PaRADIIT Project: Main Concepts and Outcomes

Frédéric Rayar, Pascal Bourquin, Jean-Yves Ramel, Rémi Jimenes, Toshinori Uetani, Sandrine Breuil, Marie-Luce Demonet

PaRADIIT Project: Main Concepts and Outcomes

Frédéric Rayar*, Pascal Bourquin*, Jean-Yves Ramel*,
Rémi Jimenes**, Toshinori Uetani**, Sandrine Breuil**, Marie-Luce Demonet**
*Laboratoire Informatique - EA 6300, **CESR –BVH UMR 7323
Université François-Rabelais de Tours
frederic.rayar@univ-tours.fr

***Abstract*—This paper presents the main concepts of the PaRADIIT project, dedicated to indexing and transcribing historical books, and its most relevant outcomes.**

Keywords—Historical documents; Digital Humanities; Layout Analysis; Pattern Redundancy; Clustering; AGORA; RETRO

I. INTRODUCTION

PaRADIIT is one of the research projects the Center for Advanced Studies in the Renaissance (CESR, http://cesr.univ-tours.fr/) and the Computer Science Laboratory of Tours (LI) are carrying on since more than 10 years, in order to build up a digital library of primordial documents of the Renaissance period: the "Virtual Humanist Libraries". To make accessible these ancient books online both in facsimile and text mode, the LI, works in close collaboration with the CESR, and develops image processing tools which participates in a full processing chain, including *(i)* layout analysis, *(ii)* text/graphics separation, and *(iii)* text transcription.. PaRADIIT focuses on these last 3 steps of the processing chain.

Indeed, standard layout analysis and OCR techniques do not handle successfully old or "noisy" documents due to their high levels of degradation. Our research project studies alternative techniques to traditional OCR in order to provide indexation of images and text transcription of the ancient imprints.

The originality of the work relies upon the analysis and exploitation of *pattern redundancy* in image documents to allow efficient indexing and transcription of books as well as identification of typographic materials. This pattern redundancy is mainly obtained via *clustering* methods on patterns extracted during the layout analysis step.

The project has been mainly funded by two Google Awards in Digital Humanities. The results of this project are available on https://sites.google.com/site/PaRADIITproject .

II. PATTERN REDUNDANCY

Clustering is the task of dividing a set of objects into subsets (called *clusters*) so that objects in a same cluster are highly *similar* to each other. Such clustering algorithms may be applied for content analysis and recognition, to reformulate the traditional indexing problem into a cluster labeling one.

A document, be it ancient or not, is made up of sequences of symbols that may appear several times in the document. We aim at leveraging this text redundancy at image level. 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 (see Figure 1). Without prior knowledge about the meaning of these symbols, application of a clustering assigns thumbnails with a similar shape to the same cluster.

Once the clustering is done, a user (or a computer) could assign a label to each cluster using a Graphics User Interface (GUI). These labels are then automatically propagated to each clustered pattern, thus achieving the indexing and transcription of the whole book. In this way, if 90% of patterns are detected as redundant, *i.e.* only one character in ten will be labeled by the user in order to transcribe the book.

The application of clustering techniques to the processing of characters in old books was initially introduced in the framework of the DEBORA project[1] to compress image dataset for storage purposes. One can also note that this approach today constitutes one of the main axes of the IMPACT project (http://www.impact-project.eu/).

III. PaRADIIT OUTCOMES

Some of the main outcomes of the PaRADIIT project are presented below.

A. *AGORA: an open source software for Layout analysis and interactive content extraction*

During the PaRADIIT project, the software AGORA has been developed. It simultaneously allows page layout analysis, text/graphics separation and specific pattern extraction in an interactive manner. This software offers to users the possibility to build interactive scenarios of incremental analysis. We call this new method "*user-driven analysis*" as opposed to data-driven or model-driven methods. AGORA can be used to analyze the page layout of historical books or to easily extract and index specific elements of content, such as initial letters, portraits, or notes in margins.

The CESR has processed several complete books using AGORA with customized scenarios of block classification. Thus, the CESR has quickly increased the quantity of valuable data offered to users, such as researchers and scholars, in its Virtual Library.

[1] F. LeBourgeois, H. Empotz, Document Analysis in Gray level and typography extraction using Character Pattern redundancies, ICDAR, Bengalore India, p177-180. 1999.

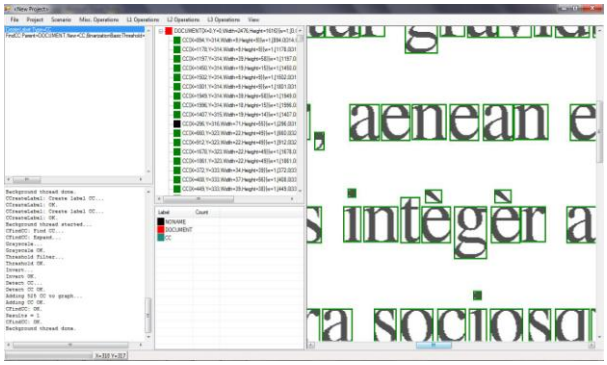

Figure 1: GUI of AGORA

B. *RETRO: an open source software for content clustering, recognition and description*

The patterns extracted with AGORA can then be processed using a second software, RETRO, to process the clustering, to visualize the current results, and to do the effective transcription.

Thus, RETRO allows users to transcribe the clusters with little effort, using an interactive labeling approach of frequent patterns. Figure 2. shows the transcription interface.

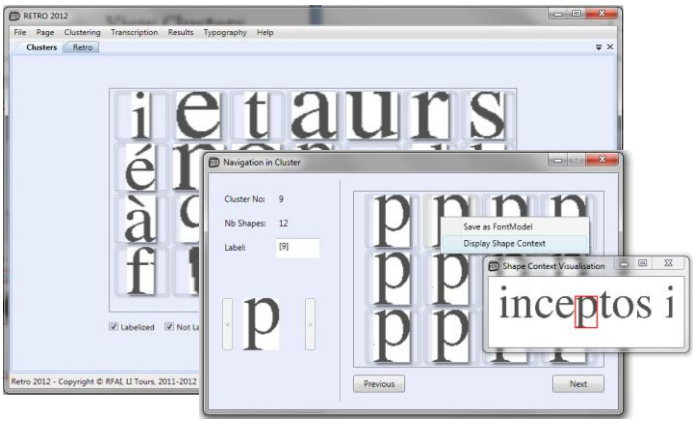

Figure 2: Interactive transcription with RETRO;

Moreover, this transcription method allows us to easily deal with special characters which appear frequently in old books (see Figure 3). For example, most of the European fonts before the 19th century use the ligature "ct" (concatenation of a "c" and a "t", see Figure 1, last item). As this double character cannot easily be split into two characters by standard OCRs, they are often improperly identified. Thanks to the proposed method, such ligature-based patterns can be transcribed into character couples without any modification of the recognition process.

C. *Pattern redundancy for Font analysis and for improving the OCR learning step*

It is also possible to use the clustering approach to extract and create new font packages from specific printing material (*e.g.* rare books printed with particular plug sets). These new font packages could be incorporated during the training step of Optical Fonts Recognition (OFR) methods, in order to improve the recognition results of OCRs on rare or specific books. This work is done in collaboration with specialists of typography of the Renaissance period (CESR). Such information could be added in a database, or encoded in a XML-TEI file, and used by researchers working on linguistic or literary field.

Consequently, while using AGORA and RETRO, it also becomes possible to construct new learning sets or new fonts of characters which are directly extracted from the clusters of characters coming from specific books.

Figure 3: Font analysis and model creation with RETRO

D. *Different GUI for the exploitation of the results*

Proof of concept applications on user-friendly interfaces have also been developed during the project to promote it and to make outcomes of our projects available on digital libraries to researchers or more general users. An online library and a Microsoft PixelSense platform (Figure 3) are currently available. It is possible to read and navigate into the digitized books distributed over remote servers, to search for specific contents at different granularities, to annotate documents or to export element of contents previously extracted and described using Agora and Retro.

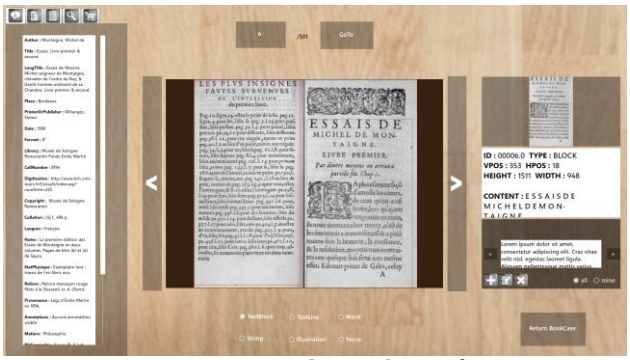

Figure 4: Multi-touch interface

ACKNOWLEDGEMENT

The PixelSense application has been mostly developed by Sébastien Guillon, former student of Polytech'Tours.