A new Method for Segmenting Newspaper Articles

B. Gatos [1], N. Gouraros [1], S. Mantzaris [1], S. Perantonis [2],
A. Tsigris [1], P. Tzavelis [1] and N. Vassilas [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”,
15310 Athens, Greece
sper@estia.iit.nrcps.ariadne-t.gr

Digital preservation of old newspapers contributes greatly to the historical register of a country's social, political and economical events. At the same time, newspaper preservation is an imperative necessity because of the fast paper deterioration and difficulty in tracing the overwhelming amount of information. 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 date. 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 from the printed material to an accessible digital archive (verification and quality control, digitization, cataloguing, search and retrieval, design and content presentation). The final digital documents form the foundation of our digital library.

Preservation and processing of this precious material can be achieved by focusing on a series of problems related to the digitization of the printed material, such as: image enhancement by noise removal, isolation of newspaper articles by document understanding techniques (segmentation - labeling). The successful tackling of these problems allows the subsequent efficient cataloguing by employing OCR, full text retrieval and information extraction techniques along with manual indexing.

In our paper we will present the results of our research associated with the stage of segmentation of the various regions - the image consists of - as well as the identification of text regions which have to be separated from other regions, i.e. figures, drawings or line regions. The main region segmentation techniques are based on two fundamental approaches: firstly, on the smearing and labeling of regions [1-2], and secondly on the image profiling in various directions [3-4]. Both techniques have not been successful in achieving newspaper segmentation because of the haphazard lay out of newspaper articles and their very close contact. Furthermore, the first approach results in great computational cost. Aiming at a solution of these particular problems accruing from the newspaper segmentation, we suggest a new technique based on horizontal and vertical image projections which provides a quick region segmentation as well as identification of text areas.

The proposed technique consists of three main stages: a) the calculation of horizontal and vertical smoothed profiling of the image, b) the indication of the various image regions using the local minima of the horizontal and vertical profiles and c) the indication of text regions by analyzing the FFT of the horizontal projections of segmented regions. More precisely, during the first stage, the image is projected horizontally and vertically using for each point the information obtained from the application of a mask whose dimensions depend on the approximate average of letter size. The result of this process is the projection of an image block in a massive rectangular region. Because newspaper images usually have clear vertical segments, we first process the vertical projections and secondly we get horizontal projections at the vertical extracted zones. During the second stage, the local minima of horizontal and vertical projections are indicated and they correspond to several segments into which the image can be divided horizontally and vertically. Horizontal neighboring segments are grouped together due to the existence of foreground pixels between the segments. During the last stage of our method, for each located area the horizontal projections are used. By analyzing their FFT we determine dominant frequencies which are used for identification of text areas as well as for the labeling of text segments according to their letter sizes. We identify text areas by defining a threshold at the dominant frequency amplitude. The letter size is provided by calculating the value of the dominant frequency which corresponds to the average distance between two successive lines.

The testbed is a collection of images from the newspaper "TO VIMA" published by Lambrakis Press S.A. from 1922 to 1970. The suggested method has already been applied with great success even in cases where text regions and graphs co-exist in an especially noisy environment.

References

1. Maier, M., Porinelli, R.: Separating Graphic Objects in Written Texts. Advances in Image Processing and Pattern Recognition (1986)
2. Kasturi, R., Bow, S., El-Masri, W., Shah, J., Gattiker, J., Mokate, U.: A System for Interpretetion of Line Drawings, Vol. 12. IEEE Trans. On Patt. Anal. And Mach. Intell., (1990) 978–991
3. Verikas, A., Bachauskene, M., Vilunas, S., Skaisgiris, D.: Adaptive Character Recognition System, Vol. 13. Pattern Recognition Letters, (1992) 207–212
4. Lettera, Maier, M., Paoli, C.: Character Recognition in Office Automation, Advances in Image Processing and Pattern Recognition, eds V. Cappellini and R. Marconi, Elsvier Science Publishers B. V., North-Holland, (1986) 191–197