Annals of DAAAM for 2010 & Proceedings of the 21st International DAAAM Symposium, Volume 21, No. 1, ISSN 1726-9679
ISBN 978-3-901509-73-5, Editor B. Katalinic, Published by DAAAM International, Vienna, Austria, EU, 2010
Make Harmony Between Technology and Nature, and Your Mind will Fly Free as a Bird
Annals of DAAAM International

SYSTEMATIC PRINTING SPACE RECOGNITION

BOIANGIU, C[ostin] - A[nton]; PETRESCU, S[erban] B[arbu]; MOLDOVEANU, A[lin] D[ragos] B[ogdan]; ASAVEI, V[ictor] & BUCUR, I[on]

Abstract: *This paper aims to present an effective method for recognising the useful print space of high-quality digital images obtained by means of automatic scanning devices.The algorithm purports to become a valid technical solution for cropping both single-page images, and double-page ones – in the latter case the gutter is detected and the final result is made of two distinct images representing the left-hand-side page and the right-hand-side page.*
Key words: *page-splitting, gutter-detection, double page*

1. INTRODUCTION

Playing a major role in high volume document digitization (Yacoub et al., 2005), print space recognition is very important both from the qualitative point of view, as well as from the overall time-performance and costs of the digitization process perspective (Di Zenzo et al., 1996). As shown by the tests which have been performed, the technique shown in this paper is suitable for most types of input documents.

Further research will be aimed at improving the proposed algorithm by constraining the detection of fixed-dimensional cropping frames for a series of related scanned documents. Moreover, an extension of the algorithm which takes into consideration the particularities of high-quality color digital images of documents (as opposed to greyscale) will be studied.

2. ALGORITHM OUTLINE

Scanning books and documents flat against the glass of a scanner cannot be a viable solution for many document digitization projects (Breuel, 2003). Many documents are quite fragile and pressing them firmly in order to ensure that the entire page is flat on the glass of a scanner might lead to further, unacceptable damage. Furthermore, many books, particularly textbooks, have gutters (the inner margins next to the book's spine) very close to the binding. (Zhang & Tan, 2005). This makes it very difficult to scan the pages without unbinding the book (so the pages are separated and then can lie flat on the scanning glass).

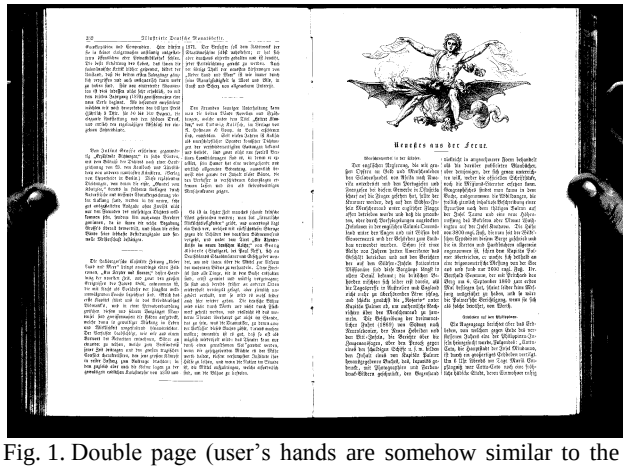

Fig. 1. Double page (user's hands are somehow similar to the clips used by automatic scanners – e.g. Kirtas)

As cutting off of the bindings of books and magazines was not an option for very old and uncommon books (Zheng et al., 2003), document-digitization companies had to turn to software driven machines and robots, developed to automatically scan books without the need of disbinding them (Kirtas, ATIZ, ScanRobot SR301, etc.). This type of machines allow for both the content and the state of a document to be preserved, by capturing high-quality digital images of each individual page. The proposed algorithm is trying to achieve print space recognition and cropping of digital images of documents, obtained using automatic scanning machines.

The first step of the algorithm is the removal of outer noise (such as scanning-machine's details or exterior margins outside the document in focus). This is performed by computing the sum of the grayscale values for each individual column/row, and then applying two thresholds as follows: all lines/columns having the sum of grayscale values lower than 60%/30% from the maximum sum of grayscale values for an individual line/column are filled with black (all pixels are turned to black).

After the first stage has been completed, a score is computed for each individual line/column by adding up the grayscale values of all pixels. These sums are altered by adding an extra value of 200 for each pixel with value greater than 200 in order to increase the score-gap between line/columns intersecting text and lines/column representing white spaces. This gap needs to be high in order to cope with brightness variations in input documents (resulting from the position of light sources in automatic scanning devices for example).

The left/rightmost margins of the document are investigated in order to determine whether the document is a single/double page one. For example, if the right margin is made a series of black columns (while the left margin is not), the document is classified as a single page one, situated on the left side. Similarly, images can be classified as single page documents sittuated on the right or double-page documents.

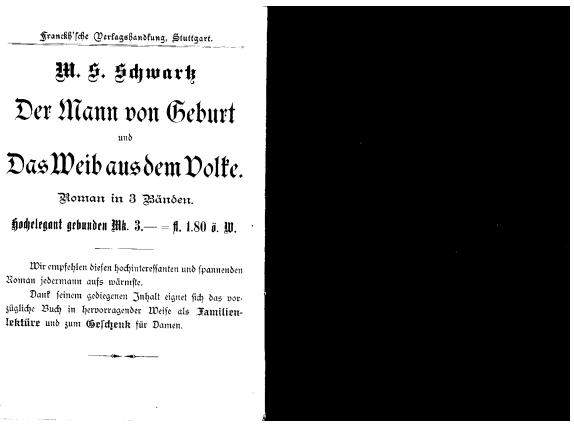

Fig. 2. Single page placed on the right

The upper/lower bounds of the cropping rectangle are determined by scanning the document from top/bottom and stopping whenever a line with a score higher than 30% of the maximum score of all lines is detected. In order to cope with a possible slight inclination of the document in the image, an extra correction of:

$$0.005 * (bottom\ bound - top\ bound) \quad (1)$$

is added/subtracted to/from the previously identified cropping.

If the input document contains a single page, the following method is used in order to determine the vertical cropping bounds. Suppose the page is situated on the right. Starting from the right margin, the image is scanned column by column until a column with a score higher than 90% from the maximum score of all columns is found – this is set as the right cropping bound. The scan than continues either:

- for further 10% of the image's width
- until a column with a very low score is encountered.

If condition of the second case is met, the scan is restarted from the current index (in this case there is a very high probability that part of the gutter of the book was encountered). Otherwise the right cropping bound remains unchanged. The same process is then repeated correspondingly in order to detect the left cropping bound, but without the second (gutter) condition. For images containing two pages, the bound detection is performed similarly.

For single page documents the final cropping is performed bounds detected so far. For the double page case, the gutter detection algorithm presented in the following section is performed on an intermediary version of the input image, obtained by cropping the input document using the cropping bounds detected so far.

3. DOUBLE-PAGE GUTTER DETECTION

If the scanned document represents a double-page, it has to be split in order to crop the two pages individually. In order to do this, the gutter of the book/magazine from between the two distinct pages must be detected and the cropping must be performed at the left/right of the gutter.

Gutter detection is performed using the histogram of the sum of grayscale values of pixels for each individual column. This histogram is first processed using a triangle filter in order to remove some of the local minimum/maximum which is the result of randomness (for example a column with a very large sum of values compared to its neighbours, within a compact paragraph). As it is trivial that the gutter is somewhere around the center of the image, we will only consider the part of the histogram 20% percent left to 20% right of the middle of the histogram. For this area we will compute the set of local minima with values lower than the mid-histogram value (the arithmetic mean between the maximum and the minimum values of the entire histogram); one of them is to be used in order to detect the gutter.

Filtering the set of local minima (from the part of the histogram corresponding to the middle of the image's width) is done using the following criteria: for each local minimum a pair of numbers (GI1, GI2) called the gutter-interval is computed as follows:

- GI1 represents the index in the histogram of the first pixel ***left*** of the considered minimum, which has a histogram value higher than the mid-histogram value (the arithmetic mean between the maximum and the minimum values of the entire histogram)
- GI2 is similar to GI1, but computed at ***right*** of the local

From the set of distinct pairs (GI1, GI2), we choose one as the gutter by selecting the pair with the largest gutter score:

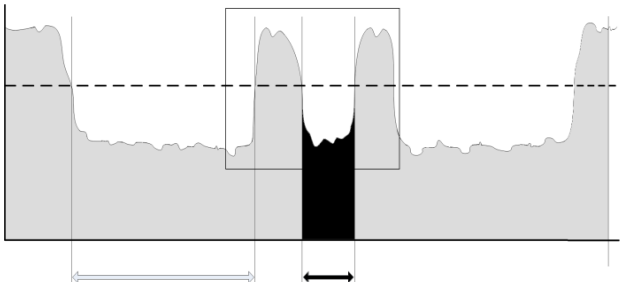

Fig. 3. Gutter detection – in black is the final gutter detected on the histogram (it can be noiced in the figure that in the middle section of the histogram there are four distinct local minim, three of which give the same interval, which represents the actual gutter)

$$GS = \frac{1}{avVal * iLength} \quad (2)$$

where:

- avVal is the average histogram value from GI1 to GI2;
- iLength is GI2 – GI1.

Finally, the two final distinct pages are used by cropping the document over the columns (0 : GI1) for the left page and (GI2 : imageWidth-1) for the right page.

4. CONCLUSION

In this paper we have shown an effective and practical method of cropping the print space from a digital image of a document. This algorithm is an essential pre-processing stage in the process of document digitization, and its success is critical for the content recognition and conversion methods that follow.

5. ACKNOWLEDGEMENTS

The research presented in this paper is supported by the national project "Excelenţă în cercetare prin programe postdoctorale în domenii prioritare ale societăţii bazate pe cunoaştere (EXCEL)", Project POSDRU/89/1.5/S/62557.

6. REFERENCES

Breuel, T.M. (2003). "An Algorithm for Finding Maximal Whitespace Rectangles at Arbitrary Orientations for Document Layout Analysis", *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Vol. 1, pp. 66, ISBN 0-7695-1960-1, Scotland, August 2003, Edinburgh

Di Zenzo, S; Cinque, L. & Levialdi, S. (1996). "Run-Based Algorithms for Binary Image Analysis and Processing", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 1, January 1996, pp. 83-89, ISSN: 0162-8828

Yacoub, S.;Saxena, V. & Sami, S. N. (2005). PerfectDoc: A Ground Truthing Environment for Complex Documents, *Proceedings of International Conference on Document Analysis and Recognition*, pp. 452-456, ISBN ISBN: 0-7695-2420-6, Seoul, August 2005

Zhang L. & Tan, C. L. (2005). Warped Image Restoration with Applications to Digital Libraries, *Proceedings of International Conference on Document Analysis and Recognition*, pp. 192-196, ISBN ISBN: 0-7695-2420-6, Seoul, August 2005

Zheng, Y; Li, H. & Doermann, D (2003). "A Model-based Line Detection Algorithm in Documents", *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Vol. 1, pp. 44-48, ISBN 0-7695-1960-1, Scotland, August 2003, Edinburgh