



# DAS 2014

11<sup>th</sup> IAPR INTERNATIONAL WORKSHOP ON DOCUMENT ANALYSIS SYSTEMS

April, 7<sup>th</sup> - 10<sup>th</sup>

# Short Paper Booklet

**TOURS, Loire Valley, France** ----->

Vinci - International Convention Centre

<http://das2014.univ-tours.fr/>

## Workshop chairs

Jean-Yves Ramel (University of Tours - France)

Marcus Liwicki (DFKI Kaiserslautern - Germany; University of Fribourg - Switzerland)

## Program chairs

Jean-Marc Ogier (University of La Rochelle - France)

Koichi Kise (Osaka University - Japan)

Ray Smith (Google - USA)

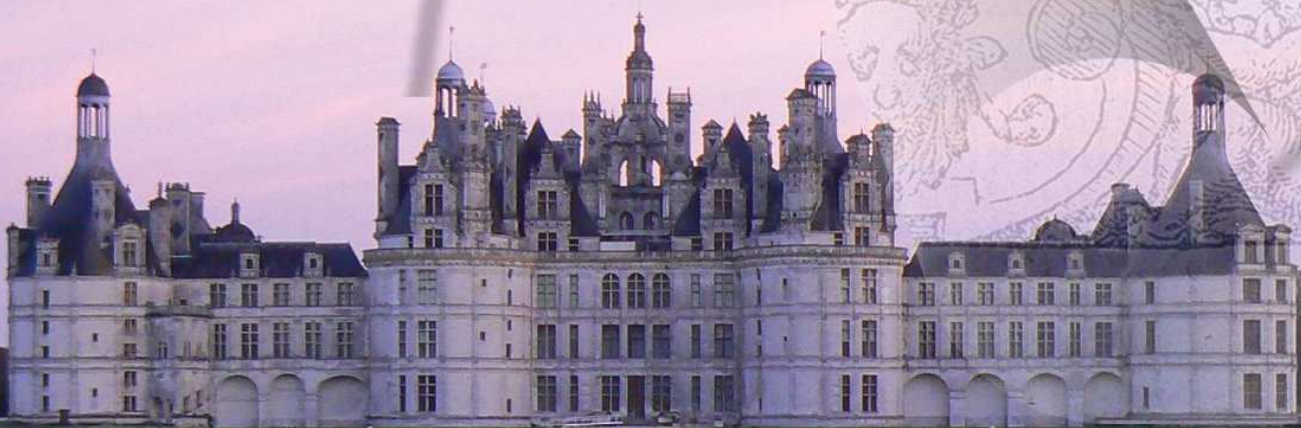
LOIRE VALLEY - TOURS - RIGNY USSE - AMBOISE - CHAUMONT SUR LOIRE - CHENONCEAU - VILLANDRY - AZAY LE RIDEAU - LANGEAIS - CHINON - BLOIS - SAUMUR - ANGERS - CHEVERNY - CHAMBORD - LOCHES - ORLEANS - SULLY SUR LOIRE - GIEN

**LI**

Laboratoire d'Informatique  
EA 6300

**IAPR** 

  
**UNIVERSITÉ  
FRANÇOIS - RABELAIS  
TOURS**





# **11<sup>th</sup> IAPR International Workshop on Document Analysis Systems**

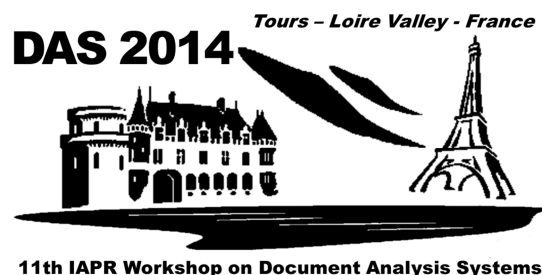


## **Short paper booklet**

**Tours - Loire Valley, France**

Vinci-International Convention Centre

April 7<sup>th</sup>-10<sup>th</sup>, 2014





## **Committees**

### **Workshop Organizing Chairs**

Jean-Yves Ramel, *University of Tours – France*

Marcus Liwicki, *DFKI Kaiserslautern – Germany; University of Fribourg, Switzerland*

### **Program Chairs**

Jean-Marc Ogier, *University of La Rochelle – France*

Koichi Kise, *Osaka Prefecture University – Japan*

Ray Smith, *Google – USA*

### **Sponsor Chairs**

R. Lins, *University of Pernambuco – Brazil*

Cheng Lin Liu, *NLPR Beijing – China*

### **Tutorial Chairs**

Josep Lladós, *CVC – Spain*

Vincent Poulain D'Andecy, *ITESOFT – France*

### **Publicity Chairs**

Venu Govindaraju, *Buffalo University – USA*

Umapada Pal, *Indian Statistical Institute – India*

### **Organizing Committee**

Alireza Alaei

Sabine Barrat

Thierry Brouard

Mathieu Delalandre

Nathalie Girard

Christelle Grange

Muhammad Muzzamil Luqman

Beatrice Pawlik

Nicolas Ragot

Romain Raveaux

Nicolas Sidere

Arundhati Tarafdar

## Program Committee

### Academic members

A. Antonacopoulos, *University of Salford - UK*  
E. Barney Smith, *Boise State University - USA*  
A. Belaid, *Université de Lorraine - France*  
M. Blumenstein, *Griffith University - Australia*  
A. Dengel, *DFKI - Germany*  
D. Doermann, *University of Maryland - USA*  
A. Fischer, *Concordia University - Canada*  
B. Gatos, *IIT Demokritos - Greece*  
V. Govindaraju, *University at Buffalo - USA*  
R. Ingold, *University of Fribourg - Switzerland*  
M. Iwamura, *Osaka Prefecture University - Japan*  
D. Karatzas, *CVC - Spain*  
F. Kimura, *Mie University - Japan*  
K. Kise, *Osaka Prefecture University - Japan*  
L. Likforman-Sulem, *Telecom-ParisTech - France*  
C.L. Liu, *NLPR - China*  
M. Liwicki, *DFKI - Germany*  
J. Lladós, *CVC-UAB - Spain*  
D. Lopresti, *Lehigh University - USA*  
S. Marinai, *University of Florence - Italy*  
J.M. Ogier, *University of La Rochelle - France*  
I.S. Oh, *Chonbuk National University - Korea*  
S. Omachi, *Tohoku University - Japan*  
U. Pal, *Indian Statistical Institute - India*  
J.Y. Ramel, *University of Tours - France*  
M. Rusiñol, *CVC - Spain*  
F. Shafait, *University of Western Australia - Australia*  
P. Shivakumara, *University of Malaya - Malaysia*  
C.L. Tan, *National University of Singapore - Singapore*  
K. Tombre, *Université de Lorraine - France*  
S. Uchida, *Kyushu University - Japan*

## **Industrial members**

D. Bloomberg, *Google - USA*  
M. Gillmann, *Insiders Technologies GmbH - Germany*  
Y. Hotta, *Fujitsu Laboratories Ltd - Japan*  
J. Hu, *IBM - USA*  
C. Kermorvant, *A2iA - France*  
F. Mamalet, *Orange Labs R&D - France*  
S. Naoi, *Fujitsu R&D Center Co Ltd - China*  
V. Poulain d'Andecy, *ITESOFT - France*  
K. Pramod Sankar, *Xerox Research - India*  
M.-P. Schambach, *Siemens AG - Germany*  
R. Smith, *Google - USA*  
V. Subramaniam, *IBM - India*  
J. Sun, *Fujitsu R&D Center Co Ltd - China*  
R. Unnikrishnan, *Google - USA*

## **Subreviewers**

Alexandros Papandreou, *NCSR Demokritos - Greece*  
Alicia Fornés, *Computer Vision Center - Spain*  
Alireza Alaei, *LI - University of Tours - France*  
Anastasios Kesidis, *TEI of Athens - Greece*  
Annegret Liebers, *Siemens AG - Germany*  
Arti Shivram, *University at Buffalo - USA*  
Bo Bai, *NLPR CASIA - China*  
Bolan Su, *Institute for Infocomm Research - Singapore*  
Chetan Ramaiah, *University at Buffalo - USA*  
Christian Clausner, *University of Salford - UK*  
Christian Schulze, *DFKI - Germany*  
Christoph Pesch, *Siemens AG - Germany*  
Christos Papadopoulos, *University of Salford - UK*  
Chunpeng Wu, *Fujitsu R&D Center - China*  
Darko Obradovic, *DFKI - Germany*  
Devansh Arpit, *SUNY Buffalo - USA*  
Fei Yin, *NLPR Beijing - China*  
Fouad Slimane, *IFN - Technische Universität Braunschweig - Germany*  
Georgios Louloudis, *NCSR Demokritos - Greece*  
Hiroaki Takebe, *Fujitsu Laboratories Ltd. - Japan*  
Hiroshi Tanaka, *Fujitsu Laboratories Ltd. - Japan*  
Jayant Kumar, *University of Maryland College Park - USA*  
Jin Chen, *Lehigh University - USA*  
Joerg Rottland, *Siemens AG - Germany*  
Jon Almazán, *Computer Vision Center - Spain*

Joseph Chazalon, *University of La Rochelle - France*  
Kostas Ntirogiannis, *NCSR Demokritos - Greece*  
Le Kang, *University of Maryland - USA*  
Li Chen, *FRDC - China*  
Liang Xu, *Fujitsu Research and Development Center- China*  
Liu Wang, *Fujitsu R&D Center Co Ltd - China*  
Lluis Gomez, *Computer Vision Center - Spain*  
Malayappan Shridhar, *University of Michigan-Dearborn - USA*  
Markus Wienecke, *Siemens AG - Germany*  
Mickaël Coustaty, *L3i - University of La Rochelle - France*  
Mingming Zhang, *Fujitsu R&D Center Co., Ltd. - China*  
Mitra Mohtarami, *National University of Singapore - Singapore*  
Nathalie Girard, *LI - University of Tours - France*  
Nibal Nayef, *University of La Rochelle - France*  
Nicholas Journet, *LaBRI - University of Bordeaux - France*  
Nicolas Sidère, *LI - University of Tours - France*  
Oriol Ramos Terrades, *Computer Vision Center - Spain*  
Pan Pan, *Fujitsu R&D Center Co., Ltd. - China*  
Panduranga Nagabhushan, *Indian Statistical Institute - India*  
Partha Roy, *Synchromedia Lab - Canada*  
Peng Ye, *University of Maryland College Park - USA*  
Petra Gomez-Krämer, *University of La Rochelle - France*  
Qi Chen, *NUS - Singapore*  
Rainer Lindwurm, *Siemens AG - Germany*  
Rajiv Jain, *University of Maryland - USA*  
Rohit Pandey, *University at Buffalo - USA*  
Roland Zimbel, *Siemens AG - Germany*  
Rong Huang, *Kyushu University - Japan*  
Saddok Kebairi, *ITESOFT - France*  
Samy Bengio, *Google - USA*  
Sergey Grosman, *Siemens - Germany*  
Shijian Lu, *Institute for Infocomm Research - Singapore*  
Shounak Gore, *University at Buffalo - USA*  
Shufu Xie, *Fujitsu R&D Center Co., Ltd. - China*  
Slim Kanoun, *MIRACL Laboratory, University of Sfax - Tunisia*  
Song Wang, *Fujitsu - China*  
Soohyung Kim, *Chonnam National University, Korea*  
Stefan Pletschacher, *University of Salford - UK*  
Sukalpa Chanda, *Indian Statistical Institute - India*  
Syed Saqib Bukhari, *Insiders Technologies - Germany*  
Tetsushi Wakabayashi, *Mie University - Japan*  
The Anh Pham, *LI - University of Tours - France*



Tomo Miyazaki, *Tohoku University - Japan*  
Trung Quy Phan, *NUS - Singapore*  
Utkarsh Porwal, *University at Buffalo - USA*  
Vlad Atanasiu, *University of Fribourg - Switzerland*  
Volkmar Frinken, *Kyushu University - Japan*  
Wael Abd-Almageed, *University of Maryland - USA*  
Wataru Ohyama, *Mie University - Japan*  
Wei Fan, *Fujitsu R&D Center Co., Ltd. - China*  
Wei Liu, *Fujitsu R&D Center Co Ltd - China*  
Xi Zhang, *NUS - Singapore*  
Xiang-Dong Zhou, *CIGI T - Chinese Academy of Sciences - China*  
Xu-Yao Zhang, *IA - Chinese Academy of Sciences - China*  
Yuan He, *Fujitsu R&D Center Co Ltd - China*  
Yun Zheng, *Fujitsu R&D Center Co., Ltd. - China*  
Zhen Lei, *Chinese Academy of Sciences - China*



## Message from the General Chairs and Program Chairs

Our heartiest welcome to DAS 2014, the 11th IAPR International Workshop on Document Analysis Systems being held in Tours - Loire Valley, France. With this eleventh edition, the workshop is held for the first time in France after successful workshops in Kaiserslautern, Germany (1994); Malvern, PA, USA (1996); Nagano, Japan (1998); Rio de Janeiro, Brazil (2000); Princeton, NJ, USA (2002); Florence, Italy (2004); Nelson, New Zealand (2006), Nara, Japan (2008), Boston, USA (2010) and Gold Coast, Australia (2012).

DAS 2014 continues well established standards and introduces novel ideas. It is a single-track peer-reviewed, 100% participation conference and it attempts to bring together industrialists and academics, as well as practitioners and theoreticians from numerous related disciplines involved in document analysis systems research and to provide opportunities for interactions between them. For the first time, an Industrial Program Committee takes part to the Workshop, which is composed of researchers coming from companies, who are very active in the field and frequently participated at previous DAS Workshops. As such DAS 2014 emphasizes the systems aspect which is already in its title.

On behalf of the organizing committee, we are happy to announce that we received 138 submissions from researchers of 32 countries around the world. The Program Committee Chairs invited 45 international reviewers (including the program committee members) to review the papers. All papers have been refereed by at least three reviewers (2 academic researchers from the Program Committee and one researcher from this Industrial Program Committee). 132 papers were reviewed by three reviewers and the other 6 papers were reviewed by four reviewers. Finally, 73 long papers were accepted, of which 27 are for oral presentation and 46 are for poster presentation. As such, the acceptance rate for oral papers is 19.6%.

These accepted papers cover diverse areas of preprocessing, feature extraction, segmentation, recognition, signature verification, text classification, image retrieval techniques, video document processing, document image decoding, graphical document processing, performance evaluation, historical and handwriting documents, different systems on document analysis etc. The final program consists of seven oral sessions, two poster sessions and one discussion session.

For the first time we have invited three distinguished keynote speakers: Prof Andreas DENGEL (DFKI Kaiserslautern - Germany), Vladimir Rybkin (Head of Character Recognition and Image Processing Group - ABBYY) and Vincent Poulain D'Andecy (Design and management of Document Analysis Systems - ITESOFT) have accepted our invitation to deliver a keynote talk at the workshop. We thank them sincerely for accepting our invitation to deliver the keynotes.

We would also like to express our sincere thanks to Ray Smith from Google, USA; to Professor C. V. Jawahar from IIIT Hyderabad, India; and to Dr. Pramod Kompalli from Xerox Research Centre, India for their very informative tutorials.

At this point we thank all the researchers who showed interest in this DAS by sending contributed papers. Thanks are also due to all chairs of various activities, program committee members, reviewers, and local organizing committee members including the Computer Science Laboratory of Tours (LI - EA6300) at the University of Tours for their strong support and active participation. The University of Tours, the city of Tours as well as the Region Centre have extended their support in organizing the workshop to a great extent. We sincerely thank all of them for their kind help.

Last but not the least; we would like to extend a special thanks to our valued sponsors of the workshop. We hope you will find your stay fruitful and rewarding. We trust that you will enjoy the exchange of technical and scientific ideas during the three days of DAS 2014 as well as getting a flavour of the city of Tours and the Loire Valley, which are one of the most famous and most beautiful tourist destinations in France. We extend our warmest welcome to you, and hope that your visit will be a memorable one!

Jean-Yves Ramel and Marcus Liwicki

**DAS 2014 General Chairs**

Jean-Marc Ogier, Koichi Kise, and Ray Smith

**DAS 2014 Program Chairs**

# Table of contents

A Hybrid Algorithm for Automatic Language Detection on Web and Text Documents. Luciano Cabral, Rinaldo Lima, Rafael Lins, Fred Freitas, Rafael Mello, Gabriel França and George D. C. Cavalcanti	3
A system for camera-based complex map image retrieval using a multi-layer approach. Quoc Bao Dang, Muhammad Muzzamil Luqman, Mickael Coustaty, Nibal Nayef, Jean-Marc Ogier and Cao De Tran	5
NOE-EDHI: A trans-disciplinary project to create a platform for demographic forms retro-conversion. Mickaël Coustaty, Alain Bouju, Pascal Chareille, Jean-Marc Ogier, Arnaud Bringé, Isabelle Seguy and Pierre Darlu	7
Categorizing a New Collection for Document Recognition Research. Barri Bruno and Daniel Lopresti	9
A simple approach to distinguish between Maghrebi and Persian calligraphy in old manuscripts. Insaf Setitra and Abdelkrim Meziane	11
Boosting-based approaches for Arabic text detection in news videos. Sonia Yousfi, Sid-Ahmed Berrani and Christophe Garcia	13
An Android Application for Browsing and Searching in Historical Manuscripts. Alicia Fornés, Pau Riba and Josep Lladós	15
PaRADIIT Project: Main Concepts and Outcomes. Frédéric Rayar and Jean-Yves Ramel	17
Efficient OCR Training Data Generation with Aletheia. Christian Clausner, Stefan Pletschacher and Apostolos Antonacopoulos	19
Dehyphenation by Classification for OCR Results. Mayce Al-Azawi and Thomas M Breuel	21



# A Hybrid Algorithm for Automatic Language Detection on Web and Text Documents

Luciano Cabral<sup>a</sup>, Rinaldo Lima<sup>a</sup>, Rafael Lins<sup>a</sup>, Fred Freitas<sup>a</sup>, Rafael Ferreira<sup>a</sup>, Gabriel Silva<sup>a</sup>,  
George Cavalcanti<sup>a</sup>, Steven Simske<sup>b</sup> and Marcelo Riss<sup>c</sup>

<sup>a</sup> Informatics Center, Federal University of Pernambuco, Recife, Brazil

<sup>b</sup> Hewlett-Packard Labs., Fort Collins, CO 80528, USA

<sup>c</sup> Hewlett-Packard Brazil, Barueri, Brazil

{lsc4,rjl4,rdl,fred,rflm,gfps,gdcc}@cin.ufpe.br, {steven.simske,marcelo.riss}@hp.com

**Abstract** — Automatic Language Detection is a research area that has gained importance with the Internet and plays a key role in information retrieval. This paper presents a hybrid algorithm to automatic language detection. Our approach relies on classical techniques such as n-gram text analysis, relative frequency and dictionaries of closed-class words. The proposed method is very fast and accurate if compare with its competitors.

**Keywords**—language detection; language identification; document engineering; comparative analysis; assessing techniques

## I. INTRODUCTION

The task of automatic language detection has recently emerged as of crucial importance because web search engines have to collect and show multilingual content to the user. Language independent search, well performed by Google, is another good example of application of specific algorithms for automatic language identification. Automatic language identification can be used as a first step towards automatic translation, which certainly increases its usability on the Web.

Some survey papers, such as [1] and [2], address this problem reaching different conclusions about the efficiency of the methods currently available. Although the initial steps in this research area date back to the mid 1960s, there are still open questions and the “birth” of the Internet has brought new challenges to the field.

This work attempts to answer some questions focusing on the language detection problem. More particularly, we want to evaluate a hybrid method called CALIM, which combines three different techniques (Dunning [3], Cavnar and Trenkle [4] and Lins and Gonçalves [5]) that gained new support to web documents. In addition, we implemented other three classical methods in the same language (Java) in order to have a fair and independent performance evaluation.

## II. THE CALIM ALGORITHM

CALIM is based on language profile dictionaries that take into account frequent short words present in all languages under study (21 in total). More precisely, the creation of such dictionaries takes into account approximately the 250 more frequent words for each language. For creating such language profiles, we used some dictionary databases provided by Lexiteria [6], which is an initiative aiming at understanding various aspects of human language.

For our research purposes, we collected some statistics for 21 languages, such as word frequency, average word length,

etc. After creating the language dictionaries, sorted by word frequency in decreasing order, we selected the top frequent 250 words.

The underlying assumption in CALIM is that the selected high frequency words are more likely to be found in the input text than low frequency ones. In addition, due to performance reasons, we only consider words that have maximum length 5 characters ( $n = 5$ ). We justify this choice because, in most languages, the most frequent words like prepositions, personal pronouns, etc. are also the shorter ones.

During the classification step of CALIM, we applied a heuristic that contributed to more accurate results in our performance evaluation. That heuristic assumes that if a word (or token) comprises very specific n-grams exclusively found in certain language (like “ão” in the Portuguese language), then the method assigns a greater value to it than it is done in normal voting schema in which the normal vote is equal to 1.

The normalized frequency of a token is simply calculated by the ratio between the token frequency and the sum of all token frequencies for a language dictionary. This simple heuristics seemed to contribute to determine the correct language of the document.

### A. Other Implementations

In addition to the CALIM algorithm, we decided to develop the *simple closed-class dictionary* method proposed by Lins and Gonçalves [5] (such method is here labelled *CALIG*), and the Language Detector [7], [3], labelled here as *LangDetect*.

- *SimpleDic*: It is a simple dictionary method, developed using as the main component, the common stop words of, commonly ignored in some Natural Language Processing (NLP) procedures. Those stop words were useful in the dictionaries formation as they are short, simple, and most of times invariant to gender or quantity (singular or plural).
- *LangDetect*: Language Detector consists of a library for language identification developed in Java. It was implemented in [7] by Shuyo, based on the techniques proposed by Dunning [3].
- *CALIG*: We revisited the work of Lins and Gonçalves [5] and used their methodology to develop new dictionaries of closed classes from the 6 original ones to 25 languages in total, using two layers of treatment. In the first layer, the lexical analyser recognizes the most common lexical features, i.e. “ñ” for Spanish, “ß” for German, “æ” for Danish, etc. The second layer takes into account the closed class dictionaries and a decision

The research results reported in this paper have been partly funded by a R&D project between Hewlett-Packard Brazil and UFPE originated from tax exemption (IPI-Law n 8.248, of 1991 and later updates).

tree heuristic to choose the best language using the ratio of the total number of tokens in the document divided by the number of recognized tokens in the document.

As already mentioned, the original paper by Lins and Gonçalves [5] covers only 6 languages, thus we had to extend it following the method originally described in [5] to cover all the 21 European languages plus four other languages with lexical features, e.g. Arabic, Hebrew, Hindi and Korean.

### III. EXPERIMENTAL SETUP

The test bed used here is an extension of the one described in reference [2]. The new test environment was adapted to recognize the format of files from the web, i.e., HTML and XML, removing tags and annotations from them, leaving for analysis only what really matters for the scope: the text part.

#### A. Environment

In our experiments, we used a laptop equipped with a processor Intel Core i3-2330M 2.20 Ghz, 4 GB RAM, and the Microsoft Windows 7 operating system. We chose Java as development language and accessing serialized files.

#### B. Test Corpora

We used the Europarl v7 corpus [8] that has two versions, the first one called “*test*” which is composed of 21,000 documents containing very small size texts. This dataset presents an equal distribution between documents and languages supported (1,000 documents per language). The second one, called “*full*”, with about 60,000 some size XML documents, with a random distribution between documents and languages (Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish).

#### C. Results and discussion

We processed the *Europarl* Test Corpus, and observed similar results compare to the ones in [2]. We have performed two test set ups, one with the 21 European languages (already mentioned), and another with the 6 most common web languages (English, French, German, Italian, Portuguese and Spanish). Table I summarizes the obtained results.

TABLE I. AVERAGE RESULTS FOR 6 AND 21 LANGUAGES (TEST)

Methods	6 languages		21 languages	
	Acc	Time(s)	Acc	Time(s)
SimpleDic	94.72	3.65	63.23	12.79
CALIM	98.93	2.87	97.08	10.06
LangDetect	99.48	3.25	99.16	11.38
CALIG	92.13	12.78	92.42	44.79

According to Table I, the best performance was obtained on the *Test Corpus* in which *LangDetect* (more accurate, with 0.55 and 2.08 percent higher than *CALIM* on 6 and 21 languages respectively), and *CALIM* (faster, with 0.38 and 1.32s smaller than *LangDetect* to 6 and 21 languages respectively).

TABLE II. AVERAGE RESULTS FOR 6 AND 21 LANGUAGES (FULL)

Methods	6 languages		21 languages	
	Acc	Time(s)	Acc	Time(s)
SimpleDic	100.00	871.76	80.123	3051.16
CALIM	100.00	793.37	99.992	2776.79
LangDetect	100.00	910.32	99.993	3186.12
CALIG	99.97	676.97	99.942	2369.40

Analyzing Table II, the result with the Full Corpus, that is more appropriate to the reality of any web document, the *CALIG* strategy was faster, with shorter processing time than *CALIM* and *LangDetect* (which had similar accuracy) in 14.67% and 25.63%, respectively, above the best time.

In addition, the accuracy for 6 and 21 languages in just 0.03 and 0.051 percentage points below the bests (*CALIM* and *LangDetect*). Thus, the results suggest that, in terms of accuracy and processing time, *CALIM* and *CALIG* strategies obtained the best performance on the *Europarl Test* and *Full Corpora* datasets, with less processing time, and higher accuracy.

### IV. CONCLUSION AND FUTURE WORK

This paper presented an assessment of some automatic language detection techniques, and most importantly, a hybrid algorithm inspired on the ideas of Dunning [3], Cavnar and Trenkle [4] and Lins and Gonçalves [5]. The importance of the reported analysis rests on the fact that all the referenced algorithms were implemented in the same hardware and software platform, and assessed on the *Europarl* corpus at different versions allowing a fair comparison amongst them.

Two of the algorithms analyzed on the Test Corpus (Plain text) experiment, namely *LangDetect* and *CALIM*, have shown competitive performance not only due to the higher accuracy, but also faster processing time. The former was more accurate, the latter was faster. On the Full corpus experiment, the aforementioned algorithms appear with the better results in terms of accuracy, with virtually the same performance score on this second dataset, and *CALIM* still running faster than *LangDetect*.

The *CALIG* algorithm strategy appears as faster in the Full corpus experiment. This result is interesting, because it has a larger quantity of dictionaries implemented (133 hash table dictionaries). The other strategies not exceed 53 dictionaries. For future work, we plan to: (i) solve common problems, such as multilingual document classification; (ii) integrate some summarization strategies aiming at language independent summarization tasks.

### REFERENCES

- [1] B. Hughes, T. Baldwin, S. Bird, J. Nicholson and A. Mackinlay, "Reconsidering language identification for written language resources," *Proceedings of LREC 2006*, pp. 485-488, 2006.
- [2] L. Cabral, R. Lins, R. Lima and S. Simske, "A Comparative Assessment of Language Identification Approaches in Textual Documents.," *Proceedings of IADIS Applied Computing 2012*, July 2012.
- [3] T. Dunning, "Statistical identification of language," Technical Report CRL MCCS-94-273, Computer Research Lab, New Mexico University, New Mexico, 1994.
- [4] W. B. Cavnar and J. M. Trenkle, "N-Gram Based Text Categorization.," *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161-169, 1994.
- [5] R. Lins and P. Gonçalves, "Automatic language identification of written texts.," *Proceedings of the ACM Symposium on Applied Computing (SAC'04)*, 2004.
- [6] Lexiteria, "Word Frequency Lists.," Lexiteria, 2002. [Online]. Available: <http://www.lexiteria.com/>. [Accessed 09 10 2013].
- [7] N. Shuyo, "Language Detection Library for Java," 2010. [Online]. Available: <http://code.google.com/p/language-detection/>. [Accessed 8 October 2013].
- [8] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," *MT Summit 2005*, 2005.



# A system for camera-based complex map image retrieval using a multi-layer approach

Q.B. Dang<sup>(1)</sup>, M.M. Luqman<sup>(1)</sup>, M. Coustaty<sup>(1)</sup>, N. Nayef<sup>(1)</sup>, C.D. Tran<sup>(2)</sup>, J.M. Ogier<sup>(1)</sup>

<sup>(1)</sup>L3i Laboratory, University of La Rochelle, France

<sup>(2)</sup>College of Information and Communication Technology, Can Tho University, Vietnam  
quoc\_bao.dang@univ-lr.fr

**Abstract**—In this paper, we present a method of camera-based document image retrieval for heterogeneous-content document using a multi-layer separating approach. We use Locally Likely Arrangement Hashing (LLAH) and PCA-SIFT for extracting features, and we show a method how to use one hash table for indexing multiple types of feature vectors. A technique of reducing the memory required for storing the hash table is also employed.

**Keywords**—camera-based document image retrieval, automatic indexing, text/graphic separation, feature extraction.

## I. INTRODUCTION AND RELATED WORK

Camera-based document image retrieval is a task of searching document images relevant to user's query that is captured by a digital camera. Recently the method called Locally Likely Arrangement Hashing (LLAH) has been known as the efficient and real-time camera-based document image retrieval methods. It is based on local combination of affine invariant calculated from feature points which are extracted from centroid of each word connected component [2]. Because of using local combination of centroids of words, accuracy of retrieval will reduce when number of word is small when it is applied to rich graphics document on one side. On the other side, graphics can be extracted by using other dedicated features such as SIFT, PCA-SIFT, SURF and shape features [1].

Text/graphics separation is a process segmenting a document image into two layers, one containing text and the other containing graphics. From several decades, many methods have been proposed to solve this problem. Color-based has been used for separating an image into many layers. Furthermore, connected component (CC) analysis was widely used for this separation. For instance, Karl Tombre et al. [5] proposed a size-histogram analysis from the bounding boxes of all CCs. By a correct threshold selection obtained dynamically from the histogram, the large graphical components are discarded, and the smaller graphics and text components are kept. In order to separate text/graphics, Winfried Höhn [6] used density of CC that is a ratio between the area of the convex hull and the number of pixels in CC. To remove the dashed and dotted lines, the CCs are filtered by their density if the density is lower than a threshold. Further, they used diameter ratio that is the ratio between minimum diameter of CC and maximum diameter of CC. They also used a combined threshold region for the density and the ratio of

maximum and minimum diameter, extended by an analysis of neighboring components to recognize text with large variations in style, size and orientations.

In this paper, we propose a method of multi-layer separating for camera-based document analysis and retrieval of complex map images which are composed of heterogeneous-content. Our method aims to separate document image into multiple layers using attributes of CC and color attributes in order to extract features adapted to the content of each layer. The result of retrieval relies on majority vote over all retrieval results of all layers from query image.

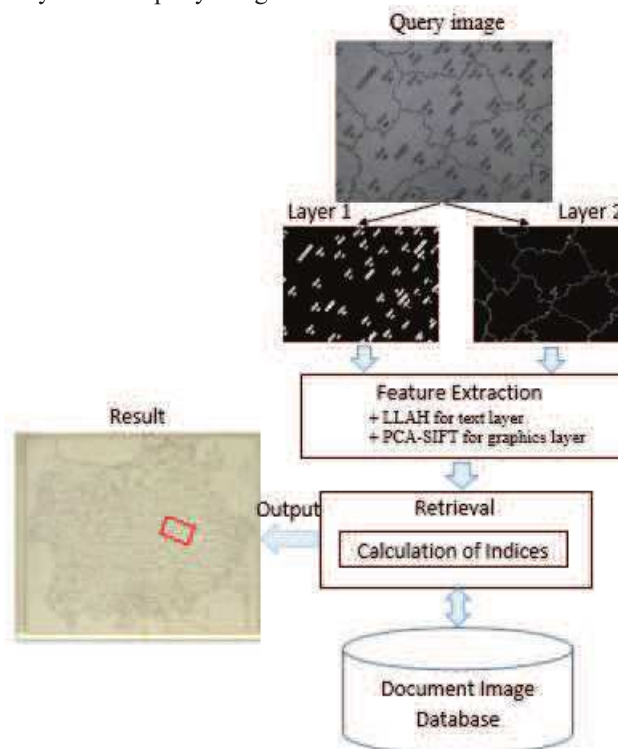


Figure 1: Retrieval phase.

## II. METHOD

In this section, we describe our system for camera-based complex map image retrieval using a multi-layer approach.

### A. Multi-layer separating

Our method is outlined in Figure 1. In both indexing phase and retrieval phase, document image is separated into many layers. We use attributes of CC, for instance, CC area, CC

perimeter, CC density, and CC diameter ratio so that CCs can be classified into several types from which the layer separation is done. Besides, the color-based can be used to separate the image into different color layers using clustering approaches (e.g. K-means and hierarchical clustering). Yet to define the thresholds to classify CCs and to choose a color-based approach will depend on the kind of document and the number of layers being separated.

### B. Feature extraction

We extract feature vector of LLAH with text layers, which can be obtained even under perspective distortion, noise, and low resolution. With graphics layers, we extract PCA-SIFT feature vector because PCA-SIFT is known as a robust and affine invariant feature. What is more important that the number of dimensions of feature vector is 36, that is effective in term of computation and indexing [4].

For LLAH, we use affine invariant as a discrete value of the invariant that is defined using 4 coplanar points, so LLAH vector is calculated from the arrangement of  $m$  points that is described as a sequence of invariants calculated from all possible combinations of 4 points taken from  $m$  points [2]. We also add additional feature using rank of area ratios of word regions at each  $m$  points [3] so that it can deal with fewer words case.

### C. Indexing phase

To add a new document ID into hash table, the existent database does not need to be reconstructed or recomputed [2, 4]. The document ID is separated into many layers firstly. Then for each layer, document feature vectors are extracted and indexed. Aiming to reduce the required amount of memory, we do not store feature vectors in hash table [2, 4]. We just store point ID where feature vector is extracted. Because of using one hash table for indexing both LLAH feature vector and PCA-SIFT feature vector, feature type is stored with point ID. So as to index, the real-valued feature vectors need to be converted into integers.

For LLAH vector  $r$ , we just normalize it similar to [2]. For PCA-SIFT vector  $r$ , we binarize each dimension by producing a bit vector  $u=(u_1, u_2, \dots, u_d)$  ( $d \leq 36$  and  $d$  is the first  $d$  value of  $r$ ) by the following formulation [4]:

$$u_i = \begin{cases} 1 & \text{if } r_i \leq 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The index  $H_{\text{index}}$  of the hash table is calculated by the following hash function [2, 4]:

$$H_{\text{index}} = \left( \sum_{i=0}^{n-1} r_{(i)} k^i \right) \bmod H_{\text{size}} \quad (2)$$

Where  $n$  is the number of dimension of  $r$ ,  $k$  is the level of quantization (e.g.  $k=15$  for normalized LLAH vector and  $k=2$  for bit vector produced from PCA-SIFT vector),  $H_{\text{size}}$  is the size of hash table.

### D. Retrieval phase

Before searching in the hash table, feature vector is converted into integral vector. Furthermore, the PCA-SIFT query vector need to be expanded into several bit vectors so that nearest neighbor can be found using the strategy in [4].

For each layer separated from captured query image, query feature vectors are extracted firstly. Next, for each feature vector, it is used for searching in hash table and voting for document ID containing it.

Since the matching between the query feature vector and document feature vectors is skipped, some voting may be wrong [4]. To ease the problem, we employ the RANSAC algorithm to find a perspective transformation  $T$  between set of query point IDs and set of document point IDs being voted, for each only document having a number of votes larger than a threshold. Then we use transformation  $T$  to remove the wrong voting if distance between point ID being voted and point being applied transformation  $T$  is less than an acceptable threshold [7]. Finally, the document with the largest number of votes is returned as the retrieval result, the transformation  $T$  is also used for spotting region of interest.

## III. EXPERIMENTATION

The experimentation is performed on a dataset of the linguistic map of France. There are 12 images of resolution 9800 x 11768 pixels in the dataset. For online retrieval phase, Samsung document camera SDP-760 is used, Figure 2. A complete sophisticated demo of the system will be shown during the workshop.



**Figure 2:** Camera.

## IV. CONCLUSION

We have presented our initial work on a multi-layer information spotting system for camera-based heterogeneous content document image retrieval. Work is in progress to extend our system to multi-layer ( $>2$ ) for automatic indexing and retrieval of scanned newspapers.

## REFERENCES

- [1] Marinai Simone, Beatrice Miotti, and Giovanni Soda. "Digital Libraries and Document Image Analysis Techniques: a Survey."
- [2] Nakai Tomohiro, Koichi Kise, and Masakazu Iwamura. "Camera based document image retrieval with more time and memory efficient LLAH." *Proc. CBDAR* (2007): 21-28.
- [3] Takeda Kazutaka, Koichi Kise, and Masakazu Iwamura. "Memory reduction for real-time document image retrieval with a 20 million pages database." In *Proceedings of the 4th International Workshop on CBDAR*. 2011.
- [4] Koichi Kise, Kazuto Noguchi, and Masakazu Iwamura. "Simple Representation and Approximate Search of Feature Vectors for Large-Scale Object Recognition." In *BMVC*, pp. 1-10. 2007.
- [5] Tombre Karl, Salvatore Tabbone, Loïc Pélissier, Bart Lamiroy, and Philippe Dosch. "Text/graphics separation revisited." In *DAS*, pp. 200-211. Springer Berlin Heidelberg, 2002.
- [6] Höhn Winfried. "Detecting Arbitrarily Oriented Text Labels in Early Maps." In *Pattern Recognition and Image Analysis*, pp. 424-432. Springer Berlin Heidelberg, 2013.
- [7] Viet Phuong Le, Muriel Visani, Cao De Tran, and Jean-Marc Ogier. "Improving Logo Spotting and Matching for Document Categorization by a Post-Filter based on Homography." In *ICDAR, 2013 12th International Conference on*, pp. 270-274. IEEE, 2013.

# NOE-EDHI: A trans-disciplinary project to create a platform for demographic forms retro-conversion

M. Coustaty<sup>(1)</sup>, Alain Bouju<sup>(1)</sup>, Pascal Chareille<sup>(2)</sup>, J.M. Ogier<sup>(1)</sup>, Arnaud Bringé<sup>(3)</sup>, Isabelle Seguy<sup>(3)</sup>, Pierre Darlu<sup>(4)</sup>

<sup>(1)</sup>L3i Laboratory, University of La Rochelle, France | <sup>(2)</sup>Centre d'Études Supérieures de la Renaissance, Tours, France

<sup>(3)</sup>Institut National d'Études Démographiques, Paris, France | <sup>(4)</sup>Museum National d'Histoire Naturelle, Paris, France

mcoustat@univ-lr.fr

**Abstract**—This paper presents the preliminary results of a trans-disciplinary project that aims at creating a platform to extract and automatically recognize the content of historical forms related to the French population from the 16<sup>th</sup> to the 19<sup>th</sup> centuries, in the context of a very large study led by the French Institute of demographic studies (INED). These contents correspond to thousands of handwritten forms that need to be automatically processed using HWR and wordspotting techniques. These analysis will allow to create symbolic links between the data and the generation of a final map. This work presents the first results of the automatic data extraction and recognition and gives an overview of the global platform that will be developed in the project.

**Keywords**—HWR, ICR, Historical documents, Word-spotting, Layout analysis

## I. CONTEXT AND RELATED WORK

The French National Institute of Demographic Studies (INED) led two surveys (in 1958 and 1982) to study the French population between the 16<sup>th</sup> and the 19<sup>th</sup> centuries. These surveys produced thousands of forms to collect vital events of French population. Some examples of these forms can be observed in Figure 1.

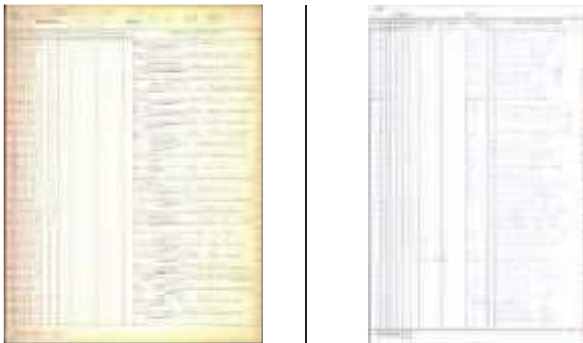


Figure 1: Examples of forms used to collect data

Supported by the French CNRS and by the INED, this project aims to study the automatic extraction and recognition of historical forms in one hand, and to interpret extracted elements in order to add semantic on the other hand.

These forms gather information related to the life of people. Generally, the left part of the sheet (col. 1 to 10) collects non-nominal data (date, type of civil status record, sex, legitimacy of children, previous matrimonial status of the spouses or of the deceased, age, place of birth and place of residence of the spouses or of the deceased, occupation, presence or lack of

signature), while the right side (col. 11 or 12) is restricted to the nominal information (name and surname of newborn, spouses or deceased, those of their parents, etc.). As a consequence, one important element to take into account is that a person could appear several times in the global database (for her/his birth, her/his first and following marriage for her/his death) and often, many more times (as parent at the birth of his/her children).

The NOE-EDHI Project aims to develop a platform to automatically process these forms by extracting their content and recognizing it. We present it in this paper.

## II. METHOD

In this section, we describe the global platform that has been used for the automatic extraction and recognition of the forms content. This step has been split in two parts: firstly the table content's extraction, and then the recognition part.

Finally, in the last part, we also present the works in progress to create some links between two registrations and then add semantic in the global database.

### A. Table content extraction

The forms used in these campaigns rely on tables, which are quite well structured and not too noisy. In order to extract the content of these tables, we focused on the structure of the table, composed of lines and columns.

First we binarize the image using the Otsu binarization process [1], and then we retain the widest connected components. As forms are composed of horizontal and vertical lines, we used the Hough transform [2,3] to search and to detect the main orientation of the document. With this transformation, we search in a parametric space for lines with orientations in the following sub-spaces ( $-10^\circ, +10^\circ$ ) ( $80^\circ, 100^\circ$ ), to be robust to the noise of the digitization process (images can be a little bit rotated).

The method used presents two main advantages. First, it relies on the Hough transform that is famous for its robustness towards degradations and noise. Secondly, the proposed method gives information on the mean orientation of tables in the image. Moreover, the use of a parametric space also allows detecting hashed lines as they correspond to small components around the mean orientation ( $\pm 5^\circ$ ).

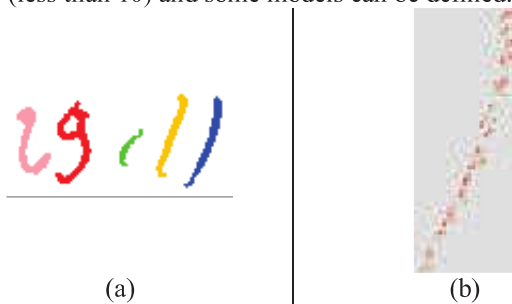
Finally, the set of detected lines gives the global structuration of the document in lines and columns.

**B. Content recognition**

In a first step, the project development has been done on dates and gender recognition. These two columns have been chosen for their interest from the demographic point of view, and also because they correspond to a limited number of characters to recognize.

1) *Date recognition*

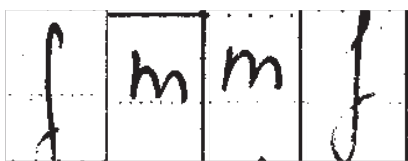
The first column of the forms contains the date of the registration. As they correspond to handwritten numbers, classical OCR techniques (Finereader, Tesseract, ...) do not give good results. Based on the work done in [4], we thus define a new number recognition technique based on an algorithm which enables us to match an image with a prototype. This is achieved by minimizing an energy function using similar ideas to the "elastic net". With this method one wishes to associate two images trying to keep unchanged the distance between the points of the prototype. In this case, the global idea is to summarize each number path by a subset of points, and to try to match this set with the model sets defined earlier. This method was chosen as the number of scripters is low (less than 10) and some models can be defined.



**Figure 2: Example of image of date (a), and the path of the yellow and blue number (b) to recognize**

2) *Gender recognition*

The third column of the forms is interesting because it allows checking the validity of the future process (family names detection and recognition). This information is interesting for researcher in demography as they can study the evolution of the population. Moreover, as first names are linked to the gender in France, this can be used in the word-spotting validations technique.



**Figure 3: Examples of gender information**

Gender cell can contain an m for "male" or an f for "female". As one can see in Figure 3, this can be seen as a binary classification problem, and the letter "f" is quite taller than the letter "m". Using a projection histogram technique of the black point on the horizontal axis, we are able to easily separate the two classes and to identify the gender of the registration.

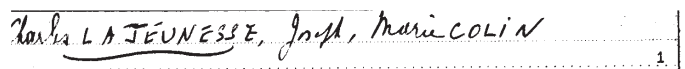
**C. Future works in progress**

The global platform of the project is summarized in Figure 4. This one will be based on the pre-processing techniques presented in the previous sections, and will also propose two features: word-spotting dedicated to family names and semantic enhancement by linking two registrations related to the same person.



**Figure 4: Global overview of the system**

For the word-spotting technique, some pre-process have been done to only retain the family names in their image format. Starting from the cell hereafter:



We binarize it, we extract the connected components and we only retain the capital letters, based on their size and position. Here comes an example of result obtained:



The next step, actually in development, will use some word-spotting techniques to retrieve two similar words in the database.

**III. EXPERIMENTATIONS**

In order to validate all the steps presented before, some experiments have been performed on a data-set of demographic forms from two French cities (Vic-sur Seilles, 1591-1862; Echeveronnes 1617-1856). 10 images of forms have been used and a complete sophisticated demonstration of the system will be shown during the workshop.

**IV. CONCLUSION**

We have presented our initial work on the NOE-EDHI project. This paper presents the first works related to a platform developed to automatically extract and recognize handwritten text in forms in one hand, and to create links between the registrations related to a same person. Mixing up various approaches, we are able to extract the content and to recognize it. The links will be integrated once the word-spotting technique will be effective, and we will present a demonstration of this platform at the conference.

**REFERENCES**

- [1] Otsu, N., A threshold selection method from Gray-Level histograms. IEEE Trans. Syst. Man Cybern. 9(1), 62-66 (1979)
- [2] Duda, R. O. and P. E. Hart, "Use of the Hough Transformation to Detect Lines and Curves in Pictures", Comm. ACM, Vol. 15, pp. 11-15 (January, 1972)
- [3] Coustaty M., Bertet K., Visani M., Ogier J.M., A New Adaptive Structural Signature for Symbol Recognition by Using a Galois Lattice as a Classifier. IEEE SMC-Part B 41(4): 1136-1148 (2011).
- [4] Bouju A. Labelling by adaptative mapping. In Neuro-Nime 1993, p291-298, France (1993)

# Categorizing a New Collection for Document Recognition Research

Barri Bruno

Department of Computer Science and Engineering  
Lehigh University  
Bethlehem, Pennsylvania  
Barri.Bruno@gmail.com

Daniel Lopresti

Department of Computer Science and Engineering  
Lehigh University  
Bethlehem, Pennsylvania  
lopresti@cse.lehigh.edu

**Abstract**—In this paper we present initial categorization and informative statistics for the Lehigh Steel Collection (LSC). The LSC is a new open dataset we have been assembling for document recognition research. The pages within the dataset represent the last few decades of research activities for the bankrupt steel giant Bethlehem Steel Corporation. Due to its potentially massive size and varied nature, we see major potential for this set of documents and have been working to make it available to the research community for non-commercial purposes.

**Keywords**—Document Analysis; Dataset

## I. INTRODUCTION

For over 100 years, the Bethlehem Steel Corporation was considered a giant in the steel-making industry. Founded in 1857 and located in Bethlehem, Pennsylvania, it was once the largest ship builder and second largest steel producer in the United States. After production business boomed during the First and Second World Wars, the Corporation invested \$25 million to build a state of the art research facility on South Mountain in Bethlehem to be called Homer Research Labs [1].

In its prime, the facility employed over 1,000 researchers and technicians, but by the 1980's, foreign competition and poor business planning began to tear the company apart. In 1986, Bethlehem Steel was forced to sell five of the eight buildings in the facility to nearby Lehigh University which converted the space into research laboratories and academic buildings [2].

In 1995, the steel manufacturing plants in Bethlehem had to be closed, but Homer Labs remained open, now employing about 75 researchers [3]. In 2001, the Corporation filed for bankruptcy. It was purchased in 2003 by the International Steel Group (ISG) and two years later, Mittal Steel Co. bought the Homer Lab facilities. One more building was sold to Lehigh University and the other two were shut down for good [4].

In May of 2013, Lehigh University purchased the remaining two buildings and opened a large work bay for educational endeavors. The building offices, however, remained untouched and contained hundreds of thousands of documents from Bethlehem Steel's last few decades. The Lehigh Steel Collection (LSC) is a fraction of these papers that

we have been working to digitize and organize for non-commercial, research purposes.

## II. OUTLINE OF PROCEDURE

### A. Collection

We have already completed the collection phase of our work. Collection took place between June and September 2013 when Lehigh University opened the two newly purchased buildings for research purposes. Documents were gathered and labeled by room number and any other defining characteristics such as cabinet numbers and drawer headings. Photographs were taken of all offices and collection sites.

### B. Digitization and Organization

We are currently working to digitize all of the collected documents. As pages are scanned, the document organizational structure is recorded and associated with the file name and location. Presently, the digitized collection consists of roughly 30,000 documents [5]. Sample pages from the collection can be seen in Fig. 1.

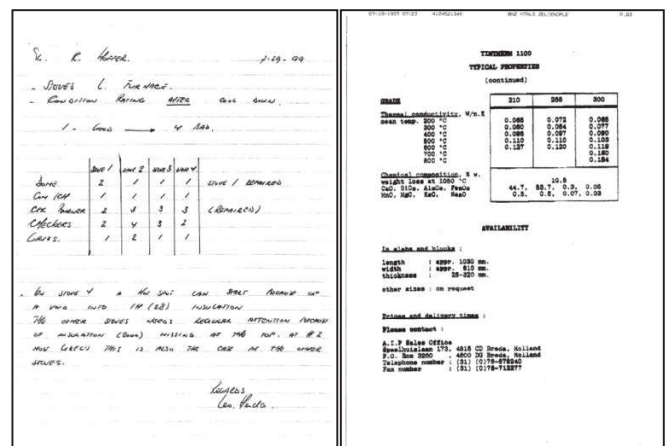


Fig. 1. Two collected pages from the LSC containing dates, rulings, machine print, handwriting, tables, and signatures.

### C. Categorization

Categorization of the digitized collection is presently our area of interest. We have adopted a two-pass approach to

categorizing the documents. The first pass occurs when documents are scanned. Because the scanning and organization procedures require concentrated attention, this pass simply allows the researcher to mark down any “flags” on a document. This most often suggests a “confidential” marking or some personally identifiable information to be checked later.

The second pass occurs after the documents have been digitized and is much more thorough. With the help of a collection interface, a researcher goes through the documents one by one and notes the attributes of each page separately. We have been working to automate this process, but at the moment, it is done by hand. The attribute values are collected in a spreadsheet for analysis and reference. Initial categorization statistics can be found in Section 3.

#### D. Release

As we make progress in the collection, digitization, and categorization of the LSC, the proper release of the data set is becoming a more prominent interest. Through work with Mittal Steel Co. and Lehigh University, we have acquired permission to release the dataset for non-commercial purposes. We plan to release the collection online for research related and academic use with a basic open license agreement. This online collection will allow the documents to be searched by attributes, thus giving researchers a vast assortment of useful sub-collections.

### III. CATEGORIZING THE LSC

Initial categorization for a small subset of the documents that have been digitized can be found in Fig 2. We have identified 19 useful attributes to record. The raw information column represents the calculated number of documents with that attribute based on the percentages.

In the first five rows, we collect the basic scanning and writing information. The values for machine print and handwriting represent the presence of each on the page, thus a page may contain both. The color statistics rely on our scanning procedures. Our digital copier, a Ricoh Aficio MP 6002/MP, scans pages automatically and senses color based on content, so the value for color represents how many pages contain some color.

The next two rows contain important release information. It is important to note that a majority of the documents can be categorized as unpublished material, or internal company material. Any document that has not been mass printed and released is considered unpublished. This includes inter-office communications and lab records. Any publicly available brochures or pamphlets would be considered published.

The bottom left two rows denote information about drawings and images. A hand drawn image is considered anything sketched on a document. This category mainly includes sketches on notes or memos. Formal drawings are considered neat and accurate. These include official plant schematics and engineering drawings.

	Raw	Percent		Raw	Percent
Machine Print	194	97%	Graphs	2	1%
Handwriting	20	10%	Photos	2	1%
Color	160	80%	Lists	78	39%
Grayscale	36	2%	Highlighting	2	1%
Black & White	4	18%	Tables	24	12%
Unpublished	196	98%	Dates	196	98%
Published	4	2%	Signatures	4	2%
Hand Drawn					
Images	22	11%	Rulings	6	3%
Formal Drawings	46	23%	Logos	128	64%

Fig. 2. Initial categorization statistics for the LSC.

Finally, the right column shows statistics for attributes that have been found in currently existing datasets. These will likely be the categories used for searching the digitized collection and organizing sub-sets. It is important to note that even the attributes that seem underrepresented in this collection, such as graphs which appear in only one percent of the current subset, still represent a large group. Considering the ultimate size of the dataset, one percent representation may yield hundreds of documents. It is also likely that other subsets of the collection will yield different statistics, a point to be confirmed as we work through the collection.

### IV. DISCUSSION

The initial statistics discussed represent a portion of the dataset, but more work must be done in order to fully represent the entire collection. As we progress through the categorization and release phases, we must make the full digitization of the LSC a priority, and the opportunity to discuss this work-in-progress at DAS will greatly benefit our efforts. Currently, we plan to make documents and categorizations available by June of 2014. A small sample of the collected and organized documents can be found at:

<http://www.cse.lehigh.edu/~lopresti/LehighSteelCollection/>

#### REFERENCES

- [1] Shope, Dan, and Kurt Blumenau. "Vaunted Steel Research Lab Closing." *Morning Call* [Bethlehem, PA] 23 04 2005, Web. 22 Jul. 2013.
- [2] Shope, Dan, and Kurt Blumenau. "New Owners Shut Down Former Homer Labs." *Morning Call* [Bethlehem, PA] 22 04 2005, Web. 22 Jul. 2013.
- [3] Shope, Dan. "Homer Research Laboratories Has Only a Glorious History to Carry On." *Morning Call* [Bethlehem, PA] 31 12 2005, Web. 22 Jul. 2013.
- [4] Loomis, Carol. "The Sinking Of Bethlehem Steel A hundred years ago one of the 500's legendary names was born. Its decline and ultimate death took nearly half that long. A FORTUNE autopsy." *Fortune* [New York, NY] 05 04 2004, Web. 22 Jul. 2013.
- [5] Bruno, Barri, and Daniel Lopresti. "The Lehigh Steel Collection: A New Open Dataset for Document Recognition Research." *Document Recognition and Retrieval Conference XXI Proceedings*. (2014): in press.

# A simple approach to distinguish between Maghrebi and Persian calligraphy in old manuscripts

Insaf Setitra

Research Center on Scientific and Technical Information  
University of Sciences and Technology Houari Boumediene  
Algiers, Algeria  
[isetitra@cerist.dz](mailto:isetitra@cerist.dz)

Abdelkrim Meziane

Research Center on Scientific and Technical Information  
Algiers, Algeria  
[ameziane@cerist.dz](mailto:ameziane@cerist.dz)

**Abstract**— *Manual annotation of images is usually a mandatory task in many applications where no knowledge about the image is available. In presence of huge number of images, this task becomes very tedious and prone to human errors. In this paper, we contribute in automatic annotation of Arabic old manuscripts by discovering manuscript calligraphy. Arabic manuscripts count a very large number of Persian and Maghrebi writing especially in North Africa. Distinguishing between these two calligraphies allow better classifying them and so annotating them. We use background constructing followed by extraction of simple features to classify Arabic manuscript calligraphies using a Quadratic Bayes classifier.*

**Keywords**—*manuscript; handwriting; calligraphy classification; supervised learning;*

## I. INTRODUCTION

In an ideal setting, an annotation of Arabic old manuscript images would provide the user with several features about the manuscript including the author, the year/century of writing, the topic, the calligraphy and so on. However, since manuscripts are usually prone to degradations, having this information is very difficult. Usually, old manuscripts miss pages which contain the author name, the date and other information. To non experts, knowing origins of the manuscript is very difficult. The task in that case is harder for a system which aims to annotate automatically the manuscript. In order to overcome these issues works on pattern recognition can be applied such as character recognition. Literature counts a big amount of works dedicated to character analysis applied to Arabic manuscripts and handwriting [1] [2] [3] [4]. Pervez and Al-Ohali give in [1] a good overview of Arabic character recognition. In [2] binarization using a new feature representation is presented. An adapted segmentation to the nature of Arabic handwriting is presented in [3]. In [5] authors are interested in presenting features for Arabic handwritten recognition. Our main contribution in this work is to use simple yet effective techniques to distinguish between two widespread calligraphies in Arabic old manuscript. We use simple features based on perimeter and simple classifier based on quadratic Bayes to do so.

## II. OUTLINE OF OUR APPROACH

Our algorithm has many details, so we first present a high-level summary. We divide, as most classification problems, our processing in two phases. A training phase and a test

phase. Processing in both phases is quite similar. First, regions of a manuscript are extracted. Then, for each region shape features are extracted. Shape features include centroid, perimeter, area, convexity, solidity, and orientation. A detailed description of all image features can be found in [6]. Our visualization of Maghrebi and Persian calligraphy showed that it is easy to distinguish between Persian and Maghrebi by looking at regions perimeters. We use this clue to decide on the perimeter feature used to distinguish between both calligraphies. Moreover, in order to isolate regions we construct a background that we use to replace regions not belonging to the same word in an extracted region by propagating neighboring of the detected region. After getting regions features, we construct our feature vector formed only of perimeter. We label manually our manuscript images by dividing them into two classes: Persian calligraphy images and Maghrebi calligraphy images. In the test phase, the same process is performed, i.e. extract perimeter feature and histogram of the test image manuscript. To achieve classification a simple Bayes decision rule was applied. We detail each step in following paragraphs.

## III. OUR APPROACH

### A. Initial Regions extraction

Input of this step is a manuscript colored image. First the image is converted to grayscale using (1).

$$I_{i,Gray}(x,y) = 0.2989 * I_{i,R}(x,y) + 0.5870 * I_{i,G}(x,y) + 0.1140 * I_{i,B}(x,y) \quad (1)$$

where  $I_{i,Gray}(x,y)$  is the gray level pixel,  $I_{i,R}(x,y)$ ,  $I_{i,G}(x,y)$  and  $I_{i,B}(x,y)$  are red, green and blue pixel of the  $i$ th manuscript image pixel respectively. After converting the image to grayscale, we extract its contour using a canny edge detector.

### B. Background construction

Previous step provides us with regions of the manuscript images with their bounding boxes. However, due to bounding boxes overlap, extracting sole regions will be difficult. Indeed, since our study focus on region perimeters, having a part of a region in a bounding box of another region will false our results. In order to overcome this previous issue, we construct a background that we apply only on the connected region detected in the previous step.

We first constructed background by propagating the mean of the 3x3 neighboring pixels of a detected region into the latter as follow:

$$B(x,y) = \begin{cases} I(x,y) & \text{if } mask(x,y) = 0 \\ \text{mean}(\text{neighboring}(I(x,y))) & \text{if } mask(x,y) = 1 \end{cases} \quad (2)$$

Where  $B(x,y)$  and  $I(x,y)$  are pixels of background and original image at location  $x,y$  respectively.  $mask(x,y)$  is the binary mask figured in fig. 9 and 10 where pixels having value 0 are background and pixels having 1 value are regions detected.  $\text{neighboring}(I(x,y))$  is the mean of the 3x3 neighbors of pixel  $I(x,y)$ . Update is done by replacing the original pixel image in location  $x,y$  with that mean and replacing the mask pixel in the same location with 0. This avoids having the same result as the original image and not constructing any background.

We noticed that background with this approach was extracted poorly and contours of regions only are highlighted with the mean. We explain that with the fact that all pixels of a region are replaced with the same value which is the mean and no continuity of texture is respected. In a second approach, we instead, propagate only value of one pixel not belonging to this or another region as follows:

$$B(x,y) = \begin{cases} I(x,y) & \text{if } mask(x,y) = 0 \\ \text{First}(\text{neighboring}(I(m,n))) & \text{if } mask(m,n) = 1 \end{cases} \quad (3)$$

Where  $B(x,y)$  and  $I(x,y)$  are pixels of background and original image at location  $x,y$  respectively.  $mask(x,y)$  is the binary mask where pixels having value 0 are background and pixels having 1 value are regions detected.  $\text{first}(\text{neighboring}(I(x,y)))$  is the first pixel found in the 3x3 neighbors of pixel  $I(x,y)$  which neither belongs to the same region as  $I(x,y)$  nor to another region. Update of the original and mask image are done the same as the previous method but in all neighboring window of  $I(x,y)$  with the value of the selected pixel for the original image. Note that we are not replacing values of original images and masks directly, we use a copy of them. Note also that we are processing directly with the original image in its three channels. The major problem of this approach is that background cannot be constructed in the border of the image if border pixels with their neighboring are detected regions. We leave this improvement to a further work. We further improved the background resulting image by applying an averaging filter.

### C. Final region and feature extraction

After having the manuscript image mask, the background and the bounding box of each region, we simply apply the image mask to the background for each region. By having all regions stored, we can extract our feature vector. The initial feature vector is a multidimensional vector including several shape features (centroid, bounding box, area, convexity, solidity and perimeter). However, our observation on both calligraphies conducted us to choose only perimeter which looks discriminative for both calligraphies (Maghrebi and Persian). Feature vector is then a two dimensional vector where first dimension is the region and second is the perimeter.

### D. Classification

As mentioned above, we use a simple Bayes decision rule. We chose Quadratic Bayes classifier because of its simplicity, its fast computations and its small number of parameters. Normal distribution is used for modeling each class.

Covariance matrix and mean are computed in the training phase and are used in the testing phase to classify new observations as in (4).

$$p(x | \omega_j) = \frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j)\right)$$

Where  $\Sigma_j$  and  $\mu_j$  are the covariance matrix and the mean of the class  $\omega_j$

## IV. CONCLUSION

In this paper we presented briefly our approach for distinguishing between Maghrebi and Persian calligraphy. Our approach involves training and classification using background subtraction, perimeter feature and histogram and Bayes classification rule for classification. The approach gave good results but we aim to improve them by including more robust features so that we can classify more calligraphy. We understand that the work presented lacks a quantitative evaluation. A ROC curve will be added in the final version of the paper.

### ACKNOWLEDGMENT

The authors would like to thank Pr. Mohamed Cheriet and Dr Rachid Hedjam for thought-provoking discussions. This work was supported by the Algerian National Project of Research on Old manuscripts restoration.

### REFERENCES

1. Mohamed Cheriet, Reza Farrahi Moghaddam, Rachid Hedjam. A learning framework for the optimization and automation of document binarization methods. *Computer Vision and Image Understanding*. March 2013, Vol. 117 Issue 3, pp. 269-280.
2. Mohammad Tanvir Parvez, Sabri A. Mahmoud. Arabic handwriting recognition using structural and syntactic pattern attributes. *Pattern Recognition*. January 2013, Vol. 46, Issue 1, pp. 141-154.
3. Pervez Ahmed, Yousef Al-Ohali. Arabic Character Recognition: Progress and Challenges. *Journal of King Saud University - Computer and Information Sciences*. 2013, Vol. 12, pp. 85-116.
4. Al-jawfi, ashad. "Hand writing Arabic Character Recognition Using Neural Network". LeNet, 2007.
5. Chherawala, Y., Roy, P.P. et Cheriet, M. Feature Design for Offline Arabic Handwriting Recognition: Handcrafted vs Automated? *12th International Conference on Document Analysis and Recognition (ICDAR)*. Aug. 2013, pp.290,29
6. Mohamed Cheriet, Nawwaf Khrama, Cheng-Lin Liu, Ching Suen Character Recognition Systems: A Guide For Students And Practitioners ISBN: 978-0-470-17652-8



# Boosting-based approaches for Arabic text detection in news videos

Sonia Yousfi, Sid-Ahmed Berrani

Orange Labs – France Telecom  
35510 Cesson-Sévigné, France  
{sonia.yousfi, sidahmed.berrani}@orange.com

Christophe Garcia

University of Lyon, INSA-Lyon, LIRIS, UMR5205 CNRS  
69621 Villeurbanne, France  
christophe.garcia@liris.cnrs.fr

**Abstract**—In this paper, we propose two boosting-based approaches for Arabic embedded text detection in news videos. The first approach uses Multi-Block Local Binary Patterns features whereas the second one relies on Haar-like features. Both approaches learn text and non-text classes using a multi-exit asymmetric boosting cascade. Bootstrap has also been used in order to improve the rejection ability of the classifiers. Text localization is performed by a sliding window search on a multi-scale pyramid of the input image. The proposed approaches have been evaluated on a large database of images coming from 4 different Arabic TV channels.

## I. INTRODUCTION

Embedded text in video frames provides valuable information on what is being shown. It is therefore very important for structuring and indexing news videos. This paper focuses on the detection of embedded text within the frames of Arabic news videos. It is a very challenging task due to text variations (e.g. size, style) and acquisition conditions (e.g. background complexity and variability). The focus on Arabic text is motivated by many reasons. First, this language is used by more than half of a billion people in the world and many big Arabic news channels appeared in the last two decades. Second, Arabic text has many specific properties different from other types of texts (cursive text, more strokes in different directions, different aspect ratio...). Finally, there are only very few works that have addressed this problem.

We propose two approaches based on a multi-exit asymmetric boosting cascade. The first approach selects text features with the Multi-Block Local Binary Pattern representation and performs training using the Gentleboost algorithm. The second method relies on Haar-like features and Adaboost. Experiments show that both methods capture different text features at the same time. Moreover, a good precision rate has been obtained thanks to the bootstrapping technique.

## II. RELATED WORK

There are three main approaches for embedded text detection in videos: (1) Machine learning-based methods aim to learn discriminative features from a training text data-set using for example neural networks [1]; (2) Heuristic-based methods apply manually inferred rules directly on the low-level features based on the observation of text characteristics [2]; (3) Hybrid methods make use of both heuristics and machine learning techniques [3]. All of these techniques are dedicated to Latin text detection. Existing methods dedicated to Arabic

text detection in video images are very few. Ben Halima et al. [4] used Multi Frame Integration process to decrease background variations and extract lines and columns that probably contain text. They used then a neural network to learn edge and color features to refine previous results. Moradi et al. [5] proposed a purely heuristic approach. The method is based on morphological operation applied on edge maps. Text localization and elimination of false alarms are achieved by profile projection analysis and empirical rules. Ahmad et al. [6] used almost the same strategy but it is based on the Laplacian operator to extract edges and the k-means algorithm to cluster text regions.

## III. THE PROPOSED APPROACH

Our method for embedded Arabic text detection in video frames is based on a boosting procedure. First, text/non-text machine learning-based classifiers are constructed. Secondly, a multi-scale research procedure is applied on video images using the trained classifiers in order to perform text localization. Our classifiers are built using a multi-exit asymmetric boosting cascade that learns to distinguish text and non-text using Multi-Block Local Binary Patterns (MBLBP) and Haar-like features.

**Feature extraction:** MBLBP features consist in encoding rectangular regions using the Local Binary Pattern operator. Then, a binary code is produced by comparing the average intensity of the central rectangle  $r_c$  with its  $3 \times 3$  neighborhood  $X = \{r_0, \dots, r_8\}$ . Haar-like features, popularized by Viola and Jones [7] for face detection, are based on the difference measure between the average intensities in rectangular regions. To improve computational efficiency of these two descriptors, we classically use the integral image technique [7].

**Multi-exit asymmetric boosting:** In order to build the text/non-text classifier, we propose to use the multi-exit asymmetric boosting cascade introduced in [8]. In this cascade, intermediate strong classifiers are represented by a set of nodes having indices  $\aleph$ . Each classifier (node) makes a decision to pass or reject the input subwindow. Each strong classifier is constructed from a sequence of  $n$  weak classifiers. Unlike conventional cascade proposed in [7], nodes are able to use overlapped sets of weak classifiers, i.e. each node exploits these weak hypotheses from the beginning to a number  $n$  which corresponds to a fixed training detection rate and false positive rate. The final classifier resulting from the cascade is



Fig. 1. Examples of images used for training

expressed as follow:

$$F(x) = \begin{cases} +1, & \text{if } \sum_{i=1}^n w_i h_i(x) \geq 0, \forall n \in \mathbb{N}; \\ -1, & \text{otherwise} \end{cases} \quad (1)$$

where  $h_i$  are weak classifiers and  $w_i$  their weights.

Under this general approach, we propose two schemes. In the first one, MBLBP features are learned at each node of the cascade by Gentleboost algorithm. In the second one, relevant Haar-like features are selected by the Adaboost procedure. In order to improve the false alarms rejection ability of our systems, we use a bootstrap procedure.

**Multiscale text search:** To localize text, a multi-scale search is performed by a sliding window. We apply, then, a kmeans-like algorithm to construct text regions with different densities. By applying a threshold on that density, we can eliminate a considerable number of false alarms.

#### IV. EXPERIMENTS

1) *Datasets:* To train our classifiers, we have used 30,000 text patches with an aspect-ratio of 3 selected from 3 Arabic TV channels: Al Jazeera, Al Arabiya and France24 Arab. The initial training set contain 20,000 non-text patches. By applying bootstrap technique along training and using a set of scene images, we reach 110,000 negative examples (Figure.1). Final detection results have been evaluated using 2 other sets: ES1 that contains 201 frames from Al Arabiya, Al Jazeera and France24 Arabic with 959 annotated texts, and ES2 that contains 164 frames collected from the BBC Arabic channel. This channel has not been used during training. It uses font and text styles that are very different from the first three channels.

2) *Evaluation results:* To evaluate our methods in term of recall an precision, we have used the metric proposed in [9].

As for the evaluation of the final detectors, results obtained on ES1 are reported on Table I. The third column presents the harmonic mean (also called f-measure) that combines precision and recall in one value. The forth column reports the number of the obtained detections. These results reflect the good detection capacity of the proposed methods specially with M1. Combining these values with the amount of detections reflects the good rejection abilities of false alarms. This is a result of using the bootstrap procedure during training.

Evaluation results on ES2 are presented in Table II. The obtained values reflect the good generalization of our methods. There are almost no significant difference between these results and the previous ones (reported in Table I). This can be explained by the efficiency of the used features that capture general characteristics of the Arabic text. This is also partly due to the generalization capacity of the boosting-based machine learning itself. It is based on the inference aspect of

	Recall	Precision	F-measure	# detections
HAAR+Adaboost (M1)	0.77	0.72	0.74	1026
MBLBP+GentleBoost (M2)	0.7	0.32	0.44	1400

TABLE I. EXPERIMENTAL RESULTS ON ES1.

	Recall	Precision	F-measure	# detections
HAAR+Adaboost (M1)	0.75	0.66	0.70	804
MBLBP+GentleBoost (M2)	0.72	0.25	0.37	1522

TABLE II. EXPERIMENTAL RESULTS ON ES2.

general strong hypotheses for classification at each step of the cascade. Both results show that M1 outperforms M2 method which can be explained by the fact that MBLBP is dedicated to capture large scale structures which is not very suitable for text texture. Few results generated using the M1 scheme are shown in Figure 2. We can see in image (b) that some scene texts are also detected (red bounding boxes) but they are considered as false alarms by our evaluation procedure and this explains partly the low value of precision.



Fig. 2. Examples of detection results using M1.

#### V. CONCLUSION

We have presented in this study two boosting-based approaches for embedded Arabic text detection in news videos. Experimental results highlight the good detection abilities of our methods. In our future work, we will address Arabic text recognition in videos with all difficulties related to Arabic text.

#### REFERENCES

- [1] M. Delakis and C. Garcia, "Text detection with convolutional neural networks," in *Proc. of the Int. Conf. on Computer Vision Theory and Applications, Funchal, Madeira, Portugal*, January 2008.
- [2] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition, San Francisco, USA*, June 2010.
- [3] M. Anthimopoulos, B. Gatos, and I. Pratikakis, "A two-stage scheme for text detection in video images," *Image Vision Computing*, vol. 28, no. 9, pp. 1413-1426, September 2010.
- [4] M. B. Halima, H. Karray, A. M. Alimi, and A. F. Vila, "NF-savo: Neuro-fuzzy system for arabic video ocr," *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 10, pp. 128-136, November 2012.
- [5] M. Moradi, S. Mozaffari, and A. Orouji, "Farsi/arabic text extraction from video images by corner detection," in *Proc. of 6th Iranian Conference on Machine Vision and Image Processing*, October 2010.
- [6] A. M. Ahmad, A. Alqutami, and J. Atoum, "A robust algorithm for arabic video text detection," in *Proc. of the Int. Congress on Computer Applications and Computational Science, Advances in intelligent and Soft Computing*, 2011.
- [7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition, Kauai, HI, USA*, December 2001.
- [8] M.-T. Pham, V.-D. Hoang, and T.-J. Cham, "Detection with multi-exit asymmetric boosting," in *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA*, June 2008.
- [9] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *Int. Journal on Document Analysis and Recognition*, vol. 8, no. 4, pp. 280-296, August 2006.

# An Android Application for Browsing and Searching in Historical Manuscripts

Alicia Fornés & Pau Riba & Josep Lladós

Computer Vision Center – Dept. Ciències de la Computació

Universitat Autònoma de Barcelona

08193 Bellaterra (Cerdanyola) Barcelona, Spain

{afornes, josep}@cvc.uab.es

**Abstract**—In this paper we present a prototype system running on Android devices for browsing and searching words through historical handwritten documents. The idea is to adapt the eBook paradigm, replacing digitally born documents by scanned historical manuscripts. Concretely, we have selected a collection of historical handwritten marriage licenses. Thanks to the implemented document analysis techniques, the user can easily cast queries (e.g. people names, occupations, places, etc.) and follow genealogical links.

## I. INTRODUCTION

In the last years, mobile devices like eBooks, tablets and smart-phones have become very popular. Therefore, the reading paradigm is moving from printed and static support, to electronic devices, which allow active reading: hyper-links, annotations, searching in dictionaries by clicking at words, etc.

Digitally born documents can be easily generated with their corresponding meta-data. However, historical books have to be digitized and processed by document analysis techniques in order to allow active reading. The adaptation of scanned historical documents to mobile devices has been addressed in some works, especially concerned on document reflowing, i.e. the adaptation of the layout presentation to the output device [1]. But interesting challenges are envisioned on the access to contents using these devices. In this paper we present a pioneering application for searching and browsing in handwritten historical documents collections. Technically, we have integrated handwritten word spotting in an Android device.

As use case scenario, the proposed application incorporates a collection of historical handwritten marriage licenses [2]. The reason for using such type of documents is the emerging interest in genealogy research. Queries about people and their connections will allow genealogical and demographic research studies (e.g. a person's life line, ancestors and descendants crosslinking, events, etc.). From a functional point of view, it is a good example of focused retrieval using words. In addition, from an ergonomics and efficiency point of view, the use of mobile platforms may represent a step forward towards the universal access to the public heritage residing in archives. Whereas physically browsing through different books in a genealogical search can require significant time and effort, the use of a portable device will dramatically improve the task.

In this paper we describe an Android mobile application for browsing historical archives. It integrates the functionality of retrieving, while browsing, the pages containing instances

of queried words, such as names, surnames, occupations or places. The main difficulty is to keep the performance of the implemented document analysis and recognition tasks, coping with the resources limited of such devices (processing power and memory). Next, in section II we overview the main functionalities of the application. In section III we overview the key technical details of the application. Finally in section IV we draw the conclusions and present the future improvements.

## II. FUNCTIONAL AND NON FUNCTIONAL REQUIREMENTS

The application has been designed with an intuitive front-end interface so the user can easily browse and search by using intuitive gestures.

### A. Browsing

The user can browse through the document collection by touching the screen. The next or the previous page is showed by a swiping movement to the left or right. Zoom in and zoom out are performed by 2-finger press, moving outwards or inwards. It follows the standard gesture language of touch devices like tablets and smartphones.

### B. Searching

The search interface divides the screen in two panels. When a page is displayed in the main panel (the browsing one), the user can select a word by a long press in the screen. Then, the system shows the selected word in a red box, and asks for confirmation. If the user agrees, the system shows, in a right side panel, the list of retrieved words in the rest of the pages. Whenever the user touches a word in the list, the application shows in the main panel the page in which that word is contained. In case there are more instances of that word in the page, they are all displayed in red bounding boxes (see Figure 1). In this figure, the user has selected the word "rebere". In the right panel the system shows the snippets in the ranked list of correct matches.

### C. Software and Hardware Constraints

We have used a Samsung Galaxy Note 10.1 tablet for implementing the Android application. The specifications are the following: Quad-core processor at 1.4GHz, 2GB of RAM, Screen size of 10.1 inches, and Screen resolution of 1280x800 pixels. It runs Android 4.0.

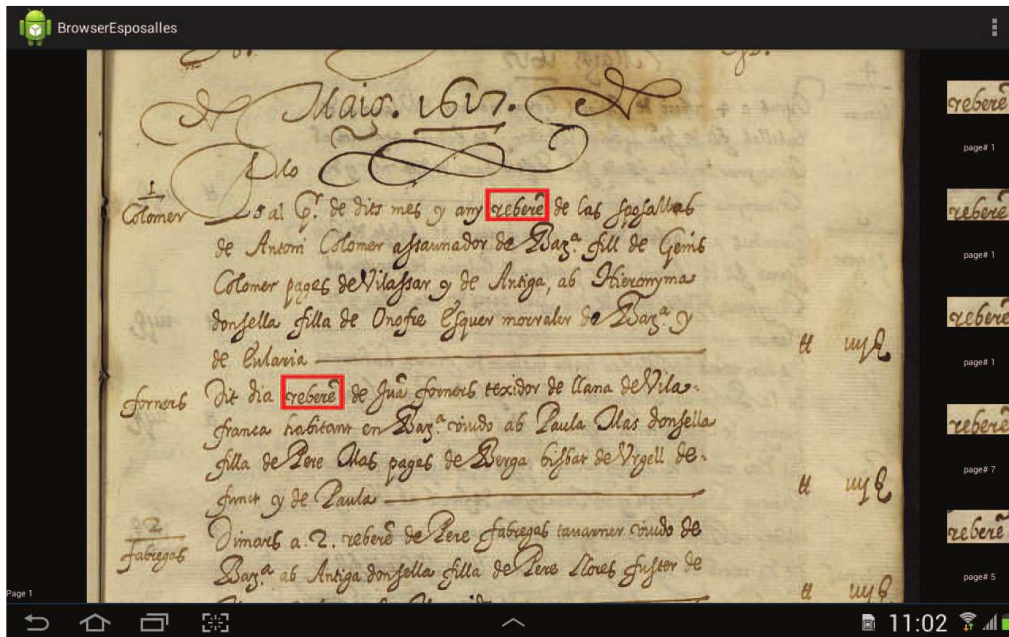


Fig. 1: A snapshot of the application when searching for a word.

### III. IMPLEMENTATION OUTLINE

Technically, the described functionalities are implemented by a query by example word spotting paradigm. For the sake of real time retrieval, the images of the database are preprocessed and indexed in the device. First, individual words are segmented and a signature describing each word image is computed.

#### A. Word Segmentation

In this step, all words contained in the documents are segmented. For this purpose, we have applied the word segmentation method described in [3]. This method is specially designed for noisy historical documents. First, text lines are segmented using a seam carving approach. Then, each text line is filtered using an anisotropic Gaussian filter, which is applied at several scales in order to obtain blobs that correspond to characters and words. In this way, the approach is tolerant to different scanning resolutions and word sizes. Finally, the resulting segmented words have been manually checked and corrected if necessary.

#### B. Word Spotting

Once words have been segmented, we have applied the word spotting method described in [4]. This method is based on attributes, which leads to a low-dimensional, fixed-length representation of the word images. The main advantage is that this approach is fast to compute and compare, suitable for a real-time application for mobile devices. It must be noted that the hardware specifications of mobile devices are lower than the ones for personal computers.

### IV. CONCLUSION AND FUTURE WORK

In this paper we have presented a mobile application for searching and browsing historical document collections in a very intuitive manner. In this way, the user can browse the documents, zoom in/out, and touch the query word for Query-by-Example word spotting. As future work, we are planning to incorporate Query-by-String and Query-by-Sketch searches, with the purpose of increasing the functionality. Preliminary promising results have been obtained for a Query-by-String word spotting.

The application can be installed on any Android device. In the near future, we are also planning to adapt it to iOS and Windows Phone-based devices.

#### ACKNOWLEDGMENT

This work has been partially supported by the Spanish project TIN2012-37475-C02-02 and the European project ERC-2010-AdG-20100407-269796.

#### REFERENCES

- [1] S. Marinai, A. Anzivino, and M. Spampini, "Towards a faithful visualization of historical books on e-book readers," in *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, ser. HIP '11, 2011, pp. 112–119.
- [2] V. Romero, A. Fornés, N. Serrano, J. A. Sánchez, A. H. Toselli, V. Frinken, E. Vidal, and J. Lladós, "The *Esposalles* database: An ancient marriage license corpus for off-line handwriting recognition," *PR*, vol. 46, no. 6, pp. 1658 – 1669, 2013.
- [3] R. Manmatha and J. L. Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1212–1225, 2005.
- [4] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Handwritten word spotting with embedded attributes," in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 1017–1024.

# 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](mailto: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<sup>1</sup> 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.

<sup>1</sup> F. LeBourgeois, H. Empotz, Document Analysis in Gray level and typography extraction using Character Pattern redundancies, ICDAR, Bangalore India, p177-180. 1999.

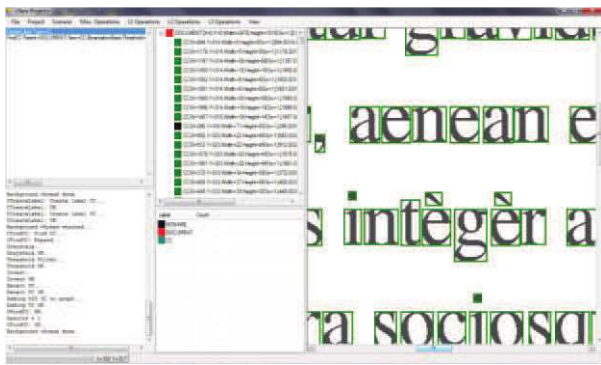


Figure 1: GUI of AGORA

*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.

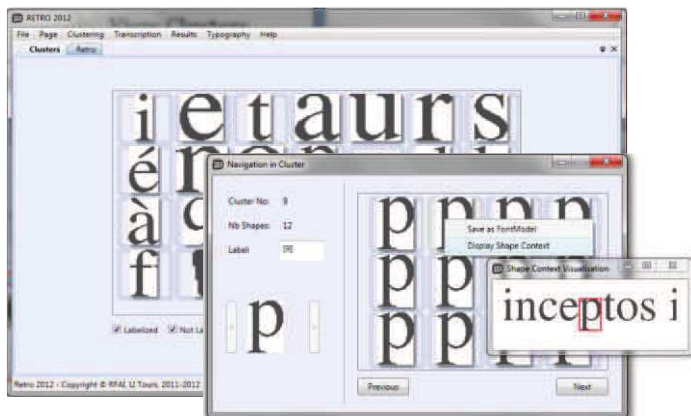


Figure 2: Interactive transcription with RETRO;

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.

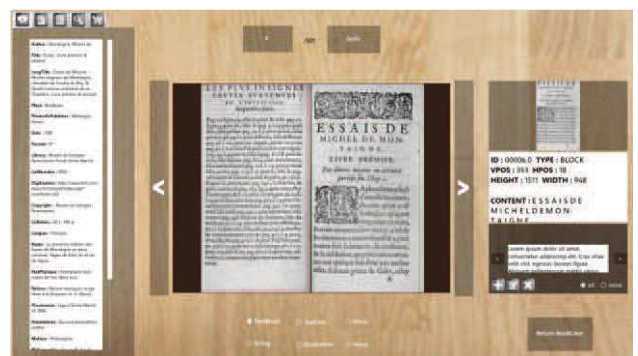


Figure 4: Multi-touch interface

ACKNOWLEDGEMENT

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

# Efficient OCR Training Data Generation with Aletheia\*

Christian Clausner, Stefan Pletschacher and Apostolos Antonacopoulos  
PRImA Lab, School of Computing, Science and Engineering, University of Salford,  
Greater Manchester, M5 4WT, United Kingdom  
<http://www.primaresearch.org>

**Abstract**—We present how the ground-truthing tool Aletheia can be used to efficiently create training data for an open-source text recognition engine. The labelling process is sped up considerably through a top-down approach. Text content is thereby entered on region level. The characters are then propagated automatically to glyph objects. In addition, segmentation is simplified by several semi-automated tools.

**Keywords**—Optical character recognition; training data generation; ground truthing

## I. INTRODUCTION

Methods for Optical Character Recognition (OCR) often need to be trained for new fonts or symbols. Training data can be either synthesised or extracted from document images (ground truth). Especially in the case of historical documents a synthesis is usually not feasible because font descriptions are not available.

Extracting training data of sufficient quality and quantity is cumbersome. It requires a precise representation of shape and class (character code) for a large amount of glyphs.

In this paper we demonstrate the use of the ground-truthing tool Aletheia [1] to generate training data for the Gamera open source OCR toolkit [2]. The same principle can be applied to train other OCR engines as well but may require conversion to the corresponding training data formats.

## II. TRAINING DATA GENERATION

### A. Ground Truth Creation

Aletheia is an advanced tool for creating page layout and text ground truth for document images. It supports top-down (from regions to glyphs) as well as bottom-up (from glyphs to regions) workflows.

Ground truth is stored in the PAGE XML format [3] wherein text objects are represented by (arbitrary) polygons and Unicode text content. Four levels of objects are available: Regions (blocks), text lines, words, and glyphs.

Semi-automated tools help to reduce manual labour and thereby increase the efficiency. Figure 1 shows the final stage of the top-down approach where word objects are split into glyph objects. In most cases this requires just one click in the

gap between each pair of adjacent glyphs. The outlines are then automatically calculated based on the pixel information.

Wrongly generated object outlines can be corrected by directly manipulating the polygons. The degree of manual intervention required depends on the quality of the document image (noise, scanning artefacts, etc.).



Fig. 1. Splitting word objects into glyph objects in Aletheia (top-down approach).

Under certain circumstances (for instance heavily degraded images), the bottom-up workflow is favourable. Glyphs are marked and subsequently grouped into words, text lines, and regions.

Text region, text line, and word objects are usually not necessary for OCR training data generation. They can, however, significantly speed up the task of assigning the character classes to glyph objects. Text can be entered conveniently at region level and can then be propagated down to glyph level. Aletheia automatically matches layout objects with their corresponding text content. Figure 2 shows an example for text entry.

The matching requires a consistent handling of white spaces and ligatures. If, for example, a punctuation character is not separated from the adjacent word by a space, the corresponding glyph object should also be part of the respective word object. Similar considerations are required for ligatures. They can either be represented by one single pre-composed character or using a decomposed sequence of characters. In both cases, the glyph segmentation has to match the number of characters given in the transcribed text. Aletheia highlights segmentation inconsistencies to speed up their correction.

A virtual keyboard simplifies the input of special characters and symbols. The selection and layout of the keys is fully customisable.

\*This work has been supported in part through the EU 7<sup>th</sup> Framework Programme grant Europeana Newspapers (Ref: 297380).

III. EXPERIMENTAL VALIDATION

For proof of concept a document page with about 3000 glyphs has been processed. On average it took 2.1 seconds to mark and label one glyph. The result has then been applied to train a classifier using the Gamera interactive classifier tool. The application of this classifier to a document image that was not part of the training is shown in Figure 3.

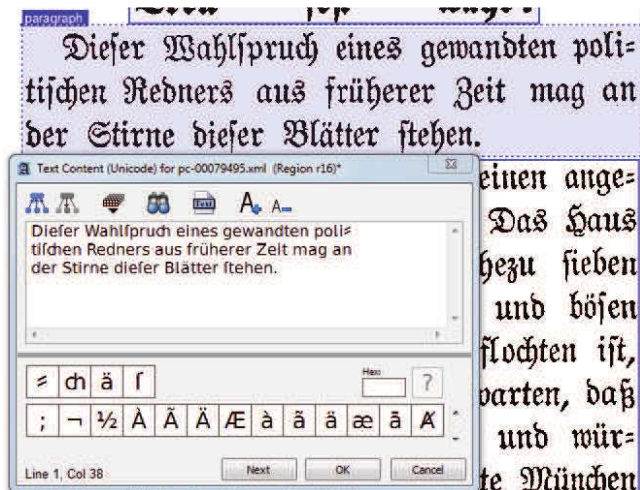


Fig. 2. Entering text content in Aletheia.

B. Conversion to Gamera Format

Generated ground truth needs to be converted to an XML format specific to Gamera [4]. For this purpose, a tool has been developed that transforms a PAGE XML file (the output of Aletheia) and the corresponding document image to a valid Gamera training data description.

Glyph shapes need to be translated from polygons to run-length encoding. This is done by scanning the image pixel data of the area inside a polygon. A bi-level image is therefore mandatory. Aletheia provides a basic set of binarisation and noise removal methods to this end.

Character classes are represented hierarchically in Gamera using a dot-separated name pattern, which usually corresponds to the respective textual Unicode descriptions (see Table I for examples). A look-up table has to be defined, containing all characters that may occur in the documents that are to be processed.

TABLE I. EXAMPLES OF CHARACTER CLASSES IN PAGE XML AND GAMERA XML

PAGE (Unicode)	Gamera (generic)
004E	latin.capital.letter.n
002C	latin.punctuation.comma
0030	digit.zero

Once converted, the training data can be applied to a Gamera classifier. This process is not limited to one file. Data from multiple pages can be added incrementally.

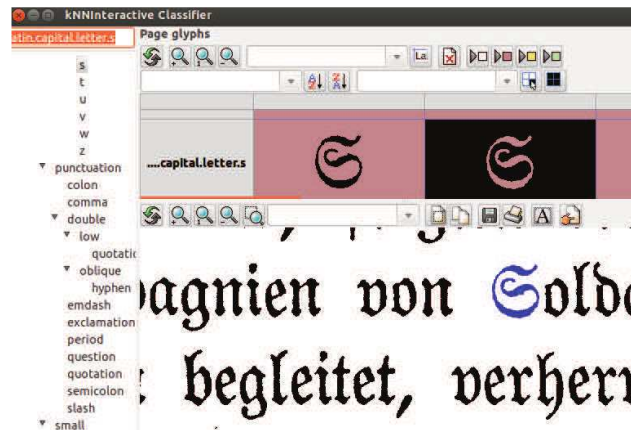


Fig. 3. Classification in Gamera after training with data that has been extracted using Aletheia.

IV. CONCLUSION AND FUTURE WORK

It has been presented how the Aletheia ground-truthing tool can be used for efficient generation of OCR training data. Semi-automated tools in combination with a mature user interface can speed up the extraction process in comparison to other tools (for instance the Gamera interactive classifier tool).

Future work will include the investigation and development of the conversion to formats required for training further OCR engines such as Tesseract.

REFERENCES

- [1] C. Clausner, S. Pletschacher, A. Antonacopoulos, "Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments", Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011), Beijing, China, pp. 48-52, September 2011
- [2] M. Droettboom, K. MacMillan, I. Fujinaga, "The Gamera framework for building custom recognition systems", Symposium on Document Image Understanding Technologies, pp. 275-286, 2003.
- [3] S. Pletschacher, A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010), IEEE-CS Press, pp. 257-260, Istanbul, Turkey, August 23-26, 2010.
- [4] Documentation for Gamera XML format version 2.0, [http://gamera.sourceforge.net/doc/html/xml\\_format.html](http://gamera.sourceforge.net/doc/html/xml_format.html) (date accessed: 18/12/2013)



# Dehyphenation by Classification for OCR Results

Mayce Al Azawi and Thomas M. Breuel  
 University of Kaiserslautern  
 Postfach 3049  
 D-67653 Kaiserslautern, Germany  
 Email: ali,tmb@cs.uni-kl.de

**Abstract**—We propose various methods for dehyphenation of the OCR results. Our purpose is to remove the hyphen and convert the OCR results to readable and reflowable format so that they can be used for different purposes such as OCR post-processing and digitalization of the transcriptions with the corresponding historical books. For historical documents, the transcription is not always available and it is very time and money consuming to transcribe them. Therefore the OCR systems are used to provide the transcriptions. In OCR post-processing, a corresponding transcription is mostly required for different purposes such as n-grams language modeling for OCR corrections or aligning the image with the transcription which has different line breaks from the original book. A hyphen is not necessarily used as line break, for languages such as German, a hyphen could be used frequently to combine two tokens and produce a word. OCR could also have misrecognized symbols as a hyphen which is original not. Hyphens are also used to stand for a common second element in all but the last word of a list. We propose dehyphenation models using different classifiers such as: Decision Trees, Naive Bayes and Maximum Entropy classifiers. The methods are language-independent. We applied our models on ocred data from historical German Fraktur. The accuracy of our dehyphenation model is 97% while the based-line is 74%. We applied on English ocred data from UWIII and the performance of our model is 98% while the based-line returned 64%. We show how the construction and performance of significant, fast models.

## I. INTRODUCTION

The purpose of dehyphenation is to provide the output of the OCR systems in readable, reflowable and clean format by removing the hyphen which is used as line break or any symbol misrecognized as hyphen and deliver a cleaned and easy to adapt text format of the OCR results to be used in different purposes. Accurate OCR raw outputs is most important resource for digitizing book, specially for historical documents the transcription is not available or available with different text modification and hyphenation. The dehyphenation is used for delivering the raw output for different OCR post-processing tasks. After applying the dehyphenation, the results can be used as transcription and aligned with the corresponding image to generate data for the training purposes. In this situation, the transcription which is provided after the dehyphenation, is easy to adapt to any different line breaks and hyphenation. The transcription could be used in representing n-grams language modeling to be used for OCR correction specially for historical documents which has no lexicon or text corpus for linguistics purposes. The hyphenated patterns and the original TEX algorithm of [1] is frequently used. Hyphenation can be carried out using a word list with hyphenation points usually derived from an electronic dictionary. The drawback of this method is that it will only cover words explicitly listed in the dictionary.

In the work of [2], they eliminated the hyphen from the text and rejoins the hyphenated word to its second half on the next line. The proposed regular expression depends on the tab or space followed by line break information. The method produced merged tokens for combined words with hyphen which does not required remove the hyphen. For example, produced “rockcarved” instead of “rock-carved”. It can not deal with capital letters, digits and letters with umlaut, i.e. “ä”, “ü”. In our approach, we treat the dehyphenation as a classification problem. Hyphens are spaced apart enough so that we can treat each hyphen/no-hyphen decision as a separate problem. The machine learning classifiers are widely used in natural language processing. The classifiers performance well with large data in short time. They are simple to understand and interpret, can handle both numerical and categorical data. For classification, we constructed de-hyphenation model using decision trees, naive Bayes and maximum entropy from the NLTK toolkit[3] and apply it on the OCR output. The classifiers were widely used, for example, in text mining, in part-of-speech tagging. We proceed by discussing the de-hyphenation problem in Section II by describing the standards use of the hyphen and the problem of the hyphenation in OCR, describing our proposed methods in Section III. Section IV presents the experimental results. Finally, Section V concludes the paper.

## II. DE-HYPHENATION PROBLEM IN OCR

In this section, we describe the specific of German and English hyphenation and the de-hyphenation Problem in OCR. To de-hyphenate OCR's results, we need a robust method against OCR errors which include recognition errors, misrecognition due the substitution errors of symbol with a hyphen or the misrecognition of the dashes as hyphen. It required also to distinguish that the end of the line is hyphenation of compound words with hyphen. The purpose is to classify the hyphen is a hyphenation of line break or hyphenation of combined words and terms. The UWIII dataset has a lot mathematical expression which is combined with a hyphen such as: “NS-1039-11” and “1073-1078”. OCR error can also occur “wHOLE-BODY”, “high-r~lution” and “Go~back-N”. Compound words appear in three tokens “Electro-Magnetic-Acoustic”. Regular hyphenated words is “distri- bution”. Compound words which hyphenated at the line break “plystyrene-poly- methylmethacrylate”. In the Fraktur documents, few compound words could appear with none capitalisation “wissenschaftlich-praktisch”.

## III. CLASSIFICATION-BASED METHODS FOR OCR DEHYPHENATION

For classification, we use decision trees, naive Bayes and maximum entropy [3]. The appropriate features which are

extracted from Fraktur were word length, and the length of both tokens joined by a hyphen because German words are longer than a hyphenated part. Capitalization is one of the most appropriate feature because the second part of the compound word could start with capital which make it very distinguishable. German nouns start with capital letter, it is more frequent that a word appear after a hyphen as second part of the compound word more than a second part of a hyphenated word at the next line.

However the compound words in English start with lower-case and some compound words have short length of words. Therefore using part-of-speech tagging support to distinguish the words from the hyphenated parts. For example, the two tokens which represent a compound word can have a POS-tagging because it could be a combination of noun-noun, adjectives-adjective, noun-participle and adjective-participle. However, the two tokens which represent a hyphenated word because of line break can not have a POS-tagging because it is not a dictionary word. Most of the hyphenated words in English when they divided to syllables, they end with tokens with no POS-tagging. We used the NLTK available tagger which is using the Penn Treebank tagset [3]. The hyphen divides the words between syllables i.e. (“bas-ket”, “pic-ture”). It can be avoided to carry over two-letter syllables to the next line (“fully”, not “ful-ly”). It can not divide a word of one syllable. Refrain from dividing any word that will result in a single-letter syllable (“again”, not “a-gain”). The hyphen divides between double consonants i.e. (“equip-ping”, not “equipp-ing”). Therefore, we use the feature vowel to learn the characteristic of the syllables. The features set is fed into the model, which generates predicted labels. The input features are:

- **wordLength**: is the length of the two tokens which connected with the hyphen.
- **part1Length**: is the length of the first part of the hyphenated or command word.
- **part2Length**: is the length of the second part of the hyphenated or command word.
- **hasTag**: is binary (“yes” or “no”) and provided by checking the the tag of the two tokens. The value is “yes” if first and second token have POS-tagging.
- **part1Vowel**: is binary (“yes” or “no”) and provided by checking the last syllable of the first token.
- **part2Vowel**: is binary (“yes” or “no”) and provided by checking the first syllable of the second token.
- **isCapital**: this value is binary (“yes” or “no”). The value is “yes” if the second token starts with upper case letter.
- **Frequency**: is an integer value. It refers to the frequency of occurrence of the tokens within some given text corpus.

#### IV. EXPERIMENTAL RESULTS

##### A. OCR and Materials

The datasets are generated from the raw output of OCRopus system after applying it on English UWIII dataset and German Fraktur Documents obtained from “Wanderungen durch

TABLE I. PERFORMANCE EVALUATION RESULTS OF THE APPROACHES APPLIED ON CLEANED DATA AND OCR’S RESULTS.

Test Sets	Based-line	Naive Bayes	Decision Trees	Maxent
Cleaned English	0.60	0.97	0.99	0.98
OCR English	0.64	0.97	<b>0.98</b>	0.97
Cleaned Fraktur	0.76	0.97	0.98	0.97
OCR Fraktur	0.74	0.96	<b>0.97</b>	0.96

die Mark Brandenburg” volumes (1862-1889) by Heinrich Theodor Fontane. The purpose is to learn the tokens which appear with or without hyphen, if the hyphen is related to line breaks or combing words or misrecognition of symbol as hyphen by the OCR. Therefore, it is also important to contain words with hyphen in different positions based on the number of the syllables. The ground truth is represented the words and corresponding labels that refer to the word if it is with hyphen or not. All classifiers were implemented in Python and applied using the NLTK toolkit 2.0 [3] under Linux. Running times for the classification and prediction are fast on a single CPU on a modern desktop PC.

##### B. Experimental Setup and Results

We applied the approaches on the Fraktur documents. We choose three volumes for the training purposes and one volume for obtaining the OCR of those documents and used them for testing. We trained on 11,239 combination of hyphenated and compound patterns. In order to evaluate our models, we must reserve a portion of 1,248 tokens for testing. Then, we evaluated our models on 2,425 tokens which are extracted from OCR’s output. The results in Table I show the performance of the best classifier decision trees with accuracy 0.97. We applied the approaches on the UWIII datasets. We split the extracted combination of hyphenated and compound patterns in training set of 6,262 and test set 932. We evaluated our models on 2,623 tokens which are extracted from OCR’s output. The results show the best performance is by using decision trees classifier 0.98.

#### V. CONCLUSIONS

In our approaches, we solved the following issues: support the Unicode and can be used for different languages and solve the problem of processing hyphenation with combined words, digits which were not handled with the other approaches. They are robust against OCR’s errors which include spell errors and substitution of symbols, i.e. dashes with hyphen due the misrecognition. We use particular features of language data for classifying type of a hyphen in OCR’s results such as: word length, length of the hyphenated tokens, capitalization, part-of-speech tagging which have not been used before, handle several various cases of hyphenation and give significant results. No OCR correction is applied before the dehyphenation.

#### REFERENCES

- [1] F. M. Liang, “Word hy-phen-a-tion by com-put-er,” in *Ph.D dissertation, Stanford University*, 1983, p. Available: <http://www.tug.org/docs/liang/>.
- [2] G. Grefenstette and P. Tapanainen, “What is a word, what is a sentence? problems of tokenization,” 1994, pp. 79–87.
- [3] E. Loper and S. Bird, “Nltk: The natural language toolkit,” in *In Proc. of the ACL on Effective Tools and Methodologies for NLP*, 2002.



# Sponsors

**ABBYY**<sup>®</sup>

MyScript  
Labs

Google<sup>™</sup>

**ITESOFT**

**e2ia**

  
**Spigraph**

**ARKHÊNUM**  
*Patrimoine du futur*

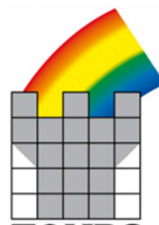
**worldline**  
e-payment services

  
**Valconum**  
CENTRE EUROPÉEN DE VALORISATION NUMÉRIQUE

**casden**   
BANQUE POPULAIRE

Région  
  
Centre

COMMUNAUTÉ D'AGGLOMÉRATION  
**Tour(s)plus**

  
TOURS

  
UNIVERSITÉ  
FRANÇOIS - RABELAIS  
TOURS

**IAPR** 

**LI**  
Laboratoire d'Informatique  
EA 6300