Projet ANR- Blanc International II -

SIMI 2 - Science informatique et applications

GUWENSHIBIE

Programme Blanc International II 2012

Work Package n°3

Deliverable n°8

Layout Analysis

State of the art on layout analysis, comparison of existing methods from the literature.

Author: Thomas Konidaris

Delivery date: 01 January 2015

1 Introduction

There is a vast amount of information available in document collections both ancient/historical as well as contemporary. The reliable processing of such collections is vital in order to provide access to their invaluable content. A key process is the reliable layout analysis of these documents. The term document layout refers to the segmentation of the documents into regions of the same class, e.g. text regions, graphical elements, headings, etc. However, there are many cases where the segmentation of documents into homogeneous regions can become a very challenging task due to the complex layout of the document collections. Graphics, images and other non-textual information can be found blended with the textual content of the documents. This information needs to be separated successfully in order to be able to process the remaining textual information. In this report we will present representative works concerning the analysis of document layouts in both modern and historical document collections.

2 Related Work

In literature there are two main categories for document layout analysis namely, top-down and bottom-up, respectively. However, there are also methods that are featured based. In the following subsection we present a detailed description of these approaches with representative works on each of them.

2.1 Top-Down Approaches

The methods that fall into this category imply that there is a priori knowledge of the underlying layout of the document. This means that there exists an initial model that describes the layout. Using this model we can specify the regions of interest such as text, graphics, etc.

2.2 Bottom-Up Approaches

Contrary to the above methods, bottom-up approaches do not require any knowledge of the document layout. The processing begins on the pixel level and gradually the information that makes up the entire document is extracted. In these methods the input is the entire page.

2.3 Feature-based

In the category fall methods that use features in order to determine the different parts a document page consists of. Features can be either global, where the entire document page is under consideration or local, where in this case only regions of the document page are processed.

2.4 Hybrid

Hybrid methods for document layout analysis use a combination of the aforementioned categories in order to produce the desired results.

3 Related Work

A bottom-up approach is proposed in the work present in [15]. The method extends document layout analysis in order to be able to cope with mixed images, that is images containing text and various graphical elements. In particular, the authors extend the functionality of the RAST algorithm found in OCRopus software in order to be able to analyse mixed text documents. They also propose a text classification method using the Voronoi algorithm in order to perform OCR in that kind of documents. The work presented in [8] uses an EM-based method to estimate the shape of the different regions in the documents. The authors use a model that labels each pixel of the document with a number that corresponds to a region. At each region the algorithm tries to fit a set of Gaussian mixtures according to the logical distribution along the page and eventually derive the resulting shape of the regions. Figure 1 illustrates an example of the method.

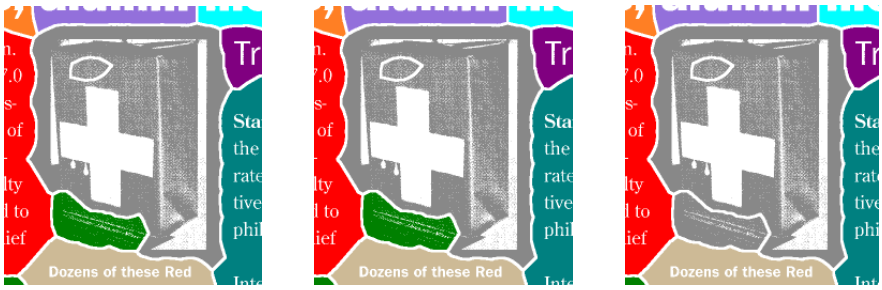
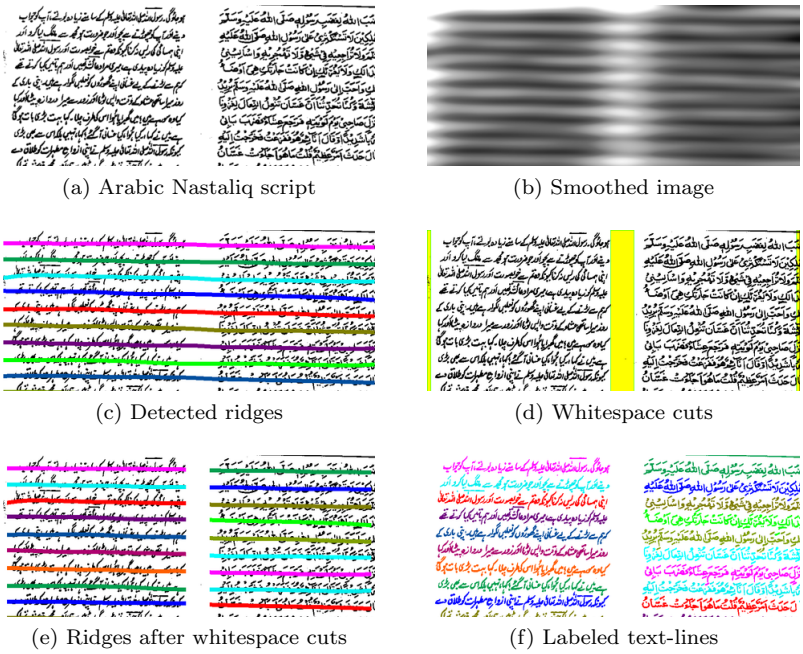


Figure 1: Region detection. Image taken from [8].

In [5] a method for text line extraction is presented. The work focuses on scanned documents of Arabic languages. The method is a combination of well established techniques such as ridge detection and whitespace cuts that enable the reliable segmentation between text and non-text regions. Text regions are further extracted into their corresponding text lines. Figure 2 illustrates the various steps of the method.



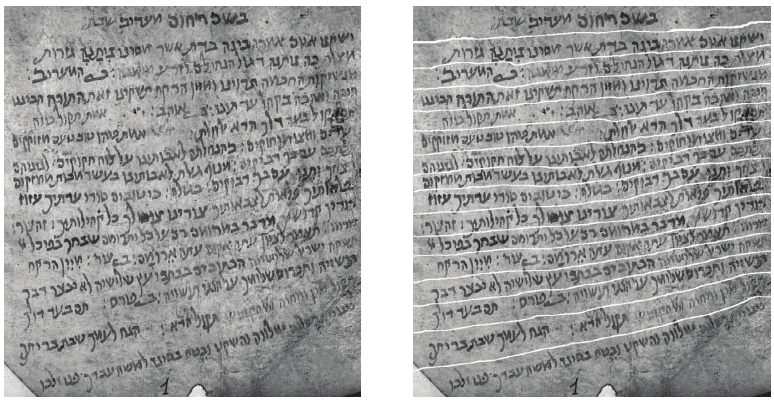
(a) Arabic Nastaliq script (b) Smoothed image

(c) Detected ridges (d) Whitespace cuts

(e) Ridges after whitespace cuts (f) Labeled text-lines

Figure 2: Text line extraction. Image taken from [5].

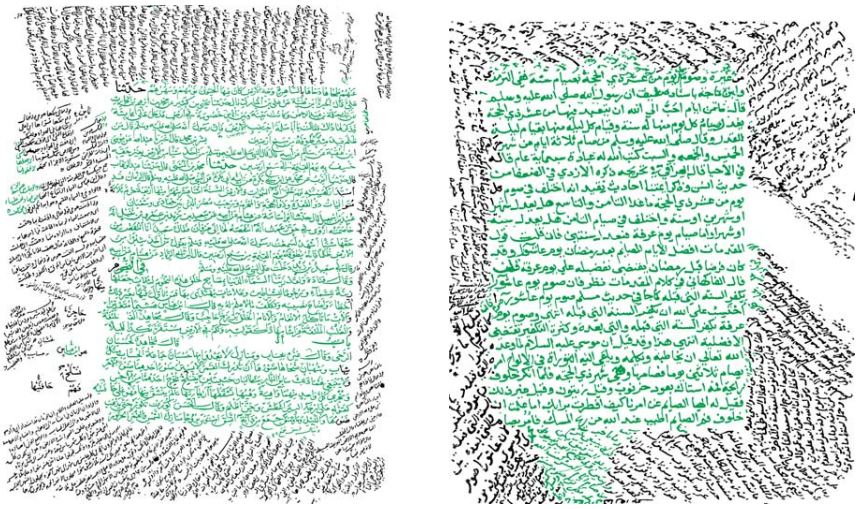
Although, traditional projection profiles has been successfully used for text line extraction, they fail when the documents are skewed. In [16], the authors introduce a method based on adaptive local projection profiles that are able to cope with various skew angles of the documents. These profiles are applied locally rather than globally like the traditional method and furthermore, can be adapted to the local skew of the document. In Figure 3 an example of text line detection is illustrated. Text lines and text regions for chinese documents is also the task in the work presented in [7]. They use a regrouping method that consists of three distinct steps. The processed documents contain a mixture of both vertical and horizontal text lines.



(a) (b)

Figure 3: Text line detection. Image taken from [16].

Bulacu et. al [6] propose a text line segmentation method based on ink transitions. The method is initialized in the center between two text lines following the contour of ascenders and descenders. If the path is blocked by overlapping ascenders/descenders, they are cut trough. The work presented in [14] falls into the feature-based category for document layout analysis. In this work a hybrid feature selection approach is adopted. An adapted greedy greedy forward selection and a genetic selection are used in a cascading way in order to process handwritten historical documents. LeBourgeois and Keileh [3] propose a system for document analysis that enables the retrieval of initials, illustrations and textual information from ancient document manuscripts. The method uses binary features concerning shape, geometry and color on a connected components level. The classification is performed using k-NN. In [4] a method is presented that segments text that appears in page margins. This approach is feature-based and it is performed on connected components level. For the classification of the text into the various classes a multilayer perception classifier is used. Figure 4 illustrates the results of the method on a document image.

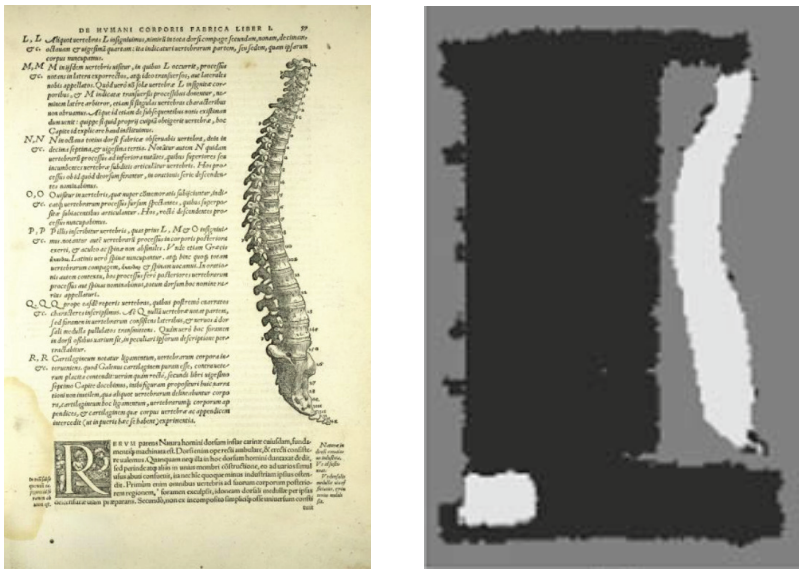


(a) (b)

Figure 4: Segmentation of marginal text. Image taken from [4].

A feature based approach is proposed in [2]. In this work SIFT local features are extracted from the document images. The method incorporates a learning system based on multiple classifiers trained through AdaBoost, in order to distinguish between the features that belong to text and to those that belong to images, drawings, photographs, etc. In [11]. Figure 5 illustrates an example of the performance of the algorithm. SIFT features are also used in the work in [9] where part-based identification of layout entities is employed.

The separation between text, decorations and images in handwritten manuscripts is the task presented in [10]. The proposed method uses circular statistics as a first step in order to extract text. Visual descriptors are then computed over the pictorial regions of the page. The separation between semantic and decorative elements is done through the use of color histograms and a novel texture feature called Gradient Spacial Dependency Matrix. Figure 6 illustrates



(a) (b)

Figure 5: Resulting image. Image taken from [11].

the results of an example manuscript page. There are also several survey papers that evaluate different document layout analysis methods. An older paper presented by Nagy [12] evaluates 99 papers, while Antonacopoulos et. al [1] evaluates several methods on scanned historical documents. Evaluation of different classifiers is presented in [13]. In particular Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP) and Gaussian Mixture Models are investigated on their performance on detecting physical structure of historical documents.

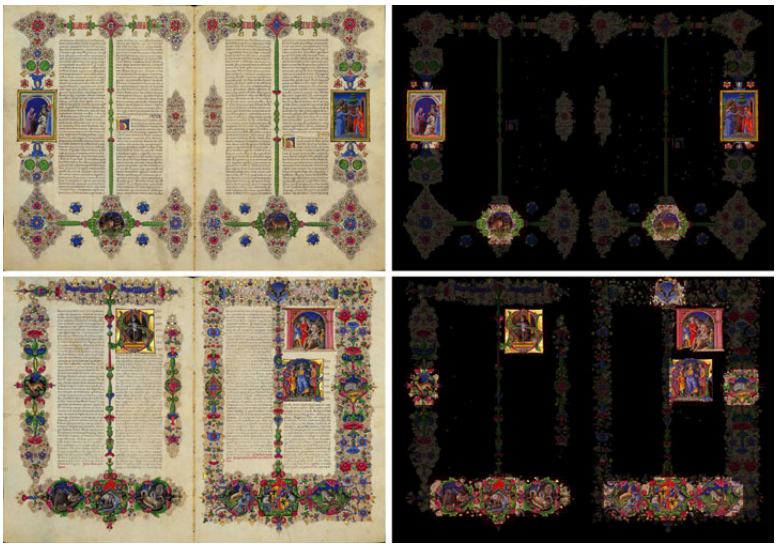


Figure 6: Resulting image taken from [10].

4 Conclusions

Document layout analysis is a very important process in document image analysis. This task can become very challenging, especially when the underlying documents are historical or when several different entities are included in the document images such as images, graphical elements, decorations, etc. Furthermore, methods that demonstrate satisfactory performance on contemporary documents is not uncommon to fail when processing the aforementioned documents. Research on the field has provided many useful tools and solutions for the robust confrontation of the problem.

References

[1] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher. Historical document layout analysis competition. In *2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Beijing, China, September 18-21, 2011*, pages 1516–1520, 2011.

[2] S. Baluja and M. Covell. Finding images and line-drawings in document-scanning systems. In *10th International Conference on Document Analysis and Recognition, ICDAR 2009, Barcelona, Spain, 26-29 July 2009*, pages 1096–1100, 2009.

[3] F. L. Bourgeois and H. Kaileh. Automatic metadata retrieval from ancient manuscripts. In *Document Analysis Systems VI, 6th International Workshop, DAS 2004, Florence, Italy, September 8-10, 2004, Proceedings*, pages 75–89, 2004.

[4] S. S. Bukhari, T. M. Breuel, A. Asi, and J. El-Sana. Layout analysis for arabic historical document images using machine learning. In *ICFHR*, pages 639–644, 2012.

[5] S. S. Bukhari, F. Shafait, and T. M. Breuel. High performance layout analysis of arabic and urdu document images. In *2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Beijing, China, September 18-21, 2011*, pages 1275–1279, 2011.

[6] M. Bulacu, R. van Koert, L. Schomaker, and T. van der Zant. Layout analysis of handwritten historical documents for searching the archive of the cabinet of the dutch queen. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September, Curitiba, Paraná, Brazil*, pages 357–361, 2007.

[7] F. Chang, S. Chu, and C. Chen. Chinese document layout analysis using an adaptive regrouping strategy. *Pattern Recognition*, 38(2):261–271, 2005.

[8] F. Cruz and O. R. Terrades. Em-based layout analysis method for structured documents. In *22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014*, pages 315–320, 2014.

[9] A. Garz, R. Sablatnig, and M. Diem. Layout analysis for historical manuscripts using sift features. In *2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Beijing, China, September 18-21, 2011*, pages 508–512, 2011.

[10] C. Grana, D. Borghesani, and R. Cucchiara. Automatic segmentation of digitalized historical manuscripts. *Multimedia Tools Appl.*, 55(3):483–506, 2011.

[11] N. Journet, R. Mullot, J.-Y. Ramel, and V. Eglin. Document image characterization using a multiresolution analysis of the texture: Application to old documents. 11(1):9–18, 2008.

[12] G. Nagy. Twenty years of document image analysis in PAMI. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):38–62, 2000.

[13] H. Wei, M. Baechler, F. Slimane, and R. Ingold. Evaluation of svm, MLP and GMM classifiers for layout analysis of historical documents. In *2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, August 25-28, 2013*, pages 1220–1224, 2013.

[14] H. Wei, K. Chen, R. Ingold, and M. Liwicki. Hybrid feature selection for historical document layout analysis. In *14th International Conference on Frontiers in Handwriting Recognition, ICFHR 2014, Crete, Greece, September 1-4, 2014*, pages 87–92, 2014.

[15] A. Winder, T. L. Andersen, and E. H. B. Smith. Extending page segmentation algorithms for mixed-layout document processing. In *2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Beijing, China, September 18-21, 2011*, pages 1245–1249, 2011.

[16] I. B. Yosef, N. Hagbi, K. Kedem, and I. Dinstein. Line segmentation for degraded handwritten historical documents. In *10th International Conference on Document Analysis and Recognition, ICDAR 2009, Barcelona, Spain, 26-29 July 2009*, pages 1161–1165, 2009.