times 100\% \tag{5}$$

where DW_j is the measure which reflects the performance of the rectification technique with respect to the j^{th} selected text line. DW_j equals to one when the j^{th} selected text line in the rectified document image is a horizontal straight text line that is the expected optimal result. It shows that the rectification technique produces the best result. On the other hand, DW_j equals to zero when the rectified document image is equal to or worse than the original image. Therefore, DW ranges in the interval [0,…,100] and the higher the value of DW, the better is the performance of the rectification technique. Table 5 illustrates the average DW measure of all rectification methods. It is worth mentioning that the overall comparative ranking is the same with the one which is produced with the experiment that takes into account OCR performance. The proposed rectification method outperforms all the others methods.

Table 5. Comparative Results Using the Semi-Automatic Evaluation Methodology

Rectification Technique	**DM**
Gatos et. al [14]	79.35
Proposed method TSD.ver1	82.53
BookRestorer [30]	84.12
Proposed method TSD.ver2	91.90

4.3 Text Line Segmentation

To verify the validity of the proposed method we use two complementary line segmentation methods, projection profiles based on [35] and Adaptive RLSA based on [24]. In [35], each minimum of the profile curve is a potential segmentation point. Potential points are then scored according to their distance to adjacent segmentation points. The reference distance is obtained from the histogram of distances between adjacent potential segmentation points. The highest scored segmentation point is used as an anchor to derive the remaining ones. In [24], Makridis et. al propose the adaptive RLSA which is an extension of the classical RLSA in the sense that additional smoothing constraints are set in regard to the geometrical properties of neighbouring connected components. The replacement of background pixels with foreground pixels is performed when these constraints are satisfied.

We apply each method to a set of 50 historical documents images (1633 text line segments) as well as to a set of 50 handwritten documents (1144 text line segments). Then, using the two different segmentation results for each image, we generate a new segmentation result according to the proposed combination method [21].

For the purpose of the evaluation, we manually marked the correct line segments in the document images. The performance evaluation was based on counting the number of matches between the lines detected by the segmentation algorithms or their combination and the lines in the ground truth [36]. Finally, we calculate the detection rate (DR), the recognition accuracy (RA) as well as the F-Measure (FM). As depicted in Tables 6 and 7, the new segmentation result outperforms the two others methods and it increases the overall evaluation measure about 20%.

Table 6. Comparative Results - Historical Document Images

Segmentation Technique	GT regions	Result regions	One-to-one matches	DR (%)	RA (%)	FM (%)
Projection Profile [35]	1633	1577	1327	81.26	84.15	82.68
Adaptive RLSA [24]	1633	1594	1358	83.16	85.19	84.16
After combination using the proposed method [21]	1633	1605	1529	93.63	95.26	94.44

Table 7. Comparative Results - Handwritten Document Images

Segmentation Technique	GT regions	Result regions	One-to-one matches	DR (%)	RA (%)	FM (%)
Projection Profile [35]	1144	1248	841	73.51	67.39	70.32
Adaptive RLSA [24]	1144	1314	860	75.17	65.45	69.98
After combination using the proposed method [21]	1144	1152	1071	93.62	92.97	93.29

4.4 Character Segmentation

In order to record the efficiency of the proposed character segmentation method we followed a well established evaluation approach that is also employed by several document segmentation contests. The performance evaluation method is based on counting the number of matches between the entities detected by the algorithm and the entities in the ground truth. Finally, we calculate the detection rate (DR), the recognition accuracy (RA) as well as the F-Measure (FM). We used a set of 51 historical document images and compared with the commercial products ABBYY FineReader Engine 8.1 [33] and with the open source OCRopus libray [37] as well as with two state-of-the-art methods based on RLSA [22] and on Projection Profiles [23]. Table 8 presents the evaluation results.

Table 8. Character Segmentation - Evaluation Results

Segmentation method	GT regions	Result regions	One-to-one matches	DR (%)	RA (%)	FM (%)
Projection Profiles [23]	71818	71948	49449	68.85	68.73	68.79
RLSA [22]	71818	69065	56361	78.48	81.61	80.01
ABBYY FineReader Engine 8.1 [33]	71818	74721	52782	73.49	70.64	72.04
OCROpus [37]	71818	79575	53648	74.70	67.42	70.87
Proposed method [24]	71818	75955	62425	86.92	82.19	84.49

5 Concluding Remarks

In this thesis, we studied the preprocessing stages which are performed before the recognition process and we focused on the enhancement and segmentation of historical documents. Preprocessing stages play an important role in document image processing since they affect the performance of subsequent processing, such as optical character recognition. At the enhancement stage, we focused on the border removal as well as on the dewarping of document images, which are common problems associated with historical documents. Two methodologies that detect and remove black borders as well as noisy text regions are proposed. Furthermore, optimal page frames of double page document images are detected. Concerning the warping problem, a coarse-to-fine rectification methodology to compensate for undesirable document image distortions is proposed. To verify the validity of the proposed methodology, experiments have been carried out using indirect evaluation techniques as well as a novel semi-automatic evaluation methodology. At the document image segmentation stage we proposed a novel combination method of complementary text line segmentation techniques. Furthermore, a methodology for character segmentation in historical documents is suggested. Finally, in order to ease the construction of document image segmentation ground-truth that includes text-image alignment we presented an efficient technique.

References

1. D.X. Le, G.R. Thoma and H. Wechsler, "Automated borders detection and adaptive segmentation for binary document images", International Conference on Pattern Recognition, Vienna, Austria, 1996, pp. 737-741.
2. C. Fan, Y.K. Wang and T.R. Lay, "Marginal noise removal of document images", Pattern Recognition, vol. 35, no. 11, 2002, pp. 2593-2611.

3. B.T. Avila and R.D. Lins, “A New Algorithm for Removing Noisy Borders from Monochromatic Documents”, ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004, pp. 1219-1225.
4. B.T. Avila and R.D. Lins, “Efficient Removal of Noisy Borders from Monochromatic Documents”, International Conference on Image Analysis and Recognition, Porto, Portugal, 2004, pp. 249-256.
5. F. Shafait, J.v. Beusekom, D. Keysers and T.M. Breuel, “Document cleanup using page frame detection”, International Journal on Document Analysis and Recognition, vol. 11, no. 2, 2008, pp. 81-96.
6. N. Stamatopoulos, B. Gatos and A. Kesidis, “Automatic Borders Detection of Camera Document Images”, 2nd International Workshop on Camera-Based Document Analysis and Recognition, Curitiba, Brazil, 2007, pp. 71-78.
7. N. Stamatopoulos, B. Gatos and T. Georgiou, “Page Frame Detection for Double Page Document Images”, 9th International Workshop on Document Analysis Systems, Boston, MA, USA, 2010, pp. 401-408.
8. M.S. Brown and W.B. Seales, “Image restoration of arbitrarily warped documents”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 10, 2004, pp. 1295-1306.
9. C.L. Tan, L. Zhang, Z. Zhang and T. Xia, “Restoring warped document images through 3D shape modeling”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 2, 2006, pp. 195-208.
10. L. Zhang and C.L. Tan, “Warped image restoration with applications to digital libraries”, 8th International Conference on Document Analysis and Recognition, Seoul, Korea, 2005, pp. 192-196.
11. H. Ezaki, S. Uchida, A. Asano and H. Sakoe, “Dewarping of document image by global optimization”, 8th International Conference on Document Analysis and Recognition, Seoul, Korea, 2005, pp. 302-306.
12. A. Ulges, C.H. Lampert and T.M. Breuel, “Document image dewarping using robust estimation of curled text lines”, 8th International Conference on Document Analysis and Recognition, Seoul, Korea, 2005, pp. 1001-1005.
13. S.J. Lu, B.M. Chen and C.C. Ko, “A partition approach for the restoration of camera images of planar and curled document”, Image and Vision Computing, vol. 24, no. 8, 2006, pp. 837–848.
14. B. Gatos, I. Pratikakis and K. Ntirogiannis, “Segmentation Based Recovery of Arbitrarily Warped Document Images”, 9th International Conference on Document Analysis and Recognition, Curitiba, Brazil, 2007, pp. 989-993.
15. M.S. Brown and Y.C. Tsoi, “Geometric and shading correction for images of printed materials using boundary”, IEEE Transactions on Image Processing, vol. 15, no. 6, 2006, pp. 1544-1554.
16. N. Stamatopoulos, B. Gatos, I. Pratikakis and S.J. Perantonis, “Goal-oriented Rectification of Camera-Based Document Images”, IEEE Transactions on Image Processing, vol. 20, no. 4, 2011, pp. 910-920.
17. N. Stamatopoulos, B. Gatos, I. Pratikakis and S.J. Perantonis, “A Two-Step Dewarping of Camera Document Images”, 8th International Workshop on Document Analysis Systems Nara, Japan, 2008, pp. 209-216.
18. N. Stamatopoulos, B. Gatos and I. Pratikakis, “A Methodology for Document Image Dewarping Techniques Performance Evaluation”, 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 2009, pp. 956-960.
19. J.C. Handley, "Improving OCR accuracy through combination: a survey", International Conference on Systems, Man, and Cybernetics, California, USA, 1998. pp. 4330-4333.

20. S. V. Rice, J. Kanai, T. A. Nartker, "An Algorithm for Matching OCR-Generated Text Strings", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 8, Issue 5, 1994, pp. 1259-1268.
21. N. Stamatopoulos, B. Gatos and S.J. Perantonis, “A Method for Combining Complementary Techniques for Document Image Segmentation”, Pattern Recognition, vol. 42, no. 12, 2009, pp. 3158-3168.
22. K. Khurshidt, “Analysis and Retrieval of Historical Document Images”, PhD Thesis, Université Paris Descartes, 2009.
23. A. Antonacopoulos and D. Karatzas, “Semantics-based content extraction in typewritten historical documents”, 8th International Conference on Document Analysis and Recognition, Seoul, Korea, 2005, pp. 48–53.
24. N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos and N. Papamarkos, “Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths”, Image and Vision Computing, vol. 28, no. 4, 2010, pp. 590-604.
25. A. Toselli, V. Romero, E. Vidal, “Viterbi based alignment between text images and their transcripts” Workshop on Language Technology for Cultural Heritage Data, 2007, pp.9-16.
26. C.V. Jawahar, A. Kumar, “Content-level Annotation of Large Collection of Printed Document Images” Int. Conference on Document Analysis and Recognition, 2007, pp.799-803.
27. N. Stamatopoulos, G. Louloudis and B. Gatos, “A Comprehensive Evaluation Methodology for Noisy Historical Document Recognition Techniques”, 3rd Workshop on Analytics for Noisy Unstructured Text Data, Barcelona, Spain, 2009, pp. 47-54.
28. POLYTIMO project, http://iit.demokritos.gr/cil/POLYTIMO.
29. IMPACT project, European Community's Seventh Framework Programme under grant agreement N° 215064, http://www.impact-project.eu/
30. BookRestorer: http://www.i2s-bookscanner.com/
31. WiseBook: http://www.cadcam.org/wise-book.php
32. ScanFix Xpress: http://www.accusoft.com/scanfix.html
33. ABBYY FineReader: http://finereader.abbyy.com/
34. G. Lowe, “Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision, vol. 60, no. 2, 2004, pp. 91-110.
35. A. Antonacopoulos, D. Karatzas, "Document Image analysis for World War II personal records", First International Workshop on Document Image Analysis for Libraries, Palo Alto, 2004, pp. 336-341.
36. B. Gatos, N. Stamatopoulos and G. Louloudis, “ICDAR2009 Handwriting Segmentation Contest”, International Journal on Document Analysis and Recognition, vol. 14, no. 1, 2011, pp. 25-33.
37. The OCROpus open source document analysis system: http://code.google.com/p/ocropus/