Unified layout analysis and text localization framework

Article *in* Journal of Electronic Imaging · January 2017

DOI: 10.1117/1.JEI.26.1.013009

2 authors, including:

Ergina Kavallieratou
University of the Aegean
70 PUBLICATIONS **1,124** CITATIONS

Complex layout analysis based on contour classification and morphological operations

Nikos Vasilopoulos Ergina Kavallieratou

Dept. Information and Communication Systems Engineering,
University of the Aegean, Samos 83200, Greece
{nvasilopoulos, kavallieratou}@aegean.gr

ABSTRACT

In this paper, a technique appropriate for document image layout analysis is presented. The technique is appropriate for colored and complex layouts of newspapers and journals. It is a hybrid technique that makes use of the contour classification method and also applies morphological operators. Detailed experiments on 2000 scanned images from newspapers gave an accuracy of more than 95% while the computational cost per page is less than a half second.

CCS Concepts

• Information systems→Information retrieval→Document representation→Document structure.

Keywords

Document images; Page Layout Analysis; Contour classification; Morphological operators

1. INTRODUCTION

Layout analysis is the process of analyzing document images in order to identify physical (e.g. text, pictures etc.) and/or logical (e.g. titles, paragraphs etc.) structures. The performance of layout analysis methods depends heavily on the page segmentation algorithm used. The page segmentation methods that have been reported in the literature can be categorized into foreground analysis, background analysis and hybrid ones.

Foreground analysis techniques use a bottom-up approach. They start from pixel level and merge regions together into larger components to form document structures (e.g. characters, then words, text lines, paragraphs and so on). Wong et al. classify connected components into text and non-text zones, after linking together neighboring black areas by performing a run-length smearing algorithm (RLSA) [1]. Fletcher et al. group together components into logical character strings using the Hough transform [2]. Gorman and Lawrence find the K-nearest neighbors for each connected component and use distance thresholds to form text blocks [3]. Simon and Pret also use a distance-metric between the components to construct the page structure [4], while Koo et al. group connected components as well into text-lines [5].

Background analysis techniques use regions of white pixels to split the page into blocks which are subsequently identified and further subdivided. Nagy et al. recursively split the document at the valleys along the horizontal and vertical projection profiles [6]. Kise et al. thin the white areas to form connected thin lines or chains and then find the loops enclosing printed areas [7]. Breuel uses tall whitespace rectangles as obstacles in order to detect text-lines [8].

Hybrid techniques analyze both foreground and background regions. Pavlidis and Zhou group background column gaps into column separators after horizontal smearing of foreground pixels [9]. Antonacopoulos and Ritchings also perform smearing first and then detect streams of white tiles whose sides encircle printed regions [10]. Chen et al. incorporate foreground and background information in order to filter whitespace rectangles progressively so that remaining rectangles form column separators [11].

2. THE PROPOSED SYSTEM

In this paper a layout-independent hybrid method, for complex (newspapers, magazines etc.) layout analysis is proposed. No prior knowledge of the document is required. Morphological operations are applied to both the foreground and the background, in order to connect neighbouring components and separate lines/columns. Contour is simultaneously used for the extraction and classification of images and text blocks.

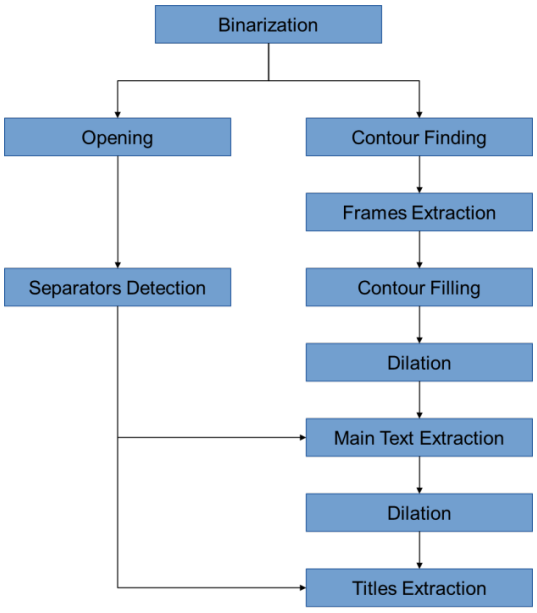

Figure 1. The proposed System.

SETN '16, May 18-20, 2016, Thessaloniki, Greece

DOI: http://dx.doi.org/10.1145/2903220.2903246

The proposed system is presented in Figure 1, while the detailed description of the tasks follows.

The method is applied on binary images. Therefore the image is first transformed to grayscale and then binarized, so that the background is white and the foreground (text, images etc.) pixels are black (Figure 2). The better the discrimination between the background and the foreground, better are the results. The background intensity range is used as the threshold value (maximum value). First, the grayscale histogram of the image is calculated using 64 bins and the bin with the highest (maximum) value is located. Then a binary image is created, where all the pixels of the peak value bin are white and the rest of the pixels are changed to black. This way, the method can be applied to pages with any background and foreground colors, since in the document images the majority of pixels belong to background.

2.1 Background processing

First, the background is processed in order to localize long white columns and rows. These straight white areas are classified as separators and can be used during the foreground processing phase to split overlapped components and improve the page segmentation results. The binary image is morphologically opened with two structuring elements: a horizontal line and a vertical line. The length of the first one is equal to the image width while the length of the second one is equal to the image height. A new image mask is created, containing all the column and row separators of the document (Figure 3).

Figure 2. The original (left) and the thresholded image (right).

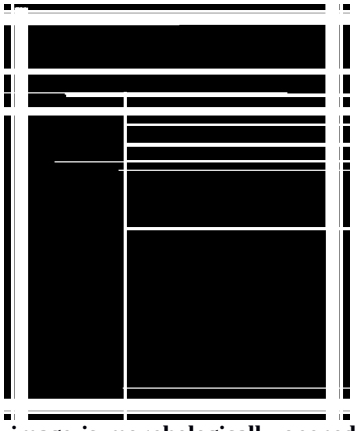

Figure 3. The image is morphologically opened with (image-width long) horizontal and (image-height long) vertical lines.

2.2 Foreground processing

Before proceeding further, the image is inverted (Figure 4). Then, a classical border-following algorithm [12] is applied and the external contours of all the connected elements are detected. The contour dimensions are used to determine the size of the main body of small text characters (x-height). That is the distance between the baseline and the mean-line of the lower-case letters (Figure 5). The minimum size between width and height of most lower-case letters is equal to the main body. Since most of the connected components in the document images of printed text are lower-case letters, the main body can be easily calculated as follows: a) all the contours are classified according to the minimum value between their bounding rectangle width and height and b) the value of the most numerous class is considered as the x-height value.

Figure 4. Foreground processing is applied on the inverted image.

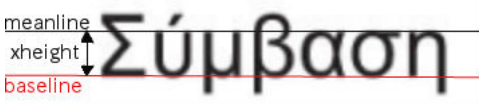

Figure 5. Definition of x-height

The extraction of frames and images follows. The frames and the images are large, compared to text, and are first extracted. A new image (layout) is created by filtering out the bigger contours (Figure 6). The big contours are classified as frames/images in the case that either their width or height is greater than ten times the main body. If their width is more than fifty times their height or vice versa, they are classified as separators and are added to the separator mask.

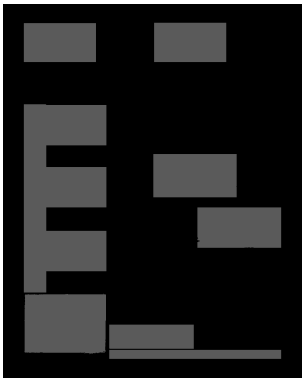

Figure 6. The dimensions of the external contours of the inverted image are calculated and large frames and/or images are extracted.

The remaining contours are filled before proceeding to the next step, in order to better distinguish between the small text and the large text regions. As described earlier, the contours have already been classified according to the minimum value between their bounding rectangle width and height. This is the value of the filling color. The bigger the text, the higher this value is. The result is a grayscale image, showing small text in dark gray and bigger text in lighter color (Figure 7).

Figure 7. The contours are filled with the minimum of their width-height value, so that smaller text is dark while larger text is lighter.

Last is the extraction of text blocks. The grayscale image containing the filled contours is dilated by a square structuring element (Figure 8). The size of the structuring element is equal to one third of the main body height. This way the small letters are connected and form text blocks, while the larger letters are not fully connected yet. The same border following algorithm [12] is applied again and the external contours of the connected elements are detected. This time the contours are classified as dark, containing small text, and lighter ones, containing large text. For each contour, the included image region is thresholded. Pixel values less than two times the main body height are zeroed. If the region mostly contains non-zero pixels, after thresholding, it's classified as large text. All other regions are extracted and added to the layout image.

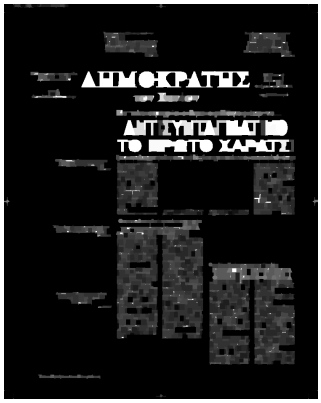

Figure 8. In order to connect small text, the image is dilated by a rectangle element of size one third of the main body height.

Sometimes, the text blocks may be very close to each other and dilation can merge regions from neighboring articles. For that reason, in order to split merged regions and improve segmentation results the logical AND is applied to the dilated image and the invert separator mask (see §2.1 background processing).

Figure 9. The bright regions are dilated and the larger text blocks are formed.

One more dilation is required, in order to cover the remaining space between the larger letters (Figure 9). The size of the square structuring element is now equal to half the main body height. This way large letters are also connected and form blocks. The dilated image is logically ANDed with the invert separator mask. Finally, the blocks are extracted as previous and added to the layout as well.

3. EXPERIMENTAL RESULTS

The algorithm has been coded in C# language. The morphological operations and the contour finding functions of the OpenCV library have been applied. About 2000 document images of high resolution newspaper scanned pages have been used for testing. The images are resized at 20% of their initial size before processing to reduce the computational cost. The results have shown that the method detects accurately more than 95% of the page components in less than half a second per page.

Figure 10. The final layout (left) shows the frames (dark gray), the small text (light gray) and the large text (white) blocks. The outlines of their bounding rectangles are drawn over the original image (right) in green (frames), blue (small) and red (large text).

4. CONCLUSION

A novel technique for page layout analysis of document images from newspaper and journals was presented. The technique can work with complex layouts and colored images. Our experimental results proved accuracy of 95% with computational cost less than 0.5 sec. For the future, the experimentation of the technique with more databases and comparison with other techniques is planned.

5. AKNOWLEDGMENTS

Nikos Vasilopoulos gratefully acknowledges financial support from the Hellenic Artificial Intelligence Society (EETN) for attending this conference.

6. REFERENCES

[1] Wong, Kwan Y., Richard G. Casey, and Friedrich M. Wahl. "Document analysis system." IBM journal of research and development 26.6 (1982): 647-656.

[2] Fletcher, Lloyd Alan, and Rangachar Kasturi. "A robust algorithm for text string separation from mixed text/graphics images." Pattern Analysis and Machine Intelligence, IEEE Transactions on 10.6 (1988): 910-918.

[3] O'Gorman, Lawrence. "The document spectrum for page layout analysis." Pattern Analysis and Machine Intelligence, IEEE Transactions on 15.11 (1993): 1162-1173.

[4] Simon, Anikó, and Jean Christophe Pret. "A fast algorithm for bottom-up document layout analysis." Pattern Analysis and Machine Intelligence, IEEE Transactions on 19.3 (1997): 273-277.

[5] Koo, Hyung Il, and Duck Hoon Kim. "Scene text detection via connected component clustering and nontext filtering." Image Processing, IEEE Transactions on 22.6 (2013): 2296-2305.

[6] Nagy, George, Sharad Seth, and Mahesh Viswanathan. "A prototype document image analysis system for technical journals." Computer 25.7 (1992): 10-22.

[7] Kise, Koichi, O. Yanagida, and Shinobu Takamatsu. "Page segmentation based on thinning of background." Pattern Recognition, 1996., Proceedings of the 13th International Conference on. Vol. 3. IEEE, 1996.

[8] Breuel, Thomas M. "Two geometric algorithms for layout analysis." Document analysis systems v. Springer Berlin Heidelberg, 2002. 188-199.

[9] T. Pavlidis, J. Zhou, Page segmentation and classification, CVGIP: Graphical Models and Image Processing, vol. 54, pp. 484-496, 1992.

[10] Antonacopoulos, A., and R. T. Ritchings. "Flexible page segmentation using the background." Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on. Vol. 2. IEEE, 1994.

[11] Chen, Kai, Fei Yin, and Cheng-Lin Liu. "Hybrid page segmentation with efficient whitespace rectangles extraction and grouping." Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013.

[12] Suzuki, Satoshi. "Topological structural analysis of digitized binary images by border following." *Computer Vision, Graphics, and Image Processing* 30.1 (1985): 32-46.