"PlaIR" : A System to Provide Full Access to Digitized Newspaper Archives

Thomas Palfray, Stéphane Nicolas, Thierry Paquet, Pierrick Tranouez; LITIS Laboratory; Rouen, France

Abstract

This paper presents a platform dedicated to the analysis and the online consultation of historical newspaper archives. This platform has been designed to provide a user experience as intuitive as possible by using mature open source tools. All the features are implemented thanks to the Spring framework. To meet this goal, we created a system to display tiled high-resolution images operating without a plug-in but based on an open source solution called IIPImage. The platform also allows for full-text searches thanks to the Java search library Apache Lucene and displays the results in the form of newspaper articles. In addition, we established collaborative features to provide the users with the ability to correct the content automatically generated by our document processing workflow and accessed through the browsing platform. The system is able to store all the corrections of the users, by using the couple Hibernate/MySQL. The aim is to enable continuous improvement of both the content quality and the search accuracy, by exploiting the ability of the users to recognize significant errors, in order to enhance the digital objects representing the newspaper issues.

The proposed system is designed to generate metadata describing the physical layout, but also the logical structure of newspaper documents. Our article segmentation analyses a newspaper issue and recognizes articles, even if they straddle more than one page or if they spread in a complex structure. The workflow can also consider as input data, the results of optical character recognition (OCR) engines in order to provide a textual indexation of the segmented articles.

By using this system, we want to create a true and representative digital object using standard formats (i.e. METS / ALTO) and containing the logical description of the content, making easier reading and understanding by the users.

Introduction

Over the last twenty years, libraries and archives around the world have developed many digitization programs for preserving and making more accessible historical records.

Newspaper archives are old documents with particularly rich historical information, but they require special processing to help finding information because of their complex layout and its evolution throughout the ages. We face particular challenges because of the nature of these documents including the poor scanning results of paper material that was originally of low print quality or has deteriorated through time. Furthermore, the content and the layout of the newspaper issues could fluctuate a lot depending of the time period, as we can see on the examples of the newspaper "Journal de Rouen" presented on figure 1.

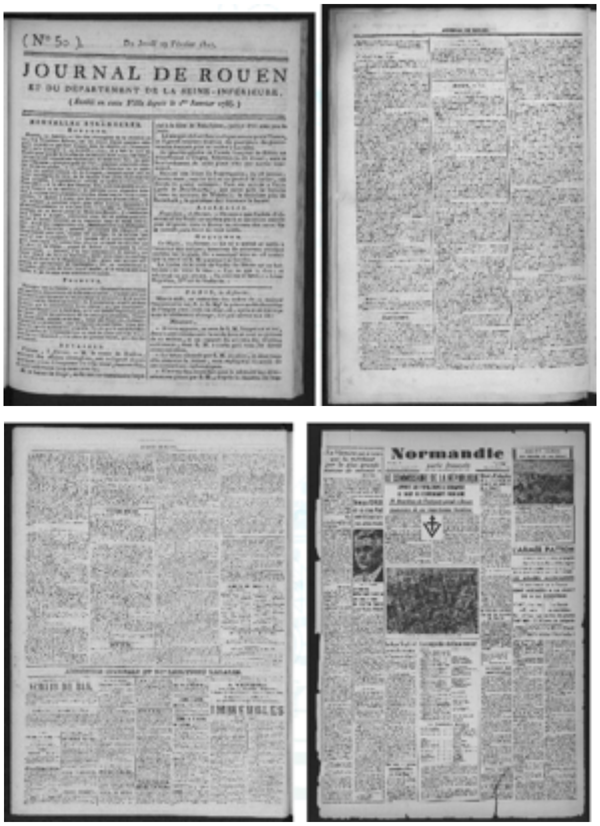


***Figure 1**. examples of various newspaper layout with different complexities*

Searching for and accessing particular information in such large corpora of complex documents is not easy for the users of digital archives. Most of the libraries and archives propose only an online access to the images of their digitized collections, or in the best case an access through PDF files offering some visualization facilities, but often only in image mode. Furthermore, due to time transfer and bandwidth limitations inherent to online access, the resolution of the images is sometimes too poor to allow the users to read comfortably the documents. It is important for digital archives and libraries to be much more than simple repositories of digitized document images, and thus to propose and to allow a better, simpler and more accurate access to the information contained in these documents. This is why we propose through the system we present in this paper, a solution to access the information at different levels, starting from the issue and down to the article. For that we exploit the results of automatic processing such as document image segmentation and optical character recognition (OCR), combined with manual interventions such as collaborative correction and tagging of the content. The aim is to provide in a convenient form, the precise information the users are looking for. This system is developed in the context of a project entitled "PlaIR" which aims at providing tools for electronic and digitized document indexing.

Our system consists of two main components: a workflow to process the digitized sources and extract the metadata needed to index the content, and an online consultation platform exploiting the produced metadata to facilitate the access to the digitized sources and their content.

We will now describe in further depth these two components, and conclude with the improvements planned for them.

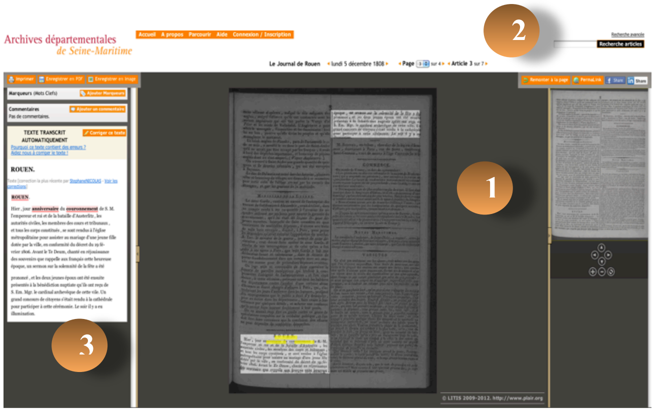
Online consultation platform

The online consultation platform is the front-end of our system. This visualization component relies on light web technologies, and thus its use does not imply any plugin installation or any system parameter setting. The user has simply to use his favorite web browser to access the documents and search the information they contain.

For the development of this consultation platform we used, integrated and adapted some open source solutions, and in particular, with their agreement, the CSS Stylesheets used by the National Library of Australia for their newspaper consultation website.

The interface of our online consultation platform is divided mainly into three functionalities as we can see on Figure 2:

1. the visualization of the digitized sources, with enlightened additional information (current selected article, search keywords, ...)
2. the access to text search functionalities
3. the view of the OCR content and the access to correction functionalities.


2
Archives départementales de Seine-Maritime
Accueil A propos Parcourir Aide Connexion / Inscription
Recherche avancée
Recherche articles
Le Journal de Rouen lundi 5 décembre 1808 Page 3 sur 4 Article 3 sur 7
Imprimer Enregistrer en PDF Enregistrer en image
Marqueurs (Mots Clefs) Ajouter Marqueurs
Commentaires Ajouter un commentaire
Pas de commentaires.
TEXTE TRANSCRIT AUTOMATIQUEMENT Corriger ce texte
Pourquoi ce texte contient des erreurs ?
Aidez nous à corriger le texte !
Remonter à la page PermaLink Share Share
1
3
© LITIS 2009-2012. http://www.plair.org

ROUEN.

Texte [correction la plus récente par StephaneNICOLAS - Voir les corrections]

ROUEN.

Hier, jour anniversaire du couronnement de S. M. l'empereur et roi et de la bataille d'Austerlitz, les autorités civiles, les membres des cours et tribunaux, et tous les corps constitués, se sont rendus à l'église métropolitaine pour assister au mariage d'une jeune fille dotée par la ville, en conformité du décret du 19 février 1806. Avant le Te Deum, chanté en réjouissance des souvenirs que rappelle aux français cette heureuse époque, un sermon sur la solemnité de la fête a été

prononcé, et les deux jeunes époux ont été ensuite présentés à la bénédiction nuptiale qu'ils ont reçu de S. Em. Mgr. le cardinal archevêque de cette ville. Un grand concours de citoyens s'était rendu à la cathédrale pour participer à cette cérémonie. Le soir il y a eu illumination.

***Figure 2**. the online consultation platform*

We describe each one of these functionalities in the following sections, before presenting the general software architecture of this online consultation platform.

Information access and visualization

Due to the great sizes and the complex non-linear layouts of the newspapers, an intuitive and ergonomic visualization module is needed to allow the users to easily browse the digitized images of the pages. The user must be able to comfortably read an article at an adapted display size without being lost in the vastness of the page, and keeping track of the context and the position within the overall document. To enhance the navigation experience of the users, we use a modified version of a high resolution image visualization tool which is called IIPImage [1]. This is an open source tool, which is used around the world by a wide variety of users, including museums, scientific imaging, astronomy, medical imaging and geographical information systems. This tool provides ultra high resolution visualization of images by using pyramidal images. In our case, it means that we can navigate fastly through the different resolutions of the images of the newspaper pages,, zooming to a part, and then zooming back to the entire page by mean of a convenient and intuitive user interface (Figure 2). Furthermore this viewer integrates a miniature view (minimap) of the entire displayed document, allowing the user to always situate the position and the context in the overall document, whatever the zoom level. One of the strength of this viewer is that it doesn't require to install any plug-in on the user's computer, because it is based on the use of the Javascript technology which is natively supported by all the web browser. We modified this open source source to fit our needs. The main modifications we introduced in the original IIPImage tool, concern the highlighting of important information in the document, as to indicate the position of an article, or a keyword in a search result.

Text and article search functionalities

Our application allows the user to access the information at different levels: issue, page and article. The information is provided in image mode and text mode using the OCR results. The access to the issues in image mode is provided through a calendar menu allowing selecting the year, the month and the day of the desired issue. This is a common functionality in almost all the digital archive systems. On the other hand, the exploitation of the full OCR results and the indexing at the article level are less current, and yet certainly much more useful for the users. The OCR results allow the users to search for full text in the corpus of newspapers. Simple or more complex text queries, with classical operators (AND, OR) used in information retrieval systems can be performed. The search results can be refined by applying multiple filters (facets) on various criteria such as the date, the length of items, the category, ... etc. Maintains the indexing module continuously updated by providing the index search for each item on the latest version thereof. This means that if a user finds it useful to correct the text of an article through the collaborative correction module, the indexing engine will assume that the corrections are qualitatively superior to the original text provided by the OCR, it will return therefore results to them.

Collaborative features

As we know it is impossible for automated tools to address all the problems inherent in the initial printing quality of the original documents or due to the inevitable degradation of the paper through time. Therefore, our tools are as good if we wanted to enable archivists to benefit from the willingness of users to the platform to correct the content thereof. We have carefully considered the feedback of other significant initiatives is this domain, in particular the experience of the National Library of Australia (NLA) [2] who proposed to its users to correct the OCR errors at the launch of a consultation website containing millions of newspaper pages. Another revealing experience in this domain, is the Digitalkoot Project launched by the National Library of Finland [3], which investigated the field of crowdsourcing OCR correction through gaming. The enthusiasm and the participative effort of the public in these experiments convinced us that it could be interesting to offer this type of functionality in our system to improve the quality of the provided information. Thus currently we make it possible to the users to correct the OCR errors through the consultation interface, and to save these corrections to refine the indexing for future searches. For that we propose a tool to regenerate automatically the METS/ALTO files using the corrections carried out. The idea is to allow an iterative process to improve and finalize the permanent storage of the XML documents in order to them to represent as accurately as possible the information contained in the original scanned sources. Furthermore the exploitation of the OCR and document segmentation results through these METS/ALTO files, also makes it possible to provide an export of the articles selected by the user into PDF files, either in image mode, or in text mode.

The improvement of the indexing, and thus of the OCR results represent a crucial point in a document consultation system, this is why our future effort will concentrate mainly on this aspect, and in particular on the integration of OCR engines adapted to the processing of degraded characters.

General architecture

In a general manner the global software architecture of our online consultation platform is illustrated on Figure 3. This architecture relies mainly on the use of open source solutions, in order to facilitate the development and the update process of the system, while supporting interoperability:

- the Lucene text search engine for text indexation
- the MySQL relational database management system for storing online data
- the Hibernate framework for querying the database
- the Spring framework for building the web application
- the ALTO and METS XML file formats to describe the physical and logical structure of the digitized documents

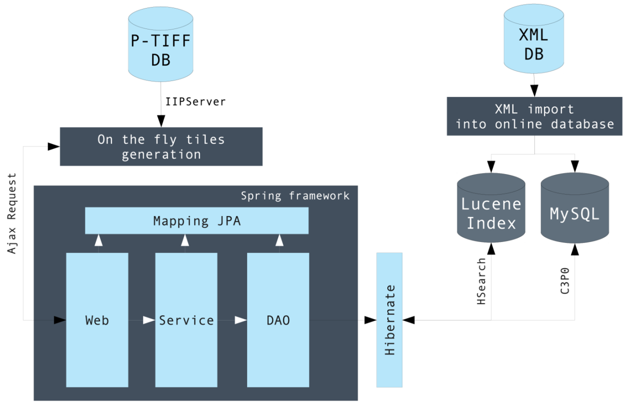


***Figure 3**. the general architecture of our online consultation platform*

After having presented the front-end of our system, and its underlying architecture, we describe now the back-end workflow which allows to process the data feeding the system.

Offline processing workflow

The data exploited by the online consultation platform are processed offline by a fully functional and automatic workflow able to process large amount of digitized data. This is not a straight forward task because of the various problems we discussed in the opening of this paper (degradations, variability, ...). For most of the digital archives, the data are fully manually processed. This is a very tedious and expensive. For this reason we chose to develop an automatic workflow as reliable as possible. The central component of this workflow is a tool dedicated to the analysis of the physical and logical structure of the documents, which aims at providing a description at various levels. Then these data are combined with OCR results to feed METS/ALTO files which will contain at the same time the content of the documents, and also the description of the structuration of this content (text line and article position, reading order, ...).

The global architecture of this processing workflow, and its constituting components are briefly described in the following sections.

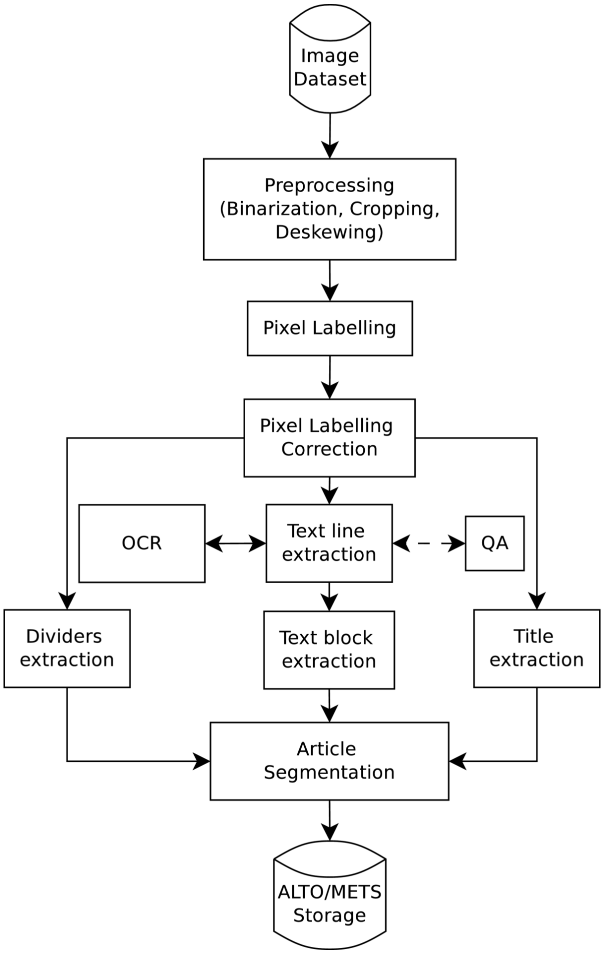


***Figure 4**. the architecture of our processing workflow*

General architecture of the processing workflow

The workflow we propose is built as a traditional document image analysis system with preprocessing, analysis and formatting stages. The analysis stage results of intensive research works completed at the LITIS laboratory on the fields of statistical modelling and machine learning applied to document analysis. For the other stages we tried to apply the most adapted techniques. As for the online consultation platform we aim at provide the most effective and efficient system. Once again we exploit several well known standards and efficient open source software elements. Thus for example our workflow uses the OpenCV1 library[1], known for its speed and robustness. This library offers a support for the latest technologies such as GPU computing and android smart-phone support. We also use the cvBlob2 library[2] because of the fast implementation of the labelling algorithm described in [4] which can obtains external and internal contours of each blob in the same pass and labels 8-connectivity components.

Globally the aim of this workflow is to have a generic system capable of detecting the problems specific to each document, and then to apply the best state-of-the-art algorithms to segment and recognize the content of these documents. In addition, we keep in mind that our system should eventually allow to quickly process large amounts of images, which means taking into account the quality / performance ratio. This section describes each step of our workflow, as illustrated on Figure 4.

Preprocessing

Binarization

The aim of the binarization process is to separate the foreground and the background of the document images. This is generally performed using local or global thresholding methods.

Due to the typical degradation of historical documents, global thresholding techniques cannot be used. That’s why the binarization step is performed through the algorithm proposed by Sauvola [5], which is implemented using integral images like in [6]. This method is particularly effective for the binarization of degraded documents because of its local nature, although our experiments have highlighted some problems limiting its use as a part of our industrial application. In order to improve the binarization while keeping in mind the industrial objective of our system, we have studied the results of the binarization competition DIBCO2009 [7] and we will implement and test in a future work, some of the top algorithms of this competition like Lu et al. [8].

Cropping/Deskewing

The next preprocessing step in our processing workflow concerns the removal of the large black edges of the digitized documents. In fact, the scanning rules of newspapers require to see the outer limits of the document to ensure that the scanner has captured the whole page. In addition, newspaper pages are kept in books, which reveals the thickness of the pages and also the binding. The resulting black edges produced by the digitization, do not bring any information for the users, and are not important for the next steps of the processing workflow. On the contrary, they increase the amount of data to be processed unnecessarily. That is the reason why we use an algorithm derived from the OCROpus project [9] to crop the images in order to remove these dark edges and the useless parts of text of them. This algorithm greatly improves the further analysis tasks by avoiding the need of categorizing useless black pixels outside the print space of the document. The newspaper images are often twisted because of the storage condition of the pages in books (weight of the binding and other pages, humidity conditions, ...). This is why we use an algorithm based on the Hough transform to automatically detect the inclination angle of the page being processed. Using the determined angle, we calculate a deskewed image based on the quadrangle for both the grayscale and the binarized images. These are the images that will be used in the next steps of the workflow.

Logical structure document analysis

This is the main component of our workflow which aims at analyzing the document images at various levels in order to provide an accurate description of the document structure which will combined with the OCR results to feed the METS/ALTO files

[1] http://opencv.willowgarage.com/wiki/

[2] http://code.google.com/p/cvblob/

needed by the online consultation platform. Our analysis method is based on the principle of multiscale image analysis. A functional labelling of the image is firstly provided at a pixel level so as to guarantee certain accuracy in the results and a very fine level of description. This image labelling which allows extracting the structural elements of the documents is then refined using labelling correction stage, and some high-level structuration rules are applied to determine the complete structure of the document. The method allowing the automatic extraction of the metadata is based on a segmentation step of the document, made possible by a novel conditional random field model which exploits multiscale quantization [10]. This system provides a fine segmentation where each pixel is assigned a label which characterize its functional role in the document (Figure 5a) The obtained segmentation is therefore not only a division into physical blocks because it also provides a logical identification of the detected entities. After a smoothing correction stage, this labelling allows to extract the rule lines, the title, the text lines and then the text blocks.

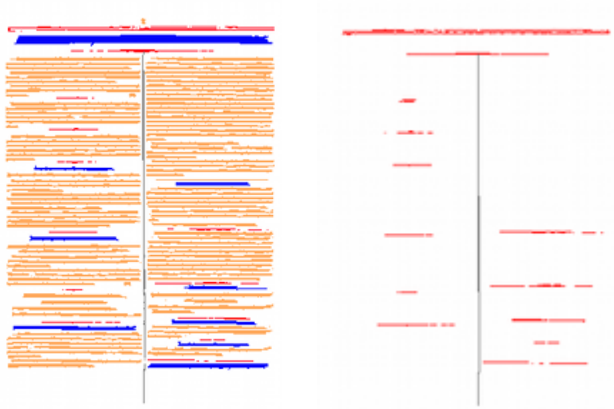


***Figure 5**. a) logical pixel labelling result, b) extracted separators*

Metadata creation

The horizontal and vertical separators (rule lines) are robust structural information in newspapers, even if they are sometimes broken due to the degradations of the document and the artifacts introduced by the digitization process. That is the reason why we use them to extract the base component articles. Indeed, the vertical separators delimit the columns of the document, while the horizontal dividers cut out the items in these columns, and thus define also the different parts of the document (Figure 5b). The title fields are also discriminative information since they indicate the beginning of an article. We can therefore consider the parsed document as a tree of blocks representing the various parts of it. The root of this tree consists of a single block representing the entire page, while the level below is constituted by the blocks delimited by the horizontal dividers broadest and so on on the next level until to obtain the full hierarchy which gives the position of all the blocks and separators in the document structure, as it can be seen on Figure 6. The set of separators on a page is therefore a logical grid describing the document structure. We exploit then this information to define a list of articles logically arranged and we classify them by reading order according to their position in the document and in the hierarchical structure of the document. Each of these items consists of one or more boxes of our grid of separators.

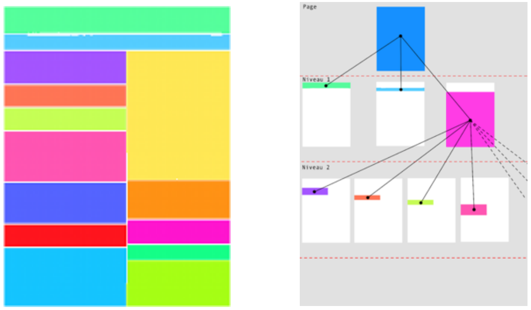


***Figure 6**. a) grid of extracted blocks, b) document structural hierarchy*

Finally all these extracted metadata are combined with the OCR results given at the input of the system, to feed the METS/ALTO files which will contain the full multi-level description of the content of the digitized documents. METS and ALTO are two metadata standards using the XML schema language. They are used more and more in the design of digital archives. METS (Metadata Encoding and Transmission Standard) encodes the descriptive and structural metadata regarding objects within a digital library, while ALTO (Analysed Layout and Text Object) describes the geometrical layout and logical structure of textual documents processed by document image analysis system and OCR. These METS/ALTO files are used by our online consultation platform to allow the users to access the contained information at various level of description, and the role of our offline processing workflow is to provide a description as accurate and correct as possible.

Conclusion and future works

We have presented in this paper a complete system dedicated to the online consultation of large corpora of old newspaper archives. This system enhances the browsing and searching experience of the users by proposing a convenient visualization interface to access the documents at different levels of description: the issue, the page and the article. To obtain such a multi-level description of the documents contained in the database, a complete processing workflow has been designed. This workflow allows to extract the logical structure of articles automatically, thanks to models and machine learning procedures, developed in the context of research works led by the LITIS laboratory. These extracted metadata feed XML files in ALTO/METS formats that are dedicated to the description of the physical and logical description of the documents, and authorize a more specialized and accurate information exploitation by the visualization component. An indexing of the textual content of the articles is performed using the well known and open source Lucene text search engine, based on the OCR input data associated to the document images. Our system is designed to consider any OCR results. As these input data cannot be always accurate and correct due to the task complexity, our system proposes collaborative functionalities to easily correct the OCR text of the articles. That makes it possible to improve the quality of the indexing and thus make more reliable search queries. To date we have a fully operational demonstrator of this system that has been applied on one year of daily regional newspaper issues. This demonstrator initially will be tested and approved by our archivist partners before being proposed online for general public evaluation. Our future works will concern the improvement of this system, and especially the automatic processing for document structure analysis and character recognition.

References

[1] D. Pitzalis, R. Pillay, C. Lahanier, "A new Concept in high Resolution Internet Image Browsing", in 10th International Conference on Electronic Publishing (ELPUB), 2006.

[2] R. Holley, "Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers", in National Library of Australia Staff Papers, 2009.

[3] O. Chrons, S. Sundell, "Digitalkoot: Making Old Archives Accessible Using Crowdsourcing", in Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011.

[4] F. Chang, C. Chen, and C. Lu, "A linear-time component- labeling algorithm using contour tracing technique," Computer Vision and Image Understanding, vol. 93, no. 2, pp. 206–220, 2004.

[5] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," Pattern Recognition, vol. 33, no. 2, pp. 225– 236, 2000.

[6] F. Shafait, D. Keysers, and T. Breuel, "Efficient implementa- tion of local adaptive thresholding techniques using integral images," Document Recognition and Retrieval XV, vol. 6815, no. 1, p. 681510, 2008.

[7] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "Icdar 2009 doc- ument image binarization contest (dibco 2009)," in Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on. IEEE, 2009, pp. 1375–1382.

[8] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," Int. J. Doc. Anal. Recognit., vol. 13, pp. 303–314, December 2010.

[9] F. Shafait and T. Breuel, "A simple and effective approach for border noise removal from document images," in Multitopic Conference, 2009. INMIC 2009. IEEE 13th International. IEEE, 2009, pp. 1–5.

[10] D. Hebert, T. Paquet, S. Nicolas, "Continuous CRF with Multi-scale Quantization Feature Functions Application to Structure Extraction in Old Newspaper", in 2011 International Conference on Document Analysis and Recognition, pp.493-497, 2011.

Author Biography

Thomas Palfray is working in a team dedicated to document analysis as a research engineer since 2007. He is primarily responsible for enhancing research results carried out within the team by creating industrial tools capable of exploiting the additionnal value of these. Moreover, he is a specialist in semi-structured languages and is currently the lead developper of the project PlaIR.

Stéphane Nicolas received the PhD degree in computer science from the University of Rouen, France, in 2006, on image segmentation using conditional random fields for document image indexing. He is currently on assistant professor at the University of Rouen since september 2007, and a researcher of the LITIS laboratory where he integrates the "Document and Learning" group. Stéphane Nicolas is a member of the french association for pattern recognition (AFRIF) and a member of the french research group on handwriting recognition GRCE. His main research interests include computer vision, image analysis, pattern recognition, machine learning, and statistical tools for signals modeling and classification, mainly applied to handwritten document layout analysis and information extraction from handwritten documents.

Thierry PAQUET received the Ph.D. degree from the University de Rouen (1992) in the field of Pattern Recognition. From 1992 to 2002 he has been appointed has a Senior Lecturer at the University of Rouen where he tought Signal and Image Processing. From 1992 to 1996 he was involved in an industrial collaboration with Matra MCS and the French Postal Research Center (SRTP) for the automation of mail sorting and bank checks reading. Thierry PAQUET was appointed as a full professor in 2002 at the University of Rouen. His current research area concern statistical Pattern Recognition and Document Image Processing with application to Handwriting Recognition, Handwritten document categorization, Biometry through Handwriting modality, robust OCR for historical documents. Thierry PAQUET was vice director of LITIS laboratory from 2007 to 2011. He is the director of LITIS Laboratory since 2012. He was the president of the French association Research Group on Document Analysis and Written Communication (GRCE) from 2002 to 2010.

After a PhD of Artificial intelligence in 2005, Pierrick Tranouez has been since 2008 a research engineer in the Document and Learning team of the LITIS laboratory, in the University of Rouen. He is a member of projects PlaIR and DocExplore. These projects try to find new innovative ways of making archives or other written materials available for digital interaction : how can we read them, but also index them, correct them, all in all enhance them ?