A structural method based on texture for ancient document image analysis

Mehri Maroua, Pierre Héroux, Petra Gomez-Krämer, Rémy Mullot

A structural method based on texture for ancient document image analysis

Student's name: Maroua MEHRI
Supervisor/s of the thesis: Rémy MULLOT, Pierre HÉROUX and Petra GOMEZ-KRÄMER
University: University of La Rochelle
Starting date of the PhD: November 02, 2011
Expected finalization date of the PhD: May 28, 2015
Email:maroua.mehri@univ-lr.fr

***Abstract*—A structural signature based on texture for the characterization and categorization of digitized historical book pages is proposed in my research work. The proposed signature does not assume *a priori* knowledge regarding page layout and content, and hence, it is applicable to a large variety of ancient books. By integrating varying low-level features (e.g. texture) characterizing the different page components (*i.e.* different text fonts or graphic regions) on the one hand, and structural information describing the page layout on the other hand, the proposed signature provides a rich and holistic description of the layout and content of the analyzed book pages. More precisely, the signature-based characterization approach consists of two stages. The first stage is extracting automatically homogeneous regions. Then, the second one is proposing a graph-based page signature, which is based on the extracted homogeneous regions, reflecting its layout and content. This signature ensures the implementation of numerous applications for managing effectively a corpus or collections of books (e.g. information retrieval in digital libraries according to several criteria, or page categorization). To illustrate the effectiveness of the proposed page signature, a detailed experimental evaluation has been conducted in this work for assessing two possible categorization applications, unsupervised page classification and page stream segmentation. In addition, the different steps of the proposed approach have been evaluated on a large variety of historical document images.**

I. Short research plan

A. Introduction

Over the last few years, there has been tremendous growth in digitizing collections of cultural heritage documents. Thus, many challenges and open issues have been raised, such as information retrieval in digital libraries, or analyzing page content of digitized historical books (DHBs). Therefore, with the support of the French National Research Agency[1], we are working on a project named DIGIDOC[2]. The ultimate goal of the DIGIDOC project is developing new ways of interacting with scanners by assisting the digitization operator to adjust automatically the best set of parameters (e.g. resolution, lightening, color calibration), detecting errors in the digitization process (e.g. blur, skewed or folded pages), providing appropriate assistance for document indexing (e.g. by recognizing automatically page types, or breaks in a sequence of pages), *etc.* There is an absolute need to design "smart" digitizers which can limit manual intervention and perform easy and high quality digitization of document images (DIs) [1]. Therefore, to achieve better interaction with scanners, we need to design a computer-aided categorization tool, able to index or categorize DHB pages according to several criteria, mainly the layout structure, graphical properties, or typographical characteristics of their content.

Several scientific works in contemporary document image analysis (DIA) have described several relevant approaches enabling multiple forms of indexing and classification based on content analysis of DIs. The current systems for categorizing digitized DIs are based on several criteria for the textual content by applying optical character recognition (OCR) or by using the interest point detection approach [2]. Nevertheless, the transposition of these tools for historical DIA, that are dedicated initially for contemporary DIA, is not a straightforward task.

As a matter of fact, the ultimate goal of my research is to propose a structural signature based on texture for DHB page characterization and categorization. An automatic approach for characterization and categorization of DHB pages is presented in this work. The proposed approach is applicable to a large variety of DHBs. In addition, it does not assume *a priori* knowledge regarding document image layout and content. It is based on the use of texture and graph algorithms to provide a rich and holistic description of the layout and content of the analyzed book pages to characterize and categorize DHB pages. The categorization is based on the characterization of the digitized page content by texture, shape, geometric and topological descriptors. This characterization is represented by a structural signature.

B. Methodology

Supported by the fact that pages of the same book usually present strong similarities in the organization of the HDI information (*i.e.* layout) and in the graphical and typographical features (*i.e.* content) throughout the DHB pages under consideration, our goal is to propose an approach that is used on an entire book instead of processing each page individually, for the segmentation and analysis of DHB content, and characterization and categorization of DHB pages. The aimed approach should not require *a priori* knowledge of the layout, typographical parameters or graphical properties of the analyzed DHB pages. It can extract automatically low-level features for discriminating the different classes of the

[1] http://www.agence-nationale-recherche.fr/en/

[2] http://www.agence-nationale-recherche.fr/en/anr-funded-project/?tx_lwmsuivibilan_pi2[CODE]=ANR-10-CORD-0020

foreground layers, through the analysis of the similarity and repetition information which is deduced from many DHB pages. Then, we aim to determine a region or group of pixels which share similar properties or characteristics on the basis of which they are grouped. These characteristics may be based on the localization of pixels and their surroundings, color, intensity or texture. In this work, we will focus only on texture-based features.

In order to ensure a distinction between different text fonts and various kinds of graphics, three assumptions are made [3]. First, the textual regions in a digitized DI are considered as textured areas, while its non-text content is considered as regions with different textures. Secondly, text with a different font is distinguishable. Finally, different types of graphics can be also separated. Thus, in this work various aspects of texture features have been explored in HDIs to assist the analysis of their content by characterizing a HDI through a set of homogeneous regions. Therefore, a data-driven or bottom-up strategy of analysis has been adopted in this work which is based on low-level data mining of pixels (e.g. texture, position, shape, geometry). This strategy investigates the texture and topology-based pixel properties (the spatial distribution of gray-levels) to determine the homogeneous or similar content regions in the analyzed HDI.

- First, faced with a large diversity of texture-based methods, few questions arise. Which texture methods are firstly well suited for segmenting graphical regions from textual ones, discriminating text in a variety of situations of different fonts and scales and separating different types of graphics ? Then, which texture approaches represent a constructive compromise between the performance (*i.e.* segmentation quality) and computational cost (*i.e.* memory requirements, processing time, numerical complexity and texture vector dimensionality) ? It is well-known that the success or failure of texture-based segmentation method tightly depends on the type of the extracted and used texture features. Thus, an experimental evaluation and benchmarking of a number of commonly and widely used texture approaches have been firstly conducted on a large corpus of HDIs, to have satisfactory and clear answers to the above questions. This work has shown the effectiveness of the different texture analysis approaches in the field of historical DIA.
- Given that there is a wide variety of DHB layouts and contents, having significant degradation levels and different noise types, proposing an approach that does not require any *a priori* knowledge, to characterize automatically DHB pages, is not a straightforward task. However, based on the hypothesis that some similarities of HDI content type can be deduced from many book pages and based on the assumption that a DI content type can be repeated on many pages of the same book [4], [5], a framework that works effectively at the entire book scale, instead of processing each book page individually, is proposed in this work. The proposed framework ensures the pixel-based characterization of the content of an entire DHB. It is automatic and it can be adapted to all kinds of books. It is independent of DI layout, typeface, font size, orientation, DI size, digitizing resolution and intensity, *etc.* It is also robust in the case of different kinds and levels of noise and degradation present in HDIs. Moreover, it does not require any manual inspection or *a priori* knowledge regarding DI content and structure or layout.
- A raising interest is noticeable recently to the use of statistical and structural pattern recognition tools to retrieve objects and classify them [6]. In DIA, the statistical and structural approaches are broadly applied for DI representation [7], [8]. A structural signature based on texture for each DHB page is proposed in this work. The proposed DHB page signature is characterized with a set of extracted homogeneous or similar content regions defined by similar texture, shape and geometric attributes and their topology. It does not assume *a priori* knowledge regarding the layout and content of the analyzed DHB pages, and hence, it is applicable to a large variety of ancient books. It integrates varying low-level features (*i.e.* texture, shape, geometric and topological descriptors) characterizing the different HDI content components (*i.e.* different text fonts or graphic regions) on the one hand, and structural information describing the HDI layout on the other hand. This rich and holistic representation of the layout and content of the analyzed DHB page can be adapted to the user preferences and specified criteria through the extracted varying levels of information (e.g. by selecting only the information characterizing the HDI layout and/or content or by retrieving any useful information available for a subsequent use). It provides a topological signature of DHB page according to several criteria, mainly the layout structure and/or typographic/graphical characteristics of the HDI content. Finally, using the obtained signatures which are modeled in the form of graphs, the similarities of DHB page structure or layout and/or content can be deduced by means of a graph dissimilarity, the graph edit distance (GED). Indeed, the DHB pages can be compared by categorizing the designed signatures which model the layout and content of DHB pages. In fact, DHB pages with similar layout and/or content can be grouped.

C. Future Work

There are many directions to proceed in the work presented in this article.

The first aspect of future work will be to use the proposed methods and studies in this dissertation on a larger database. This is ongoing and will evaluate the different studies and proposed methods more adequately with more convincing experimental results in order to help improve their scalability. We will then study and combine statistical, geometric, model-based and spectral texture-based features in order to refine the segmentation and ensure a distinction between different text fonts and various graphic types.

Historical DIA is still an open issue for both supervised and unsupervised methods due to the variability of the contents and/or layouts of historical documents. As for the supervised

methods, feature learning or representation learning [9] will be investigated for pixel-classification in future research. This helps in dealing with retrieving relevant features or representations from raw data. In addition, a feature selection step (e.g. dimension reduction technique) can also be integrated to select relevant features and remove redundant ones.

In this work, a generic signature for DHB page characterization and categorization has been evaluated on two possible applications, unsupervised book page classification and book page stream segmentation, with no hypothesis concerning page layout and content. Then, we will assess other possible applications of the proposed graph-based signature:

- Finding pages in a DHB or HDI corpus which contain a particular content component or a group of patterns that match specific criteria defined by a user (*i.e.* investigating the sub-graph isomorphism paradigm).
- Retrieving similar pages in a HDI corpus query tool by establishing a ranking based on the computed GEDs between the corpus pages and the query page. This ranking can be adjusted automatically according to the weights of each category of the computed features in the cost of the edit operations when performing the GED.
- Detecting the scanning failure occurring during the digitization process (e.g. curvature, light) to ensure effective computed-aided quality control of the digitization, *etc.*

Finally, we will investigate a finer unsupervised book page classification with different values of the number of clusters. We also intend to analyze the impact of different feature weighting schemes in the cost of the edit operations when computing the GED. In addition, further work also needs to compare the results given by using the approximate GED computed on the involved graph-based DHB page signatures with other state-of-the-art graph dissimilarity techniques.

II. Curriculum Vitae

Name: Maroua MEHRI
Date of Birth: 23 Feb. 1986
Place of Birth: Sousse - Tunisia
Address: L3i - University of La Rochelle - Pascal Building, Office 123 - Michel Crpeau Str. - 17042 La Rochelle - France
Website: http://l3i.univ-larochelle.fr/Mehri-Maroua
https://sites.google.com/site/marouamehri/
Email address: maroua.mehri@univ-lr.fr
maroua.mehri@gmail.com

A. Education

Nov. 2011 - May. 2015: Ph.D. in computer science
Laboratoire Informatique, Image et Interaction (L3i), University of La Rochelle - France
Laboratoire d'Informatique, du Traitement de l'Information et des Systmes (LITIS), University of Rouen - France
Under the supervision of Pr. Dr. Rémy MULLOT, Dr. Pierre HÉROUX and Dr. Petra GOMEZ-KRÄMER
Funded by the DIGIDOC ANR-10-CORD-020 research project

Sep. 2010 - Feb. 2011: M.Sc. in signal/image processing
Electronics and Telecommunications (ET): Signal, Image, Embedded Systems and Automatic (SISEA)
University of Rennes 1 - France

Sep. 2009 - Jun. 2010: M.Sc. in signals and communicating systems
Intelligent and Communicating Systems (SIC)
National Engineering School of Sousse (ENISo), University of Sousse - Tunisia

Sep 2006 - Jan. 2008: B.Sc. in applied computer engineering
Telecommunications and Industrial Networks (TIN)
ENISo, University of Sousse - Tunisia

Sep. 2004 - Jun. 2006: Preparatory courses for national entrance examination to engineering education cycle
Mathematics and Physics (MP)
Preparatory School for Engineer Studies of Tunis (IPEIT), University of Tunis - Tunisia

B. Experience

Oct. 2014 - Aug. 2015: Temporary lecturer and research assistant
Computer Science Department, Institute of Technology (IUT), University of La Rochelle - France

Oct. 2012 - Sep. 2014: Teaching assistant
Computer Science Department, IUT, University of La Rochelle - France

Mar. 2011 - Aug. 2011: M.Sc. internship in image registration and data analysis in DBS
VisAGeS - IRISA/INRIA - INSERM U746 - France
Under the supervision of Dr. Pierre JANNIN, Dr. Florent

LALYS and Dr. Mohamed Lassaad AMMARI
Committee members: Pr. Dr. Pascal HAIGRON, Pr. Dr. Xavier MORANDI,
Dr. Mohamed Ali MAHJOUB and Dr. Khaled KAÂNICHE

Sep. 2010 - Feb. 2011: M.Sc. internship in blind channel equalization in SISO or SIMO context
LTSI - INSERM U642 - France
Under the supervision of Dr. Laurent ALBERA

Jul 2009 - Mar. 2010: Engineer (software development)
DotFOSS company, Sousse - Tunisia

C. Publications

Journal papers*:*

1) **M. Mehri**, P. Gomez-Krämer, P. Héroux, A. Boucher and R. Mullot, A Texture-based Pixel Labeling Approach for Historical Books. Pattern Analysis and Applications, Springer-Verlag, pages 1-40, 2015.
2) F. Lalys, C. Haegelen, **M. Mehri**, S. Drapier, M. Vérin and P. Jannin, Anatomo-clinical atlases correlate clinical data and electrode contact coordinates: Application to subthalamic deep brain stimulation, Journal of Neuroscience Methods, Elsevier Science, 212 (2), pages 297-307, 2013.

International conference papers*:*

1) **M. Mehri**, P. Héroux, J. Lerouge, P. Gomez-Krämer and R. Mullot, A Structural Signature Based on Texture for Digitized Historical Book Page Categorization. *International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, 2015 [accepted].
2) **M. Mehri**, P. Gomez-Krämer, P. Héroux, M. Coustaty, J. Lerouge and R. Mullot, A Bottom-up Method Using Texture Features and a Graph-based Representation for Lettrine Recognition and Classification. *International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, 2015 [accepted].
3) J. C. Burie, J. Chazalon, M. Coustaty, S. Eskenazi, M. M. Luqman, **M. Mehri**, N. Nayef, J. M. Ogier, S. Prum and M. Rusiñol, SmartDoc: Smartphone Document Capture and OCR Competition. *International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, 2015 [submitted].
4) **M. Mehri**, P. Héroux, N. Sliti, P. Gomez-Krämer, N. E. B. Amara and R. Mullot, Extraction of Homogeneous Regions in Historical Document Images. *In Proceedings of the* 10^{th} *International Conference on Computer Vision Theory and Applications (VISAPP)*, SciTePress, Berlin, Germany, 2015.
5) **M. Mehri**, N. Sliti, P. Héroux, P. Gomez-Krämer, N. E. B. Amara and R. Mullot, Use of SLIC superpixels for ancient document image enhancement and segmentation. *In Proceedings of the* 22^{nd} *Document Recognition and Retrieval (DRR), Part of the IS&T/SPIE 27th Annual Symposium on Electronic Imaging*, SPIE, San Francisco, CA, USA, 2015.
6) **M. Mehri**, M. Mhiri, P. Héroux, P. Gomez-Krämer, M. A. Mahjoub and R. Mullot, Performance Evaluation and Benchmarking of Six Texture-based Feature Sets for Segmenting Historical Documents. *In Proceedings of the* 22^{nd} *International Conference on Pattern Recognition (ICPR)*, IEEE, pages 2885-2890, Stockholm, Sweden, 2014.
7) **M. Mehri**, P. Gomez-Krämer, P. Héroux, A. Boucher and R. Mullot, A Pixel Labeling Framework for Comparing Texture Features: Application to Digitized Ancient Books. *In Proceedings of the* 3^{rd} *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, SciTePress, pages 553-560, Angers, France, 2014.
8) **M. Mehri**, P. Héroux, P. Gomez-Krämer, A. Boucher and R. Mullot, A Pixel Labeling Approach for Historical Digitized Books. *In Proceedings of the* 12^{th} *International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, pages 817-821, Washington, DC, USA, 2013.
9) **M. Mehri**, P. Gomez-Krämer, P. Héroux and R. Mullot, Old document image segmentation using the autocorrelation function and multiresolution analysis. *In Proceedings of the* 20^{th} *Document Recognition and Retrieval (DRR), Part of the IS&T/SPIE 25th Annual Symposium on Electronic Imaging*, SPIE, San Francisco, CA, USA, 2013.
10) **M. Mehri**, F. Lalys, C. Maumet, C. Haegelen and P. Jannin, Analysis of electrodes' placement and deformation in deep brain stimulation from medical images. *In Proceedings of Medical Imaging: Image-Guided Procedures, Robotic Interventions and Modeling*, SPIE, San Diego, CA, USA, 2012.

International workshop papers*:*

1) **M. Mehri**, V. C. Kieu, M. Mhiri, P. Héroux, P. Gomez-Krämer, M. A. Mahjoub and R. Mullot, Robustness Assessment of Texture Features for the Segmentation of Ancient Documents. *In Proceedings of the* 11^{th} *International workshop on Document Analysis System (DAS)*, IEEE, pages 293-297, Tours, France, 2014.
2) **M. Mehri**, P. Gomez-Krämer, P. Héroux, A. Boucher and R. Mullot, Texture Feature Evaluation for Segmentation of Historical Document Images. *In Proceedings of the* 2^{nd} *International Workshop on Historical Document Imaging and Processing (HIP)*, ACM, pages 102-109, Washington, DC, USA, 2013.

REFERENCES

[1] F. LeBourgeois, E. Trinh, B. Allier, V. Eglin, and H. Emptoz, "Document images analysis solutions for digital libraries," in *DIAL*, 2004.

[2] O. Augereau, N. Journet, A. Vialard, and J. P. Domenger, "Improving classification of an industrial document image database by combining visual and textual features," in *DAS*, 2014.

[3] B. Julesz, "Visual pattern discrimination," *IT*, 1962.

[4] A. Piper, "Reading's refrain: from bibliography to topology," *Readings: Selected Essays from the English Institute*, 2013.

[5] E. T. Nalisnick and H. S. Baird, "Extracting sentiment networks from Shakespeare's plays," in *ICDAR*, 2013.

[6] H. Bunke and K. Riesen, "Towards the unification of structural and statistical pattern recognition," *PRL*, 2012.

[7] H. Bunke, S. Günter, and X. Jiang, "Towards bridging the gap between statistical and structural pattern recognition: two new concepts in graph matching," in *ICAPR*, 2001.

[8] S. Jouili, M. Coustaty, S. Tabbone, and J. M. Ogier, "NaviDoMass: structural-based approaches towards handling historical documents," in *ICPR*, 2010.

[9] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *PAMI*, 2013.