Integrated Algorithms for Newspaper Page Decomposition and Article Tracking

B. Gatos [1], S. L. Mantzaris [1], K. V. Chandrinos [2], A. Tsigris [1], S. J. Perantonis [2]

[1] Lambrakis Press S.A., 8 Heyden Str,
104 34 Athens, Greece
bgat@dolnet.gr
[2] Institute of Informatics and
Telecommunications,
National Research Center “Demokritos”,
153 10 Athens, Greece
sper@iit.demokritos.gr

Abstract. *The conversion of newspaper pages into digital resources is an important task that greatly contributes to the preservation and access to newspaper archives. In this paper, an integrated methodology is presented for segmenting newspaper page and identifying newspaper articles. In a first stage, a succession of image processing and document analysis algorithms is employed for segmenting newspaper page images into various objects (text, images and drawings, titles). A rule based approach is subsequently applied to the objects identified during the page segmentation phase for reconstructing individual articles. Experimental results, obtained from a large testbed of old newspaper issues, are presented which clearly demonstrate the applicability of our integrated approach to successful newspaper page segmentation and identification of newspaper articles.*

Keywords: Page segmentation, page layout analysis, document understanding, region identification, article tracking, newspaper page analysis

1 Introduction

Automatic decomposition of digitized newspaper pages can result in cost-effective solutions concerning the transformation of paper archives to digital collections. A number of techniques have been proposed in the literature describing algorithms to facilitate automatic page decomposition. However, many of these algorithms are not directly applicable to newspaper images, which present special problems. The most significant problems identified include the complex layout of newspaper pages, in particular the oldest ones, where text columns are located very close to each other in a haphazard way, as well as the poor scanning results obtained because of low print quality or deterioration through time. Another important problem was the repeated changes in layout habits through time.

Bearing all the above in mind, we proceeded in designing and implementing a set of algorithms that could in effect automate the mark-up procedure of page segments and allow, through post-processing of automatically segmented pages, a rule-based approach to article tracking and identification.

We evaluated our methods in the framework of an extensive newspaper digitization project at the Lambrakis Press S.A. The company owns newspapers ranging in publication time from 1890 to date, amounting to a total of 1.3 million pages. For this work, we used archival copies of the newspaper "TO VIMA", published daily from 1922 to 1982 and weekly thereof.

This paper is organized as follows: In Section 2 we present the algorithms for page segmentation. In Section 3, we present our rule-based system for article identification. Section 4 has experimental results, while Section 5 concludes this paper and sets out future plans.

2 Page Segmentation

At the page segmentation phase, the newspaper image is decomposed in its basic principal components, such as text areas, titles, images, lines etc. The haphazard layout of newspaper articles and the close contact of different segments hinders the application of many standard page segmentation algorithms (e.g. [5], [7]). We propose a new technique for newspaper page segmentation based on smearing and labeling of regions and on gradual extraction of image components in the following order: Lines, images and drawings, background lines, text and title blocks. Background lines are narrow foreground blank stripes between text columns. It is essential to preserve vertical background lines as a border between nearly touching text columns. Our component extraction methods are described in the susbequent subsections.

2.1 Line extraction

Horizontal and vertical foreground lines are the first elements to be identified because they usually are too close to other kinds of segments, hence all other region categories (text, images, etc.) will be more effectively extracted if lines are removed from the original image. Moreover, it is essential to extract horizontal and vertical foreground lines because they are used as tags for article identification (see Section 3). Main approaches for line identification are based on Hough Transform as well as on morphological transformations ([2]). We propose a new technique based on grayscale transformation of the binary image, assigning values for every pixel according to the leng th of the horizontal or vertical line every pixel belongs to. This technique combines fast implementation with accurate results and can be summarized in the following steps:

Step 1: Image sub-sampling with respect to foreground pixels.

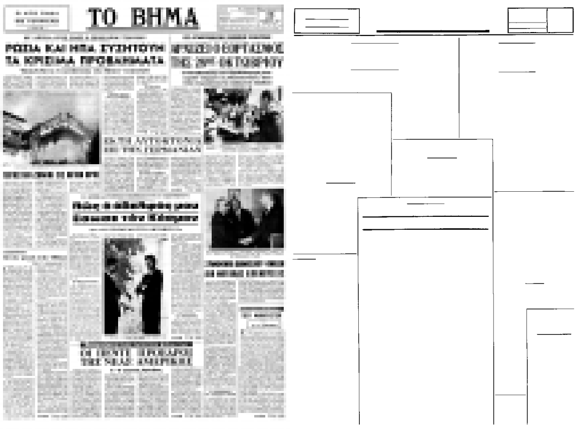
Step 2: From the resulting image we extract two grayscale images F_H and F_V, assigning to every foreground pixel the length of the vertical (for F_V) or horizontal line (for F_H) it belongs to.

Step 3: Lines that are interesting to identify are characterized by an a priori known minimum length and maximum width. Images resulting from step 2 are thresholded so that the only remaining pixels belong to lines whose length/width are respectively larger/smaller than a threshold. We extract the binary images L_H and F_V.

Step 4: Lines are defined using a Contour Following technique for all points of L_H and F_V.

The proposed method for line extraction has tolerance to slight line skewing (lines have successive horizontal/vertical pixels even if they have a slight slope) and to the presence of broken lines (due to sub-sampling with respect to foreground pixel information). Application of the method is shown in Fig. 1.

For background line extraction, the same technique is applied, this time working with background pixels.



(a) (b)

Fig. 1. (a) Original image, (b) Vertical and horizontal line extraction

2.2 Image and drawing extraction

The presence of nearly touching images (or drawings) and text segments is very frequent in newspaper pages. This limits the usefulness of many common approaches to image and drawing identification, which attempt to detect major component blocks and then use simple statistical tests to classify them as text or images ([7],[8]). For example, an image and its corresponding caption may interfere and blocks cannot be detected. For this reason, we use a novel technique which is based on a preliminary search only for image segments. In this stage, our objective is to detect composite regions whose surrounding box height will help us distinguish text and image or 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 (or drawings). Results from the preliminary stage are improved by a further check based on the FFT of the horizontal projections of segmented regions. The algorithm we use for image and drawing extraction is summarized in the following steps:

Step 1: We employ sub-sampling with respect to background pixels.

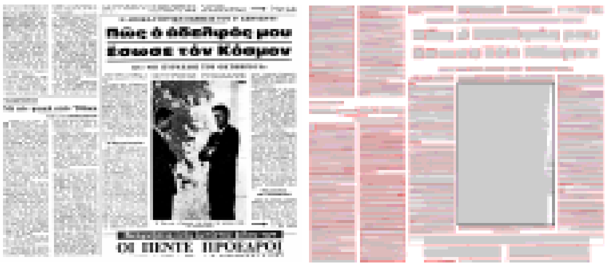
Step 2: From the resulting image we extract two grayscale images B_H and B_V, returning for every background pixel the length of the vertical (for B_V) or horizontal background line (for B_H) it belongs to. Candidate image/drawing points are 1s in the IDR table:

$$\mathrm{IDR}(x,y)=\begin{cases}1, & \text{if } B_H(x,y)<c_H \ \wedge \ B_V(x,y)<c_V \\ 0, & \text{otherwise}\end{cases} \qquad [1]$$

Parameters C_H and C_V correspond to maximum length of horizontal and vertical background lines.

Step 3: We extract segments using Connected Component Analysis ([7]) of the image IDR. From all resulting segments, we select those with significant height greater than this of the largest expected title letter. An example for this procedure is given at Fig. 2.

Step 4: For each extracted area, horizontal projections are used. By analyzing their FFT we may find dominant frequencies, signifying the periodical structure of a text region. If no dominant frequencies are found, the corresponding block is classified as image/drawing.



(a) (b)

Fig. 2. (a) Original image, (b) Image IDR with all segments defined using Connected Component Analysis.

2.3 Text and Title block extraction

At this final page segmentation stage, we extract text and title areas from the remaining image. Title blocks are distinguished to standard title and master title blocks according to letter height. We propose a novel method for text and title block extraction which is based on RLSA ([4],[7]) with adaptive parameters. We first follow step 1 of section 2.1 and steps 1,2 of section 2.2 to extract images B_H and B_V, using sub-sampling with respect to background pixels (1:2 ratio). We thus ensure that vertical neighboring letters will not connect, which is essential for discriminating text against titles. Then we proceed with connected component analysis and every foreground pixel is assigned a value according to the height of the box of its connected area. In this way, starting from the original image we extract grayscale image H_S. After this procedure, every pixel has an approximate classification to either text or title according to its grayscale value. At the same time, possible remaining image noise is rejected. A new image T_S is extracted with values: 0 for noise pixels, 1 for text pixels, 2 for standard title pixels and 3 for master title pixels. Image T_s is given by the formula:

$$T_s(x,y) = \begin{cases} 0, \text{if } H_s(x,y) < t_1 \\ 1, \text{if } H_s(x,y) \geq t_1 \wedge H_s(x,y) < t_2 \\ 2, \text{if } H_s(x,y) \geq t_2 \wedge H_s(x,y) < t_3 \\ 3, \text{if } H_s(x,y) \geq t_3 \end{cases} \quad [2]$$

where t_1, t_2, t_3 are respectively minimum expected letter height of a text segment, a title and a title region.

Next, we proceed with image smoothing using adaptive RLSA with smoothing factors depending on the first classification of every pixel to text or title (background pixels belonging to the already located vertical background lines are excluded).

Using a contour following technique, we manage to extract all text and title regions with great accuracy. The final newspaper page decomposition (see Fig. 3b) is obtained after a heuristic labeling of certain segments such as captions which are text segments lying below images and text titles which are one-line text segments.

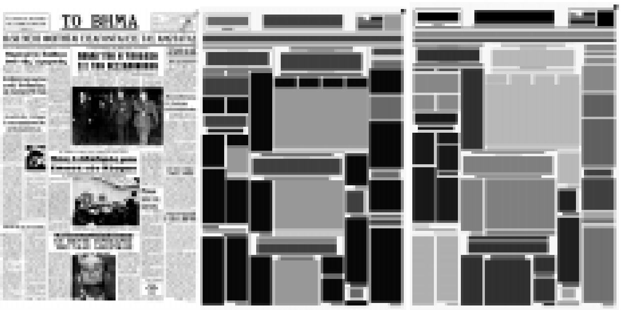


Fig. 3: Newspaper image decomposition: (a) original, (b) segmented image, (c) identified articles.

3 Article identification and recognition

The overwhelming majority of the researchers are addressing the document image understanding problems for scientific journals (for an overview see Jain et al. [1]). Their results are not directly applicable to newspapers, which 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 formulated as a set of rules. A similar approach is given by Niyogi ([3]). However, the set of rules given by Niyogi are inadequate to handle the large variety of page layout of our testbed collection.

After the segmentation stage, it is very important to automatically identify the articles on a page, since the primary unit of information in a newspaper is the article. An article consists of a headline and a collection of segments having a logical type among the following: sub-headline, title, text, picture, caption and references.

Our set of rules tries to accomplish three goals. The first goal is to distinguish, among segments of the same type, those that separate articles. For example horizontal lines can be underlines or lines that separate articles. The second goal is to label master title, title and text title segments as headlines and sub-headlines. Finally, the third goal is to group a number of segments having different types around a headline. Our rules exceed 40 in number and can be grouped as follows:

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 which exceeds a threshold length value.

b) Rules that distinguish underlines from lines separating articles. Short lines, lines under master titles, titles etc are considered as underlines. In contrast, lines below a text segment are considered as lines that separate articles.

c) Rules that locate white rectangular areas that separate articles. These rules consider, among other objects, white areas that exceed a threshold in height and in length.

d) Rules that identify segments which are in the border of an article. These rules locate special patterns which are used to separate neighboring articles.

e) Rules that find the headlines. These rules consider a group of neighboring master titles as a headline. In addition, titles and text titles below a horizontal line that separates articles are considered as headlines.

f) Rules that specify a sub-headline that possibly accompanies a headline. These rules classify titles and text tiles as sub-headlines according to their relative position to a headline.

g) Rules that group segments in order to form an article. These rules are applied last and take into consideration the results of previous rules in order to define a thin separator segment, which lay at the top of the article.

An article identification example is shown in Fig. 3c.

4 Experimental results

The proposed methods for newspaper page segmentation and article tracking were tested extensively on a large set of newspaper images. The test set contains 100 pages from the newspaper «TO VIMA» with publication dates from 1965 to 1974. It is characterized by variety in page layout and in noise contamination due to low paper quality. A Pentium II processor at 350 MHz was used and the average time for segmentation and article tracking was 9 sec. Image resolution was 6592x9890.

Experimental results are summarized in Table 1. In order to trace the correctness of the identified elements (segments or articles) we calculated a recall rate (number of correct elements found divided by the total number of elements we are looking for) and the precision rate (number of correct elements found divided by the total number of elements found). Recently, Summers ([6]) 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 in the segmentation phase. Also, results are given 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 and lines. Recall and precision rates exceed 96%.
- The article tracking methodology exhibits success rates over 75% in both recall and precision calculation. This does not seem impressive compared to segmentation results. There are two reasons which explain this behaviour. First, if a segment is incorrectly classified to an article, then two articles are incorrect, one which has an additional segment and one which 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 which is not followed faithfully by the typesetters.

		Recall (%)	**Precision (%)**	**Segments with correct article classification (%)**
Segmentation	**Title**	89,10	87,44	
	Text	97,48	96,07	
	Image	98,24	98,23	
	Line	98,28	97,35	
Article tracking		75,20	77,15	90,23

Table1: Experimental results

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 ([6]). In other applications where these measures are used, there is usually a trade-off between recall and precision. It is interesting to further explore this difference in behaviour.

5 Conclusions

This paper presents a set of integrated algorithms for the decomposition of digitized newspaper pages and a rule based approach to identify the articles of a decomposed page. Our experimental results indicate that our method is adequately efficient and presents a cost efficient alternative to the creation of digital collections compared to the manual or semi-automatic methods that are used today. However, there is always place for improvements. Regarding page decomposition, techniques for the block-based parameterization of our algorithms can be devised. Identification of special symbols and vignettes is also important. Regarding article tracking, due to the variety in page layouts additional sources of information are required in order to improve our results. A possible source is the textual content of the segments produced by page decomposition. To explore this option we are already conducting experiments with OCR modules and term extraction from text segments.

References

1. Jain, A. K., Yu, B.: Cash, G. L., Hatamiam, M.: Document Representation and its Application to Page Decomposition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No 3, 294-308 (1998)
2. 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, 163-174 (1995)
3. Niyogi, D.: A Knowledge-Based Approach to Deriving Logical Structure from Document Images. Ph.D Thesis, State University of New York, Buffalo (1994)
4. Papamarkos, N., Tzortzakis, J., Gatos, B.: Determination of Run-Length smoothing values for document segmentation. Third IEEE Int. Conf. On Electronics, Circuits and Systems, ICECS 96, 684-687 (1996)
5. Strouthopoulos, C., Papamarkos, N.: Text identification for document image analysis using a neural network. Image and Vision Computing, Vol. 16, Iss. 12-13, 879-896 (1998)
6. Summers, K.M.: Automatic Discovery of Logical Document Structure. Ph. D. Thesis, Cornell Univ. (1998)
7. 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)
8. Wang, D., Srihari, S. N.: Classification of newspaper image blocks using texture analysis. Computer Vision Graphics and Image Processing, Vol. 47, 327-352 (1989)