Interactive layout analysis, content extraction and transcription of historical printed books using Agora and Retro

J.Y. Ramel, N. Sidère

[*] *Lab. d'Informatique, Ecole Polytechnique de l'Université de Tours,*
64, avenue Jean Portalis 37200 Tours- France
Tel : +33.2.47.36.14.26   Fax : +33.2.47.36.14.22
ramel@univ-tours.fr

High level analyses of document images are mainly based on the output of a page segmentation process. For example, the extracted text regions can be the input to an OCR system to retrieve the ASCII characters printed on the pages. The spatial relationships between segmented blocks along with other features can be used in logical page organization analysis to group the extracted components appropriately and recover the correct reading order. Many techniques for page segmentation have been proposed in the literature but most of them are based on the assumption that an input document image consists of a set of rectangular blocks. Furthermore, the classification step is generally domain specific and uses static rules to automatically determine, for each block, the coherent label selected from a predefined list (title, paragraph, graphic, table,...). These limitations appear too restrictive with respect to some noisy and distorted documents and new approaches need to be developed.

In this context, we present a work achieved in collaboration with the "Centre d'Etude Superieur de la Renaissance" of Tours (CESR / http://www.cesr.univ-tours.fr). The CESR is a training and research centre which receives students and researchers who wish to work on various domains of the Renaissance using a rich library of historical books. The CESR wants to create a Humanistic Virtual Library; however, until now, only bitmap versions of several books that have been scanned or photographed are accessible. The initial objective of the CESR was to obtain an ASCII version of the text contained in the pages of these historical books. The centre first tried to use the commercial OCR software to index their books but they quickly realized that, applied to historical documents, this procedure would had been vowed to failure. So, the CESR asked our Pattern Recognition and Image

Analysis research team to help them to define a new system adapted to their needs. They have appreciated our efforts as our collaboration will lead to a system able to bring a better description and indexation of the content of their books and would also make the search and the reading of these precious historical books easier.

The poster will first describe the new hybrid method we have developed for the extraction of layout information and of specific elements like graphical parts or ornaments based on the construction of two representations of the contents of the images. A mapping of the shapes and a mapping of the background are computed. By exploiting this information, our algorithm produces and sends back a list of blocks constituting a first segmentation result. Then, this initial representation of the image is used during a more sophisticated analysis. Having an aim of genericity, the architecture of the system that we carried out authorizes an interactive installation of scenarios for analysis of the image contents. Scenarios work on the initial representation provided by the first step of the segmentation. According to its needs (localization of the ornamental letters, the notes at margins, titles,…) and using user-friendly interfaces, the user (not expert in image processing) builds scenarios allowing to label, to merge, to remove the blocks contained in the intermediate representation. One can thus locates the desired entities without taking care of the other areas of the image. The elaborated scenarios can then be stored, modified and applied to various sets of images during batch processing. The results obtained with this method are very interesting; the adjustment of the necessary parameters is straightfoward and not sensitive to variations. The originality of our approach lies in the opportunity which we offer to the users to be able to build, in an interactive way, scenarios of incremental analysis. We propose to call this new method "user-driven analysis" in opposition to data-driven or model-driven methods. The goal is, on the basis of the initial segmentation, to be able to make the representation of the images evolve in a progressive way to lead to the finest possible characterization of its contents according to the user objectives and to the type of images to be analyzed. The CESR has processed several complete books using AGORA prototype and their own scenarios of block classification. Thus, the CESR has increased the number of books offered to the users in its Virtual Library (see http://www.bvh.univ-tours.fr). Even if the system produced some errors, the processing and time saved as compared to manual processing is considerable (for exemple, the manual indexation of the page layout of an historical book of 300 pages last approximately two days instead of only two hours when using Agora), this providing to the specialists of historical books, a useful tool which they had never imagine (see Figure 1).

Concerning text transcription, the originality of our work relies upon the analysis and exploitation of pattern redundancy in documents to allow efficient and quick transcription of books as well as identification of typographic materials. This pattern redundancy is mainly obtained via clustering methods. Like this, the traditional OCR problem could be reformulated into a text transcription one. A text, be it ancient or not, is made up of sequences of symbols. The scanning process produces pictures where symbols are represented as thumbnails of patterns (a pattern could be a single character, a part of a character or a set of joined characters), which may be more or less distinct. Without prior knowledge about the meaning of these symbols, the application of a clustering approach assigns thumbnails of a similar shape to the same cluster. As an example, one cluster containing thumbnails of the lowercase letter “a,” another one the uppercase letter “A,” yet another one the letter “b” in a specific font, and so on. Once the clustering is done, a user could assign a label to each cluster using a other Graphics User Interface (software called RETRO). These labels are then automatically assigned to each pattern, thus achieving the text transcription of the whole book. In this way, if 90% of patterns are detected as redundant, only one character in ten will be labeled by the user in order to transcribe the book. This part of the work is still in progress and is corresponding to a Google Digital Award obtained by our team in December 2010.

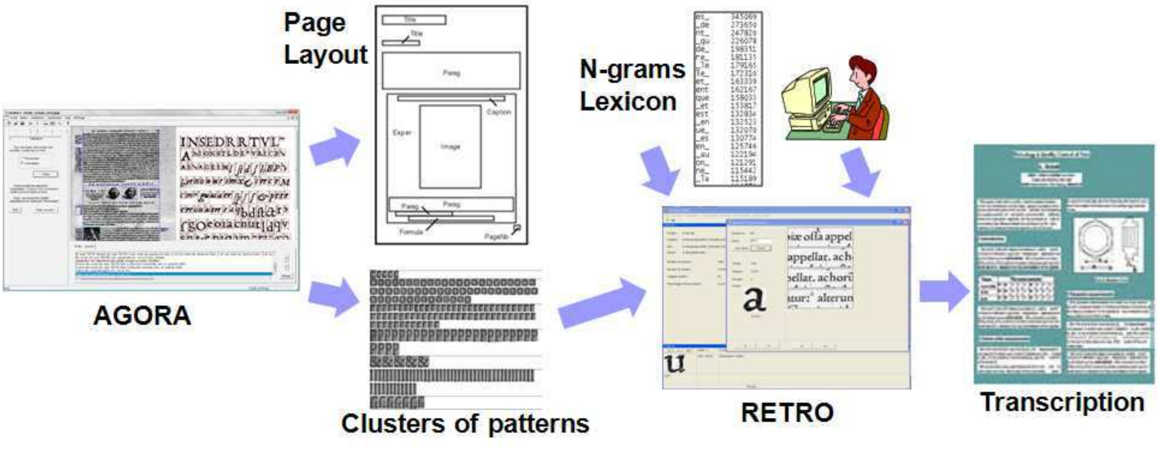

Figure 1 : A view of the proposed processing framework with Agora and Retro