Information Extraction from Document Images Using White Space and Graphics Analysis

Gerd Maderlechner and Peter Suda

Siemens AG, Corporate Technology
Otto-Hahn-Ring 6, D-81730 München, Germany
email: gerd.maderlechner@mchp.siemens.de

Abstract. The goal of this work is the fast extraction of relevant information from document images. Examples of interesting information are the type of document (e.g. form, report, letter), the title of an article or the sender of a business letter, and a logo or figure on a page. The basic idea is to use non-textual cues from the document image before any OCR/ICR or word recognition is performed. The approach is based on a compact runlength representation of the binary image and allows a document type classification by white space analysis in a time comparable with the input of the compressed image. Graphics related information extraction needs approximately the same time.

1 Introduction

The conversion of arbitrary black on white paper documents into electronic documents is an urgent need for many organizations, companies and even private persons when they use modern network communication. Large conversion projects are performed with considerable effort in cost and time, to guarantee a high quality standard, i.e. no loss of relevant information. There is no completely automatic conversion into the target document format available. The compromise is either high cost using human controlled input or semi-automatic input using document imaging, i.e. the original scanned image is stored. In this paper we propose some methods how to speed up the access to non-coded information by extracting only relevant parts of the document to the human computer user.

The goal of this work is to speed up the extraction of relevant information from document images. By fast we mean a processing time which is comparable with the input time of the document image from the compressed file or from the scanner, which is less than one second for professional systems. Current OCR/ICR systems are much slower.

Examples of interesting information are the type of document (e.g. form, report, letter), the title of an article or the sender of a business letter, and a logo or figure on a page.

Applications are in the area of browsing in large document image data bases with extraction of titles, figures, tables or chapters.

2 White Space Analysis

The basic idea is to use non-textual cues from the document image before any OCR/ICR or word recognition is performed which is slow and error prone.
The non-textual cues are part of the publishing design rules for layout and typography. Their purpose is to improve the communication from the writer to the reader through visual characteristics like placement of text, figures, surrounding white space, or graphics items.
Document image analysis starts with binary images. This is sufficient as long as the typical document is printed in black ink on white paper. Binary document images are efficiently compressed and stored using the lossless standardized CCITT G3 or G4 coding scheme.
Connected component (CC) analysis is widely used as the first step because there is a close correspondence between the black printed characters and the connected components. Ideally each character is described by one CC, only few characters consist of two or three CC's (i, j, äöüÄÖÜ, special characters like ; : = " ! ? and %). This does not hold any more for real scanned documents containing merged or broken characters. On the other hand the merging resp. breaking of characters can be used by CC analysis for a rough quality estimation (details in a future paper).
Both the size of the compressed images and the number of CC's increase if the documents contain halftone (i.e. rasterized) images or noise. Halftone images can be detected and eliminated easily and fast (for one of the fastest solutions see [1]). In high-performance systems the separation of halftone regions is done by the scanner in real time before storage of the image in a G4 compressed file (see e.g. [4]).

There are several published approaches on layout segmentation and analysis which use white space analysis (e.g. [1], [3], [5]). The speed of [1] is comparable to our results whereas the approach in [3] is slower. In [5] the analysis starts from coded information after OCR which never can reach the performance of the proposed image based approach.
The difference of our approach to [1, 3, 4] is in the representation of the binary image and the application of different resolution levels. We use an efficient representation of run length data which allows fast and flexible operations on the document image. The main operations are connected component analysis, skew detection and correction, horizontal and vertical projections, run length statistics and histogram calculations.
The resolution levels useful for white space analysis are in the interval from 8 to 32 (linear reduction in each dimension). This iconic like representation of the document page allows a very fast localization of salient information blocks and, as a side effect, a rough document pre-classification (e.g. discriminate a journal page from a business letter).

White space analysis for extraction of salient regions of both text and graphics is based on the detection and evaluation of white rectangular horizontal and vertical

regions surrounding the information containing text and graphics blocks. According to well known layout rules the emphasis of a text or graphics block increases with the size of the surrounding white space.
White space is represented by horizontal or vertical oriented rectangles surrounding the information blocks (Figure 1). White space is localized by analysis of horizontal and vertical projection profiles (after skew correction), starting with the whole page, and continuing recursively on the horizontal and vertical zones. There are parameters for the minimal width and height of a white space rectangle. This allows a robust localization of minima (characteristic for white space in min-is-white representation) in the projection profiles in the presence of orthogonal black zones. There is also a threshold of a small number of black noise pixels allowed in white space.
The white space analysis starts with the localization of the left, right, top and bottom white page margins. There is a special handling for images with black borders resulting from scanners with black background. These black zones are excluded from the following ranking process for information extraction.

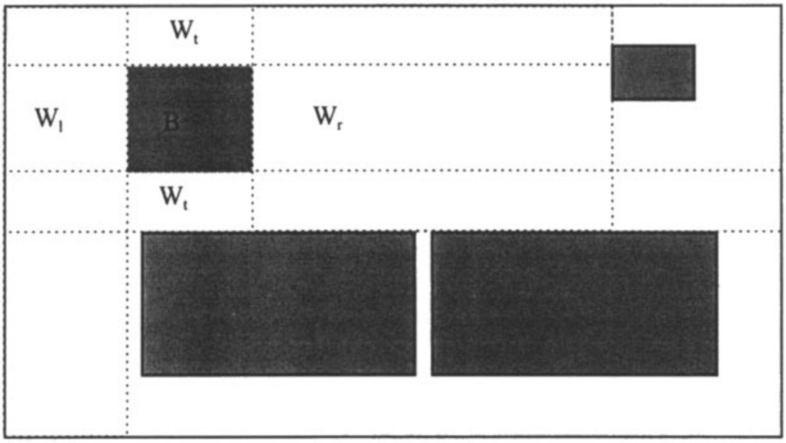


Figure 1: Definition of white space regions W_l, W_t, W_r, W_b *around a block B.*

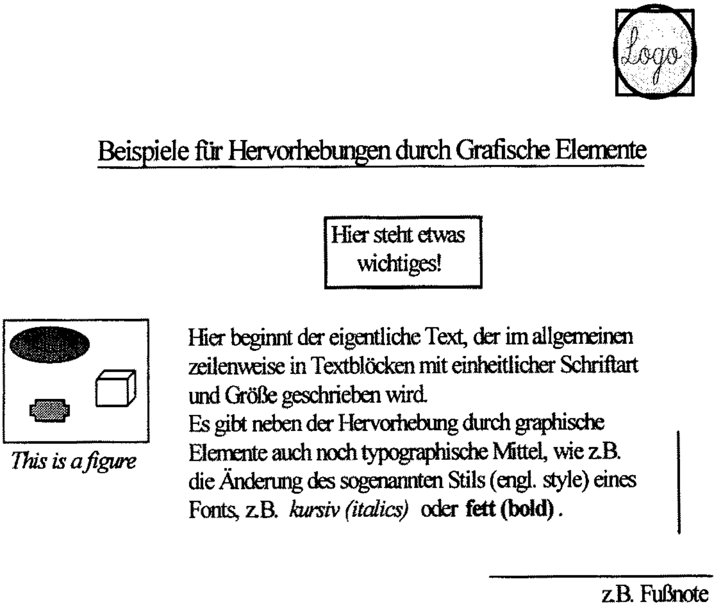
The ranking of the highlighting effect of a text or graphics block B by the surrounding white space W(B) is given by a *white space highlighting score S (B, W(B))*. The white space W surrounding B is given by the four maximal white rectangles W_l, W_t, W_r, W_b, touching B at the left, top, right and bottom side, respectively. The score S increases with the width w_i and height h_i of each W_i ($i = l, t, r, b$) and the common borders w (width) and h (height) with the block B according to the equation

$$S(B, W) = (\alpha_l * w_l * h_l + \alpha_t * w_t * h_t + \alpha_r * w_r * h_r + \alpha_b * w_b * h_b + \alpha_B * h * w) / (4 * A) \qquad (1)$$

where A is the area of the whole page, used for normalization. The α_i are weights for tuning the relative importance of left, top, right and bottom white zones with $\Sigma\alpha_i = 1$. α_B controls the influence of the block size. Blocks with the same score are ordered from top to bottom in importance because of the standard reading order.

3 Graphics Analysis

In the traditional approach using OCR systems the graphical information is ignored. In our approach we try to localize the salient graphics first in order to speed up the localization of the relevant information on the page for subsequent OCR processing. The graphics entities to be recognized are the following (see Figure 2): straight horizontal and vertical lines, rectangular frames, logos, and figures with their captions.



Beispiele für Hervorhebungen durch Grafische Elemente

Hier steht etwas wichtiges!

This is a figure

Hier beginnt der eigentliche Text, der im allgemeinen zeilenweise in Textblöcken mit einheitlicher Schriftart und Größe geschrieben wird.
Es gibt neben der Hervorhebung durch graphische Elemente auch noch typographische Mittel, wie z.B. die Änderung des sogenannten Stils (engl. style) eines Fonts, z.B. *kursiv (italics)* oder **fett (bold)**.

z.B. Fußnote

Figure 2: Examples of highlighting by surrounding white space and graphical entities.

The proposed approach to localize the salient graphics is based on a fast connected component (CC) analysis, which is performed in one pass during reading the compressed (CCITT group 3 or 4) image. In parallel to the CC's we accumulate histograms and calculate several features and neighborhood relations on the fly. In a second step the histograms are analyzed locally and globally. As a side effect we get the global skew of the document. This allows an accurate and fast determination of the horizontal and vertical lines by projection analysis. The rectangular frames are determined by aggregation of the remaining horizontal and vertical lines. Sometimes the lines are connected to characters. We have developed a new separation method which is faster than the solution described in [2].

The Logos will be segmented not only by the features of their CC's but also by analysis of their neighborhood. We observed that the salient graphics are arranged according to layout rules. One rule for localization is to surround the relevant object with sufficient white space around it (see chapter 2). Similar rules are used for the localization of figures, diagrams, and logos. The recognition of logos is described in [6]. Inside of text regions highlighting is done by typographic techniques (e.g. change of font styles like cursive and bold). Recognition approaches are under investigation.

4 Results and Conclusion

The proposed approach of white space and graphics analysis was applied to a set of about 430 binary document images (330 business letters and 100 journal pages) scanned at 300dpi. Some examples are shown in Figure 3.

Prior to the white space analysis the images were rotated into the right reading orientation and deskewed using our basis document analysis system IDA, which is described in [7].

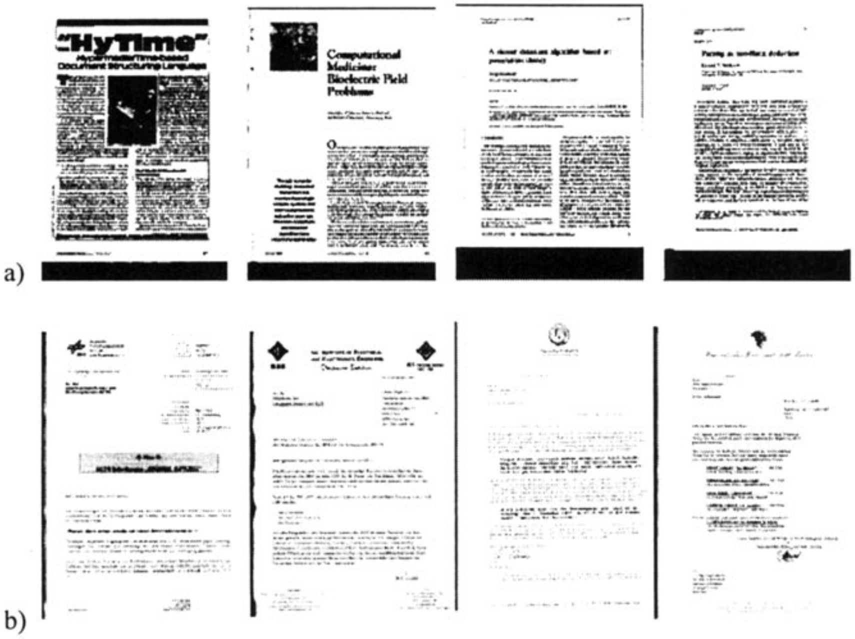


Figure 3: Examples of test documents: (a) Title pages of journals and (b) business letters.

We investigated the influence of the resolution on the extraction of white space information from factor 4 to 32. For fast discrimination between document types we found an optimum in speed and acceptable accuracy at about factor 16. For a human observer the reduced images (in gray scale display) allow the localization of salient regions. In Figure 4 we show the result of white space analysis to extract the region with the highest score independent of the type of the region (text or graphics).

We extended the white space recognition also to recognize and eliminate the black margins, caused by scanning US letter size with an A 4 scanner and black background.

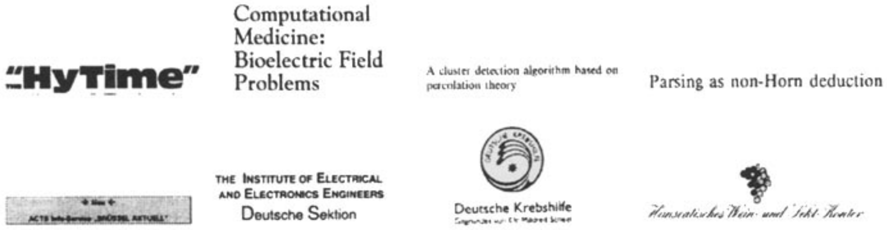

"HyTime"
Computational
Medicine:
Bioelectric Field
Problems
A cluster detection algorithm based on percolation theory
Parsing as non-Horn deduction
THE INSTITUTE OF ELECTRICAL
AND ELECTRONICS ENGINEERS
Deutsche Sektion
Deutsche Krebshilfe

Figure 4: Extracted region (text or graphics, normalized to width) from test documents with highest score from white space analysis, one for each test document in the same ordering Figure 3.

The analysis of graphics was limited to salient black objects like black copy margins (mentioned above), figures, logos, and horizontal and vertical lines. Figures and logos were identified in combination with the white space analysis, using the constraint that they are surrounded by white spaces areas (Figure 5). Here we did not discriminate between halftone image regions (upper three regions in Figure 5) and graphics regions.

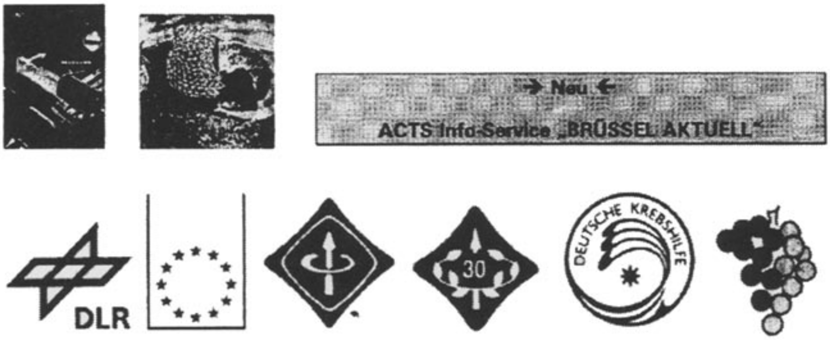


Figure 5: All extracted graphics regions in the documents from Figure 3 (Two journal pages contained no graphics, one letter page contained three, and one letter page two graphics regions)

This approach was also applied to business letters and journal pages. In business letters the following areas of interest could be localized very fast: address of recipient, footer area, and sender information containing a logo. In title pages of journals we could localize the whole title area, the title, the authors (with affiliation), the journal logo if present, and the page number.

The processing times always remain below one second on a 200 MHz Pentium PC, including skew correction and connected component analysis for a typical A4 sized page scanned at 300 dpi.

The proposed approach to use white space and graphics as visual cues to focus the OCR to relevant areas of interest is suitable to reduce the processing time. A quantitative and statistical evaluation is in progress.

5 References

[1] D. S. Bloomberg, *Textured Reductions for Document Image Analysis,* Proc. SPIE, Vol. 2660, 1996, pp. 160 - 174.

[2] Gerd Maderlechner, *Symbolic Subtraction of Fixed Formatted Graphics and Text from Filled in Forms*, Proc. IAPR Workshop on Machine Vision and Applications, Tokyo, Nov. 1990, pp. 457 - 459.

[3] M. Ozaki, *Logical Tagging of Document Images by White Space Pattern Matching,* in *Shape, Structure and Pattern Recognition* by D. Dori and A. Bruckstein (editors), World Scientific, Singapore, 1995, pp. 350 - 359.

[4] T. Pavlidis and J. Zhou, *Page Segmentation and Classification,* CVGIP: Graphical Models and Image Processing, Vol. 54, No. 6, 1992, pp. 484 - 496.

[5] D. Rus and K. Summers, *Using Non-Textual Cues for Electronic Document Browsing,* in *Digital Libraries: Current Issues* by N. R. Adam, K. Bhargava, and Y. Yesha (editors), Lecture Notes in Computer Science, Springer Verlag Berlin, New York 1995, pp. 129 - 162.

[6] P. Suda, C. Bridoux, B. Kämmerer, G. Maderlechner, *Logo and word matching using a general approach to signal registration*, Proc. ICDAR'97, pp. 61 - 65.

[7] G. Maderlechner , T. Brückner, and P. Suda, *Classification of documents by form and content,* Pattern Recognition Letters, Vol. 18, No. 11-13, 1997, pp. 1225 - 1231.