

Automatic page analysis for the creation of a digital library from newspaper archives

B. Gatos¹, S.L. Mantzaris¹, S.J. Perantonis², A. Tsigris¹

¹Lambrakis Press S.A., 10 Heyden Str., 104 34 Athens, Greece;
Fax: +30-1/8250040; E-mail: slm@dolnet.gr

²Institute of Informatics and Telecommunications, National Research Center “Demokritos”, 15310 Athens, Greece;
E-mail: sper@iit.demokritos.gr

Received: 21 December 1998/Revised: 25 May 1999

Abstract. Digital preservation of newspaper archives aims both at the salvation of endangered material (paper) and at the creation of digital library services that will allow full utilization of the archives by all interested parties. In this paper, we address a series of issues pertaining to the retro-conversion of newspapers, i.e., the conversion of newspaper pages into digital resources. An integrated approach is presented that provides solutions to problems related to newspaper page image enhancement, segmentation of pages into various items (titles, text, images etc), article identification and reconstruction, and, finally, recognition of the textual components. Emphasis is placed on the most difficult intermediate stages of page segmentation and article identification and reconstruction. Detailed experimental results, obtained from a large testbed of old newspaper issues, are presented which clearly demonstrate the applicability of our methodology to the successful retro-conversion of newspaper material.

Key words: Digital preservation – Retro-conversion – Newspapers – Page segmentation – Identification and reconstruction

1 Introduction

Newspaper archives incorporate a number of facets concerning a country's social, political and economic history. Therefore, they are rightfully considered part of a country's national heritage. Digital preservation of such archives aims both at the salvation of endangered material (paper) as well as the creation of digital library services that will allow full utilization of the archives by all interested parties. In that sense, to achieve proper digital conversion of newspaper archives, one should take steps in a direction that will suit the needs of specialists

and enhance their research capabilities (e.g., journalists, historians, sociologists, etc.) but will also cater for information access facilities offered to the citizen.

A fundamental step in the digital retro-conversion of newspaper archives is the optimization of the methods relevant to the acquisition and the storage of information. More specifically, automatic document layout analysis and understanding is of paramount importance for any serious attempt to digitize an archive of considerable size, if one wishes to achieve a viable solution both in terms of time and cost.

A number of different problems need to be tackled before this procedure can be fully automated. These problems refer both to the actual input, such as image enhancement by noise removal from the scanned pages, as well as the logical organization of the input for storage purposes, for example the isolation of newspaper articles by document understanding techniques, such as segmentation and labeling. Successful tackling of these problems allows subsequent efficient cataloguing by employing OCR, full text retrieval and information extraction techniques, greatly reducing time and financial cost of manual indexing.

In this paper, an integrated approach is presented that provides solutions to problems related to newspaper page image enhancement, segmentation of pages into various items (titles, text, images etc), article identification and reconstruction, and, finally, recognition of the textual components. At the first stage of our workflow, we deal with image preprocessing (see Sect. 2) and specifically with image filtering for the improvement of image quality, as well as with skew correction in order to restore horizontal image status. For image enhancement, we use an accurate and quick method based on the transform of the binary image to grayscale according to the density of neighboring foreground pixels inside a window. Defining a threshold, we extract a shrunk image. In the

same way, using background pixel information we extract a swelled image. Successive application of shrinking and swelling procedure results in satisfactory image enhancement. For skew detection, we use a fast Hough transform approach based on the description of binary images using rectangular blocks. At the second stage, the main image components are automatically extracted (see Sect. 3). We propose a new technique for newspaper page segmentation based on gradual extraction of image components in the following order: lines, images and drawings, background lines, special symbols, text and title blocks. At the final stage, individual articles are traced and automatically recognized using suitable optical character recognition techniques (see Sect. 4). For article tracking, we followed a novel rule-based approach, which exploits the segment relationships that exist in the page layout format of newspaper pages. For the recognition of all text blocks, we integrated an OCR module which involves character segmentation, reliable feature extraction, and finally, fast and effective classification.

The testbed of our experiments is a collection of images from the newspaper “TO VIMA” published daily by Lambrakis Press S.A. from 1922 to 1982 and weekly from 1982 to the present. Lambrakis Press S.A. owns a large collection of newspapers and periodicals that consists of 1 300 000 pages and covers a time period from 1890 up to the present. This material is divided into 600 000 A2 pages, 500 000 A3 tabloid, and 200 000 A4 pages approximately. Our team is working on all aspects of the transformation procedure necessary to turn the printed material to an accessible digital archive (verification and quality control, digitization, cataloguing, search and retrieval, design and content presentation). The methods concerning automatic page analysis that are presented in this paper have already been applied with encouraging results even in cases where text regions and graphs co-exist in a particularly noisy environment.

2 Image preprocessing

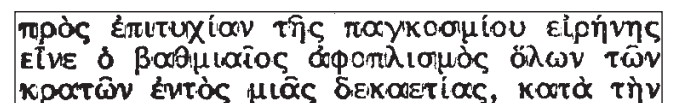
Retrospective conversion of published newspapers poses a number of difficulties concerning the actual digitization process. The hardware options on what is termed “newspaper scanners” are not wide and an effective workflow ought to be designed that will strike a balance between quality of scanning and time-efficiency if the physical archive is to be turned digital for further manipulation.

Due to the inherent low quality of old newspaper editions and to possible errors in the digitization process (the page is not positioned correctly on the scanner or the scanner automatic selection of binarization threshold is not correct) preprocessing of the image is essential before proceeding to other stages. This preprocessing mainly concerns image filtering for the improvement of image quality, as well as skew correction in order to restore horizontal image status.

Image filtering is mainly associated with problems such as noise elimination, isolated pixel removal, filling of possible breaks, gaps or holes, enhancement of character image body [17, 20]. Old newspaper images have extra noise between letters due to the old printing matrix quality or ink diffusion. The most common techniques for document image filtering are the application of morphological transformations to the original image [19]. According to these techniques, for each foreground pixel a structuring element is superimposed in order to achieve image shrinking or swelling.

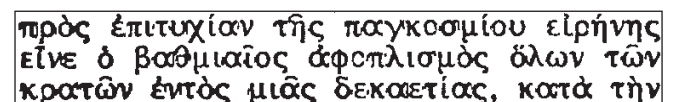
These are standard image processing operations. Shrinking refers to the replacement of foreground pixels belonging to thin image regions by background pixels, while swelling refers to the enhancement of thin regions by replacing neighboring background pixels by foreground pixels. We use a very accurate and quick method for image enhancement based on [38]. The image is transformed to grayscale as follows: a sliding window ($n \times n$) is moved over the image. For each position of the window the central pixel is assigned a gray level value proportional to the density of foreground pixels inside the window. Assigning a threshold, the extracted image is converted into binary, which is shrunk compared with the original image. Defining a threshold T , the shrunk binary image is extracted. Respectively, we use background pixel information in order to produce a swelled binary image. A window ($n \times n$) is moved over background pixels assigning for each pixel a value according to the density of background pixels inside the window. Assigning a threshold to the extracted grayscale image, the original image is swelled. Defining a threshold T , the swelled binary image is extracted.

Successive application of shrinking and swelling procedure results in satisfactory image filtering. An example of this procedure is demonstrated in Fig. 1. Figure 1a shows the original image and Fig. 1b shows the result of the above filtering procedure for a 5×5 window. In Fig. 1b, the application of a 5×5 mask leads to a satisfactory character restoration by enhancing the character and removing some of the noise. For example, notice that most of the spurious thin vertical lines originating from



πρὸς ἐπιτυχίαν τῆς παγκοσμίου εἰρήνης
εἶνε ὁ βαθμιαῖος ἀφοπλισμὸς ὅλων τῶν
κρατῶν ἐντὸς μιᾶς δεκαετίας, κατὰ τὴν

(a)



πρὸς ἐπιτυχίαν τῆς παγκοσμίου εἰρήνης
εἶνε ὁ βαθμιαῖος ἀφοπλισμὸς ὅλων τῶν
κρατῶν ἐντὸς μιᾶς δεκαετίας, κατὰ τὴν

(b)

Fig. 1. a Original image. b Image filtered with a 5×5 -window procedure, resulting in successful character enhancement and noise removal

low quality printing, which are present in Fig. 1a, have been removed in Fig. 1b.

For skew detection there are several methods, based mainly on the Hough transform [18, 27], projections [2, 8] the Fourier transform [37] as well as the correlation between vertical image lines [12, 13, 48]. The Hough transform based methods are the most popular but they are computationally expensive. Different fast Hough transform based methods have been proposed, including the creation of the grayscale “burst image” [18] and the use of only a selected square of the document where only bottom pixels of candidate objects are preserved [27]. In our pre-processing scheme, we use for skew detection a fast Hough transform approach based on the description of binary images using rectangular blocks [16, 36]. This technique has been developed by two of the authors in previous work. The Block Hough Transform (BHT) method takes advantage of the rectangular decomposition of a binary image and achieves fast evaluation of the Hough Transform field by analytically calculating the contribution to the cells in the Hough accumulator array of a whole rectangular block rather than that of each individual pixel. Newspaper images are mixed document images containing text with fonts of different sizes, images, lines, etc. A smoothing technique (such as RLSA [45]) helps the extraction of large size rectangular blocks in order to speed up the application of BHT.

3 Page segmentation

During the page segmentation stage, the main image components are automatically extracted. These components are: text, titles, images, lines (vertical and horizontal) and special symbols (such as an arrow indicating that an article continues on another page). Many algorithms for page segmentation have been proposed based on three fundamental approaches: on the smearing and labeling of regions [7, 10, 23], on the image profiling in various directions [28, 44] and on texture information [39–41, 47]. All techniques have not been successful in achieving newspaper segmentation because of the haphazard layout of newspaper articles and their very close contact. We propose a new technique for newspaper page segmentation based on gradual extraction of image components in the following order: Lines, images and drawings, background lines, special symbols, text and title blocks. It is essential to identify the background lines in order to preserve the small background vertical zones between text blocks. The methods applied for every category of image component to be extracted are as follows:

Lines. Vertical and horizontal line extraction is essential for two main reasons: first, because lines are used as tags for article identification (see Sect. 4), and secondly, because lines may be too close to text regions (as in tables or forms). Thus, text blocks will be more effectively

extracted if lines are removed from the original image. The most popular techniques for vertical and horizontal line identification are based mainly on the Hough Transform [9] as well as on morphological transformations [25]. We propose a new technique for line extraction, which combines fast implementation with accurate results. The algorithm we use is based on the following steps: a) image sub-sampling with respect to foreground pixels. Every foreground pixel of the original image is projected to the sub-sampled image Im_s (we used 1:4 sub-sampling) so even dotted lines are transformed to continuous lines; b) from the image Im_s we extract two gray scale images Im_V and Im_H , returning for every foreground pixel the length of the vertical (for Im_V) or horizontal line (for Im_H) it belongs to; c) the resulting images are thresholded so only pixels belonging to large lines remain; d) large lines are defined using the Contour Following technique [35]. The method has tolerance to slight line skewing as well as to dotted or broken lines. The parameterization of the method includes minimum length, maximum width, as well as maximum gap of the extracted line segments. Figure 2 illustrates the performance of the proposed line extraction algorithm: Fig. 2a is the original image and Fig. 2b shows the extracted vertical and horizontal lines.

Images and drawings. In newspaper pages, the existence of images and drawings within a very short distance of text is very frequent. Identifying and extracting the images and drawings from the original newspaper image is essential before proceeding to text extraction phase. Most approaches to this first attempt to detect major component blocks and then use simple statistical tests to classify them as text or images [3, 31, 45, 46]. In newspaper images, it is rather difficult to detect major component blocks because different elements are too close (an image and its corresponding caption may interfere). For this reason, we use a novel technique that is based on primary search only for image segments. In this stage, our objec-

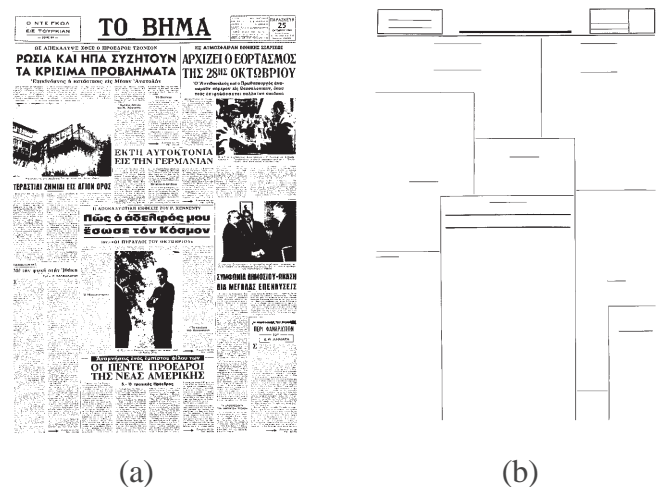


Fig. 2. a Original image. b Vertical and horizontal line extraction

tive is to detect composite regions whose surrounding box height will help us distinguish text and image/drawing regions. Therefore, we try to keep text lines separate while at the same time connecting different regions of the same image. This is done by subsampling the image only with respect to background pixels. This stage often suffices to distinguish between text and images/drawings. For possible image segment extraction we use the Run Length Smoothing Algorithm (RLSA) [34, 45] focused on background pixels. According to this technique, successive background pixels forming a segment of small length are transformed to foreground pixels. We extract all segments using Connected Component Analysis [23, 45].

After this, we ensure the existence of images or drawings by analyzing the Fast Fourier Transform (FFT) of the horizontal projections of segmented regions. The horizontal projections of text blocks exhibit a more or less periodical structure because of the presence of equally spaced text lines. Therefore, by transforming these projections to the frequency domain using the FFT, we expect the presence of dominant non-zero frequencies. On the other hand, the horizontal projections of images and drawings do not normally exhibit a periodical structure, so that a transformation to the frequency domain does not result in the presence of dominant frequencies. Figure 3 shows the difference of the corresponding FFT field for a text area (Fig. 3a), and an image area (Fig. 3b). Only in Fig. 3a (text area) is there a dominant frequency of a significant amplitude.

Background lines. The same technique as the one we used for vertical foreground lines is applied, this time working with background pixels. All background vertical lines found constitute separating border between two close neighboring text columns and are identified in order to help the text block extraction stage. This procedure is essential for processing old newspapers with dense lay-

out, i.e., very small horizontal or vertical spacing between neighboring segments.

Special symbols. Locating special symbols in newspaper images helps defining special regions (such as references, which are text regions lying on the right side of arrows indicating that an article continues on another page, as well as special symbols defining separating tags between articles). Connected Component Analysis helps us locate special symbols according to their geometric features (such as the ratio of height to width). To ensure the existence of special symbols we use a pattern matching technique [20].

Text and title blocks. At this final stage of page segmentation, we must extract text and title blocks from the remaining image (all the above mentioned components have been extracted). We propose a novel method for text block extraction that is based on RLSA with adaptive parameters. We first proceed with a connected component analysis of the image and every foreground pixel is assigned a value according to the height of the box of its connected area. After this procedure, the image is converted to gray scale and every pixel has an approximate classification to either normal text or title according to its gray scale value. At the same time, possible remaining image noise is rejected. Next, we proceed with a smoothing of the image using RLSA with smoothing factors depending on the first classification of every pixel to text or title. From this smoothing procedure we exclude all background pixels belonging to the already located vertical background lines. Using a contour following technique, we manage to extract all text and title regions with great accuracy. To speed up the entire text and title extraction process we use an image sub-sampling with respect to background pixels (we used 1:2 sub-sampling). In this way, we also ensure that vertical neighboring letters will not connect, which is essential for discriminating text against titles.

Finally, newspaper page decomposition is obtained after a heuristic labeling of certain segments such as captions which are text segments lying below images and text

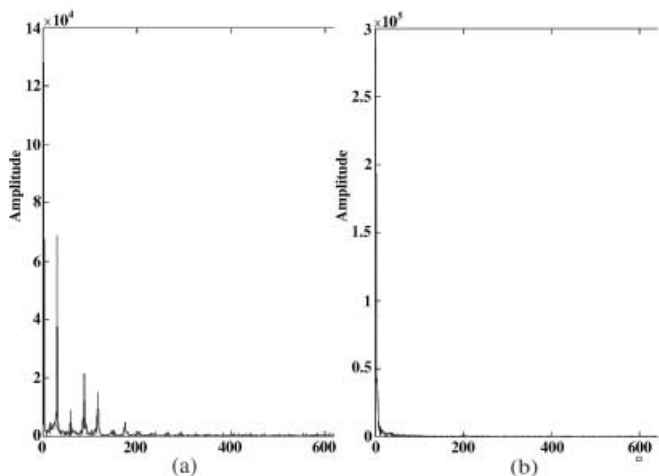


Fig. 3a,b. FFT field for a text area (a) and an image area (b). In a a dominant frequency is observed at 30 Hz, while in b there are no dominant frequencies

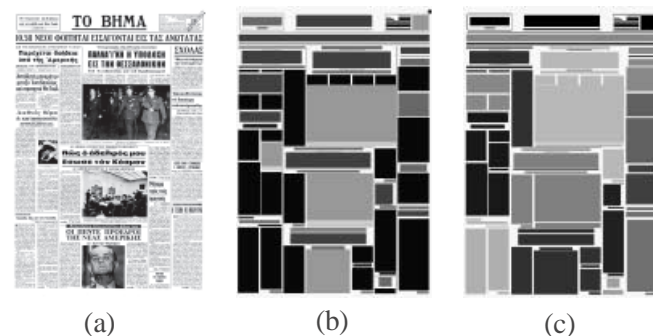


Fig. 4a–c. Final newspaper image decomposition. a Original image, b segmented image, c identified articles

titles which are one-line text segments. Figure 4 shows the final page decomposition of a newspaper image. Figure 4a shows the original image, while Fig. 4b shows the decomposition of the image to its basic components.

4 Article identification and recognition

4.1 General remarks

The primary unit of information in a newspaper is the article. The main purpose of automatic newspaper page analysis is to build suitable digital representations for the contents of an article. To this end, two successive processes must be applied to the segmented newspaper items. First, it is required to digitally reconstruct articles by identifying, gathering and representing in digital form all individual components (headline, text, images etc.) of each specific article. At a second stage, it is very important to automatically recognize the textual content of the individual components using suitable optical character recognition techniques. Using state of the art methods, it is possible to achieve very high recognition rates at the OCR stage. Therefore, it is obvious that the main burden for achieving reliable digital representation of the article content falls on the article reconstruction stage. Thus, in this paper we place particular emphasis on the first task of article components reconstruction. Our related methodology is presented in the next subsection and a full evaluation is given in Sect. 5. For reasons of completeness, we also present in Sect. 4.3 the OCR techniques used for the recognition of the textual parts of the identified articles.

4.2 Article components reconstruction

After segmenting a newspaper page, it is very important to automatically identify the articles on a page, since articles are the milestones of a digital library based on newspaper material. An article is considered to consist of a headline and a collection of segments having a logical type among the following: over-headline, sub-headline, title, text, picture, caption and article continuation indicators (references). Headlines, over-headlines and sub-headlines are labeled after sufficient evidence for the starting point of an article in the newspaper page is found. A segment that is not associated to any of the headlines of a page is considered to be unassigned. Unassigned blocks are usually the continuation of an article, which has started from another page of the newspaper issue.

An overview of the current techniques applied in order to solve document image understanding problems can be found in Jain et al., [21] and in Summers [42]. However, the overwhelming majority of the researchers are addressing the problem for scientific journals, which makes their results of less importance for understanding newspaper

pages, since newspapers have a completely different and considerably more complex page layout. Our approach exploits the segment relationships that exist in the page layout of a newspaper. These relations are depicted as a set of rules. A similar approach is given by Niyogi [32]. However, the set of rules given by Niyogi is inadequate to handle the large variety of page layout of our testbed collection. The aim of our rules is threefold. The first objective is to distinguish among segments of the same type those that separate articles. For example, if we consider horizontal lines, our rules try to distinguish them either as underlines or as lines that separate articles. The second objective is to label master title, title and text title segments as headlines, over-headlines or sub-headlines according to their segment height and position from the starting point of an article. Finally, the third objective is to group a number of segments that have different types around a headline. The aforementioned rules exceed 40 in number and can be grouped in the following categories:

- a) Rules determining the header area of the page. These rules try to identify the longest horizontal line in the top area of the page that exceeds a threshold length value.
- b) Rules that distinguish underlines from lines that divide articles. These rules consider as underlines short lines, lines under master titles, titles etc. In contrast, lines below a text segment are considered as lines that divide articles.
- c) Rules that locate white rectangular areas on the page that separate articles. These rules consider, among other things, white areas that exceed a threshold in height.
- d) Rules that identify segments which are in the border of an article. These rules locate special patterns that are used to separate neighboring articles.
- e) Rules that find the headlines. These rules consider groups of neighboring master titles, titles and text titles below a horizontal line as headlines.
- f) Rules that specify the over-headline and the sub-headline that possibly accompany a headline. These rules classify titles and text tiles as over-headlines or sub-headlines according to their relative position to a headline.

An example of a newspaper article identification is depicted in Fig. 4c.

4.3 Optical character recognition

The recognition of all text blocks is of great interest in digital newspaper library design. Using recognized text as an input to an effective Text Retrieval system boosts the efficiency and the usability of the digital library. We integrated an OCR module developed in earlier work of some of the authors [11] which involves the stages of character segmentation, the stage of extracting reliable features for every character and finally, the stage of recognition using

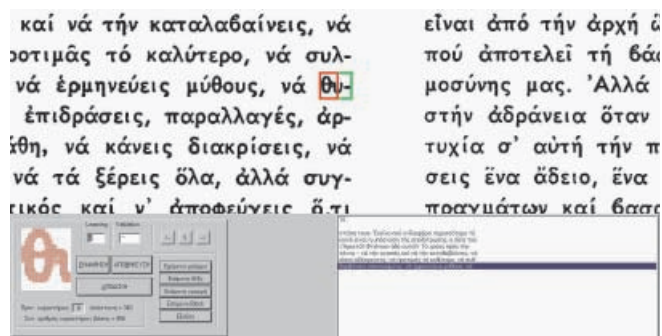


Fig. 5. OCR module for newspaper recognition

a fast and effective classifier. In Fig. 5 our OCR module is demonstrated.

The proposed techniques for character segmentation involve construction of a decision tree for resolving ambiguities in segmenting touching characters [29, 43], recursive algorithms for merged characters [5], and a Hidden Markov Model to the touching and degraded text [4]. We used a dynamic recursive segmentation algorithm specially for merged characters based on pixel and profile projections [30]. The pixel projection consists of the total number of foreground pixels in each vertical column. The profile projection is the profile of the external contour of the text line as seen from the top or bottom. The cutting points for touching characters are obtained in the horizontal positions where the segmenting objective function is maximized.

Many OCR systems have been proposed based on statistical, matching, transform and shape features [1, 6, 14, 33], combining a feature-extraction technique with a classifier scheme, such as: features based on the application of the distance transform in combination with a k-Nearest-neighbor classifier [26], end-points, corners and tees extraction as features and a binary decision tree classifier [15, 24], shape clustering feature extraction and

statistical Bayesian classifier [22]. The choice of appropriate features in OCR techniques is of utmost importance. Some of the characteristics that must be embodied in the feature extraction scheme are: discrimination, reliability, independence, small feature space and low computational cost. To this end, we selected as features appropriately weighted averages originating from the application of overlapping Gaussian masks to the character body [11]. This feature extraction scheme achieves great tolerance to noise and deformation with simultaneous low computational complexity. Using those features as an input to a k-Nearest-neighbor classifier we achieve high recognition rates which are in the range of 99.7%.

5 Experimental results

All the proposed techniques for newspaper page analysis and identification were tested extensively on a large set of newspaper images. The experimental results that follow prove the efficiency and the accuracy of the proposed methodology for segmentation and component labeling as well as for article identification.

The test set contains 100 pages from the newspaper «TO VIMA» with publication dates from 1965 to 1974. The basic characteristic of the test set is the variety in the page layout as well as noise existence due to low-quality paper. The test set is divided into first, middle and last page sets (13, 75, and 12 pages respectively) in order to extract conclusions for the behavior of the page analysis techniques in those three main newspaper page categories. All experiments were implemented on a PC with a Pentium II 350 MHz processor and the average time for page analysis (segmentation and article tracking) was 9 sec.

The experimental results are summarized into three tables. Table 1 demonstrates the efficiency of image decomposition into its basic components, Table 2 presents

Table 1. Segmentation and labeling results

	First pages		Middle pages		Last pages		Total	
	Recall %	Precision %	Recall %	Precision %	Recall %	Precision %	Recall %	Precision %
Master title	96.07	89.68	95.66	83.69	96.95	89.79	95.86	85.20
Title	88.56	92.73	84.53	86.38	92.34	91.83	85.99	87.85
Text title	94.08	88.97	83.15	88.85	90.61	92.45	85.46	89.29
Text	98.61	98.02	97.23	95.68	97.82	96.45	97.48	96.07
Image	98.07	94.87	98.00	99.20	100	95.83	98.24	98.23
Caption	79.48	89.74	97.2	97.2	58.33	41.66	90.23	89.56
Reference	82.05	100	100	100	–	–	85.66	100
Horiz. line	98.08	95.34	97.56	94.43	98.32	96.02	98.19	95.74
Vert. line	98.14	99.06	98.44	98.95	98.26	99.00	98.37	98.97
Headline	92.37	94.12	90.82	87.56	93.71	94.38	91.36	89.23
Over-headline	78.18	100	89.00	69.00	81.90	96.38	86.74	76.31
Sub-headline	81.54	81.12	73.26	75.26	80.03	91.78	75.14	78.00

Table 2. Confusion matrix

FOUND	REAL						
	Master title	Title	Text title	Text	Image	Caption	Other
Master title	85.20	11.02	0	0.9	1.80	0	1.08
Title	0.97	87.85	6.85	1.97	0	1.97	0.39
Text title	0	0	89.29	3.58	0	1.29	5.84
Text	0	0	2.41	96.07	0	0.82	0.70
Image	0	0	0	1.77	98.23	0	0
Caption	0	0	3.25	7.19	0	89.56	0

the confusion matrix among different segmentation items and Table 3 demonstrates the efficiency of the technique for identifying the articles the newspaper image consists of. In order to trace the correctness of the identified elements, (segments or articles) we calculated the recall and precision values. Recall value is the number of correct elements found divided by the total number of elements we are looking for. Precision value is the number of correct elements found divided by the total number of elements found. Recently, Summers [42] used the same criteria in order to evaluate her techniques. In Table 1, recall and precision values are calculated for all basic image segments found, analytically for the three page categories (first, middle and last pages). In Table 2, the confusion matrix for the page segmentation phase is presented. Thus, an analysis of the segment categories that are erroneously identified can be carried out. In Table 3, recall and precision values are calculated for article tracking. An identified article is regarded as correct only if all its segments are grouped together correctly. Additionally, the number of segments that are classified correctly is calculated. As we can observe from the tables:

- Image segmentation into main components has success rates over 85% in both recall and precision calculation.
- Significant success is recorded in tracing text, images, vertical, and horizontal lines. Recall and precision rates exceed 96%.
- The segmentation technique mostly confuses: master titles with titles, titles with text titles, text titles with text, and captions with text.

Table 3. Article tracking results

	Recall %	Precision %	Segments with correct article classification %
First pages	81.54	81.12	93.21
Middle pages	72.43	74.38	89.27
Last pages	85.66	90.18	93.01
Total	75.20	77.15	90.23

- The article tracking methodology exhibits success rates over 75% in both recall and precision calculation. This result does not seem impressive compared to the previous results given above. However, there are two reasons that explain this behavior. First, article tracking is very sensitive to errors; if a segment is incorrectly classified to an article, then two articles are incorrect, one that has an additional segment and one that has a missing segment. Calculating the segments that are classified correctly, we get a success rate of 90% or greater. Secondly, there are many exceptions to the page layout format of the newspaper page that is not followed faithfully by the typesetters. This is particularly true for the middle pages.

An interesting overall observation is that our techniques achieve high recall and high precision at the same time. A similar effect has appeared in Summers [42] also. In other areas where these measures are used, i.e., Information Retrieval, it is common that there is a trade-off between recall and precision. It is interesting to further exploit this difference in behavior.

6 Conclusions

In this paper, a number of problems regarding the digital retro-conversion of newspapers were addressed. Image enhancement, segmentation, article tracking, and OCR can be done economically and in a fraction of the time that manual approaches require. The experimental results presented here clearly indicate that automating the process of converting newspaper pages to digital resources is feasible with adequate success. These digital resources can be the foundation of a digital library. We believe that our approach can be extended to cover other materials, important to the preservation of cultural heritage, materials that are currently available only in libraries and archives.

References

1. Abdelazim, H.Y., Hashish, M.A.: Automatic reading of bilingual typewritten text. Proc. VLSI and Microelectronic Applications in Intelligent Peripherals and their Application Networks, 1989, pp. 140–144

2. Baird, H.S.: The skew angle of printed documents. Proc. SPSE 40th Conf. Symp. Hybrid Imaging Systems, Rochester, New York, 1987, pp. 21–24
3. Bixler, J.P.: Tracking text in mixed-mode documents. Proc. ACM Conf. On Document Processing Systems, 1988, pp. 177–185
4. Bose, C.B., Kuo, S.: Connected and degraded text recognition using Hidden Markov Model. Pattern Recognition 27(10):1345–1363, 1994
5. Casey, R.G., Nagy, G.: Recursive segmentation and classification of composite character patterns. Proc. 6th Int. Conf. Pattern Recognition, 1982, pp. 1023–1026
6. Cash, G.L., Hatami, M.: Optical character recognition by the method of moments. Computer Vision, Graphics and Image Processing 39:291–310, 1987
7. Chauvet, P., Lopez-Krahe, J., Taffin, E., Maitre, H.: System for an intelligent office document analysis, recognition and description. Signal processing 32:161–190, 1993
8. Ciardiello, G., Scafur, G., Degrandi, M.T., Spada, M.R., Rocoteli, M.P.: An experimental system for office document handling and text recognition. Proc. 9th Conf. On Pattern Recognition, 1988, pp. 739–743
9. Duda, R.D., Hart, P.E.: Use of the Hough transform to detect lines and curves in pictures. Commun. ACM, 1972, pp. 11–15
10. Fan, K., Liu, C., Wang, Y.: Segmentation and classification of mixed text/graphics/image documents. Pattern Recognition Letters 15:1201–1209, 1994
11. Gatos, B., Karras, D., Perantonis, S.: Optical character recognition using novel feature extraction & neural network classification techniques. Proc. Workshop on Neural Network Application and Tools, 1993, pp. 65–72
12. Gatos, B., Papamarkos, N., Chamzas, C.: Skew detection and text line position determination in digitized documents. Pattern Recognition 30(9):1505–1519, 1997
13. Gatos, B., Papamarkos, N., Chamzas, C.: Skew detection of digitized documents using cross-correlation. Proc. 5th Int. Conf. On Advances in Communication & Control - Telecommunications/Signal Processing in the Multimedia Era, COMCON 5:124–130, 1995
14. Gatos, B., Papamarkos, N., Chamzas, C.: Using curvature features in a multiclassifier OCR system. Engineering Applications of Artificial Intelligence 10(2):213–224, 1997
15. Gatos, B., Papamarkos, N.: A novel method for character recognition. Proc. COMCON 4, 1993, pp. 493–503
16. Gatos, B., Perantonis, S.J., Papamarkos, N.: Accelerated Hough transform using rectangular image decomposition. Electronic Letters 32(8):730–732, 1996
17. Gonzalez, C., Wintz, P.: Digital Image Processing. 2nd ed., Reading, MA: Addison-Wesley, 1987
18. Hinds, S.C., Fisher, J.L., D'Amato, D.P.: A document skew detection method using run-length encoding and the Hough transform. Proc. 10th Int. Conf. On Pattern Recognition, 1990, pp. 464–468
19. Impedovo, S., Otaviano, L., Occhinegro, S.: Optical character recognition - a survey. Int. J. Pattern Recognition and Artificial Intelligence 5:1–23, 1991
20. Jain, A.K.: Fundamentals of Digital Image Processing. Englewood Cliffs, NJ: Prentice-Hall, 1989
21. Jain, A.K., Yu, B., Cash, G.L., Hatami, M.: Document representation and its application to page decomposition. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(3):294–308, 1998
22. Kahan, S., Pavlidis, T., Baird, H.S.: On the recognition of printed characters of any font and size. IEEE Trans. on Pattern Analysis and Machine Intelligence 9:274–287, 1987
23. Kasturi, R., Bow, S., El-Masri, W., Shah, J., Gattiker, J., Mokate, U.: A system for interpretation of line drawings. IEEE Trans. on Pattern Analysis and Machine Intelligence 12: 978–991, 1990
24. Kerrick, D., Bovik, A.: Microprocessor-based recognition of handprinted characters from a tablet input. Pattern Recognition 21(5):525–537, 1988
25. Kong, B., Chen, S., Haralick, R.M., Phillips, I.T.: Automatic line detection in document images using recursive morphological transforms. IS&T/SPIE Symp. On Elec. Imag., DR-II, 1995, pp. 163–174
26. Kovacs, Z.S., Guerrieri, R.: Massively-parallel handwritten character recognition based on the distance transform. Pattern Recognition 28(3):293–301, 1995
27. Le, D.S., Thoma, G.R., Wechsler, H.: Automated page orientation and skew angle detection for binary document image. Pattern Recognition 27(10):1325–1344, 1994
28. Lettera, C., Maier, M., Paoli, C.: Character Recognition in Office Automation. Advances in Image Processing and Pattern Recognition. In: Cappellini, V., Marconi, R. (eds.): North-Holland: Elsevier Science, pp. 191–197, 1986
29. Liang, S., Shridhar, M., Ahmadi, M.: Segmentation of touching characters in printed document recognition. Pattern Recognition 27(6):825–840, 1994
30. Lu, Y.: Machine printed character segmentation - an overview. Pattern Recognition 28(1):67–80, 1994
31. Maier, M., Porinelli, R.: Separating graphic objects in written texts. Advances in Image Processing and Pattern Recognition, 1986
32. Niyogi, D.: A Knowledge-Based Approach to Deriving Logical Structure from Document Images. Ph.D Thesis, State University of New York, Buffalo, 1994
33. Papamarkos, N., Spiliotis, I., Zoumadas, A.: Character recognition by signature approximation. Int. J. Pattern Recognition and Artificial Intelligence 8(5):1171–1187, 1994
34. Papamarkos, N., Tzortzakis, J., Gatos, B.: Determination of run-length smoothing values for document segmentation. 3rd IEEE Int. Conf. on Electronics, Circuits and Systems, 1996. ICECS 96, pp. 684–687
35. Pavlidis, T.: Algorithms for Graphics and Image Processing. Rockville, MD: Computer Science Press, 1992
36. Perantonis, S.J., Gatos, B., Papamarkos, N.: Image segmentation and linear feature identification using rectangular block decomposition. 3rd IEEE Int. Conf. on Electronics, Circuits and Systems, 1996. ICECS 96, pp. 183–186
37. Postl, W.: Detection of linear oblique structures and skew scan in digitized documents. Proc. 8th Int. Conf. on Pattern Recognition, 1986, pp. 687–689
38. Schilling, R.J.: Fundamentals of Robotics Analysis and Control. Englewood Cliffs, NJ: Prentice-Hall, 1990
39. Strouthopoulos, C., Papamarkos, N., Chamzas, C.: Identification of text-only areas in mixed type documents. IEEE Workshop on Nonlinear Signal and Image Processing, 1995, pp. 162–165
40. Strouthopoulos, C., Papamarkos, N., Chamzas, C.: Identification of text-only areas in mixed type documents. Engineering Applications of Artificial Intelligence 10(4):387–401, 1997
41. Strouthopoulos, C., Papamarkos, N.: Text identification for document image analysis using a neural network. Image and Vision Computing-Special Issue on Document Image Processing and Multimedia Environments 16(12–13):879–896, 1998
42. Summers, K.M.: Automatic Discovery of Logical Document Structure. Ph. D. Thesis, Cornell University, 1998
43. Tsujimoto, S., Asada, H.: Resolving ambiguity in segmenting touching characters. First Int. Conf. On Document Analysis and Recognition, 1991, pp. 701–709
44. Verikas, A., Bachauskene, M., Vilunas, S., Skaisgiris, D.: Adaptive Character Recognition System. Pattern Recognition Letters 13:207–212, 1992
45. Wahl, F.M., Wong, K.Y., Casey, R.G.: Block Segmentation and Text Extraction in Mixed Text/Image Documents. Computer Graphics and Image Processing 20:375–390, 1982
46. Wang, D., Shihari, S.N.: Classification of newspaper image blocks using texture analysis. Computer Vision Graphics and Image Processing 47:327–352, 1989
47. Williams, P.S., Alder, M.D.: Generic Texture Analysis Applied to Newspaper Segmentation. International Conference on Neural Networks. ICNN'96
48. Yan, H.: Skew correction of document images using inter-line cross-correlation. Comput. Vision Graphics Image Process.: Graphical Models and Image Process. 55(6):538–543, 1993