Electronic Letters on Computer Vision and Image Analysis 14(3):54-56, 2015

Contextual Word Spotting in Historical Handwritten Documents

David Fernández

Computer Science Department & Computer Vision Center, Edifici O/Q Campus UAB, Cerdanyola del Valles, Barcelona, Spain
Advisor/s: Josep Lladós, Alicia Fornés
Date and location of PhD thesis defense: 14 November 2014, Autonomous University of Barcelona

Received 24th July 2015; accepted 27h July 2015

Abstract

There are countless collections of historical documents in archives and libraries that contain plenty of valuable information for historians and researchers. The extraction of this information has become a central task among the Document Analysis researches and practitioners. There is an increasing interest to digital preserve and provide access to these kind of documents. But only the digitalization is not enough for the researchers. The extraction and/or indexation of information of this documents has had an increased interest among researchers. In many cases, and in particular in historical manuscripts, the full transcription of these documents is extremely difficult due the inherent deficiencies: poor physical preservation, different writing styles, obsolete languages, etc.

Word spotting has become a popular an efficient alternative to full transcription. It inherently involves a high level of degradation in the images. The search of words is holistically formulated as a visual search of a given query shape in a larger image, instead of recognising the input text and searching the query word with an ascii string comparison. But the performance of classical word spotting approaches depend on the degradation level of the images being unacceptable in many cases . In this thesis we have proposed a novel paradigm called contextual word spotting method that uses the contextual/semantic information to achieve acceptable results whereas classical word spotting does not reach.

The contextual word spotting framework proposed in this thesis is a segmentation-based word spotting approach, so an efficient word segmentation is needed. Historical handwritten documents present some common difficulties that can increase the difficulties the extraction of the words. We have proposed a line segmentation approach that formulates the problem as finding the central part path in the area between two consecutive lines. This is solved as a graph traversal problem. A path finding algorithm is used to find the optimal path in a graph, previously computed, between the text lines. Once the text lines are extracted, words are localized inside the text lines using a word segmentation technique from the state of the art.

Classical word spotting approaches can be improved using the contextual information of the documents. We have introduced a new framework, oriented to handwritten documents that present a highly structure, to extract information making use of context. The framework is an efficient tool for semi-automatic transcription

Correspondence to: <dfernandez@cvc.uab.es>

Recommended for acceptance by Jorge Bernal
DOI http://dx.doi.org/10.5565/rev/elcvia.741
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

that uses the contextual information to achieve better results than classical word spotting approaches. The contextual information is automatically discovered by recognizing repetitive structures and categorizing all the words according to semantic classes. The most frequent words in each semantic cluster are extracted and the same text is used to transcribe all them.

The experimental results achieved in this thesis outperform classical word spotting approaches demonstrating the suitability of the proposed ensemble architecture for spotting words in historical handwritten documents using contextual information.

References

[1] David Fernández-Mota, Pau Riba, Alicia Fornés, Josep Lladós. A graph-based approach for segmenting touching lines in historical handwritten documents. *IJDAR*. 2014.

[2] Josep Lladós, Marçal Rusiñol, Alicia Fornés, David Fernández-Mota, Anjan Dutta. On the Influence of Word Representations for Handwritten Word Spotting in Historical Documents. *IJPRAI*. 2012.

[3] David Fernández-Mota, Pau Riba, Alicia Fornés, Josep Lladós. On the Influence of Key Point Encoding for Handwritten Word Spotting. In *International Conference on Frontiers in Handwriting Recognition*. 2014.

[4] Pau Riba, Jon Almazán, Alicia Fornés, David Fernández-Mota, Ernest Valveny, Josep Lladós. e-Crowds: a mobile platform for browsing and searching in historical demography-related manuscripts. In *International Conference on Frontiers in Handwriting Recognition*. 2014.

[5] David Fernández-Mota, Jon Almazán, Núria Cirera, Alicia Fornés & Josep Lladós. *BH2M*: the Barcelona Historical Handwritten Marriages database. In *International Conference on Pattern Recognition*. 2014.

[6] David Fernández-Mota, R. Manmatha, Josep Lladós & Alicia Fornés. Sequential Word Spotting in Historical Handwritten Documents. In *Workshop on Document Analysis Systems*. 2014.

[7] David Fernández-Mota, Simone Marinai, Josep Lladós & Alicia Fornés. Contextual Word Spotting in Historical Manuscripts using Markov Logic Networks. In *Workshop on Historical Document Imaging and Processing*. 2013.

[8] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez, Sergi Robles, Joan Mas, David Fernández-Mota, Jon Almazán, Lluis Pere de las Heras. ICDAR 2013 Robust Reading Competition. In *International Conference on Document Analysis and Recognition*. 2013.

[9] Lluis Pere de las Heras, David Fernández-Mota, Alicia Fornés, Ernest Valveny, Gemma Sanchez, Josep Lladós. Perceptual retrieval of architectural floor plans. In *Workshop on Graphics Recognition*. 2013.

[10] Lluis Pere de las Heras, David Fernández-Mota, Ernest Valveny, Josep Lladós, Gemma Sanchez. Unsupervised wall detector in architectural floor plan. In *International Conference on Document Analysis and Recognition*. 2013.

[11] David Fernández-Mota, R. Manmatha, Alicia Fornés, Josep Lladós. On Influence of Line Segmentation in Efficient Word Segmentation in Old Manuscripts. In *International Conference on Frontiers in Handwriting Recognition*. 2012.

[12] Jon Almazán, David Fernández-Mota, Alicia Fornés, Josep Lladós, Ernest Valveny. A Coarse-to-Fine Approach for Handwritten Word Spotting in Large Scale Historical Documents Collection. In *International Conference on Frontiers in Handwriting Recognition*. 2012.

[13] David Fernández-Mota, Alicia Fornés, Josep Lladós. Handwritten Word Spotting in Old Manuscript Images Using a Pseudo-Structural Descriptor Organized in a Hash Structure. In *Iberian Conference on Pattern Recognition and Image Analysis*. 2011.