Pattern Recognition 48 (2015) 545–555

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Efficient segmentation-free keyword spotting in historical document collections

Marçal Rusiñol *, David Aldavert, Ricardo Toledo, Josep Lladós

Computer Vision Center, Dept. Ciències de la Computació, Edifici O, Univ. Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

ARTICLE INFO

Article history:
Received 23 April 2014
Received in revised form
17 June 2014
Accepted 20 August 2014
Available online 29 August 2014

Keywords:
Historical documents
Keyword spotting
Segmentation-free
Dense SIFT features
Latent semantic analysis
Product quantization

ABSTRACT

In this paper we present an efficient segmentation-free word spotting method, applied in the context of historical document collections, that follows the query-by-example paradigm. We use a patch-based framework where local patches are described by a bag-of-visual-words model powered by SIFT descriptors. By projecting the patch descriptors to a topic space with the latent semantic analysis technique and compressing the descriptors with the product quantization method, we are able to efficiently index the document information both in terms of memory and time. The proposed method is evaluated using four different collections of historical documents achieving good performances on both handwritten and typewritten scenarios. The yielded performances outperform the recent state-of-the-art keyword spotting approaches.

1. Introduction

Nowadays, in order to grant access to the contents of digital document collections, their texts are transcribed into electronic format so users can perform textual searches. When dealing with large collections, automatic transcription processes are used since a manual transcription is not a feasible solution. In the context of digital collections of historical documents, handwriting recognition strategies [1] are applied to achieve an automatic transcription since most of those documents are manuscripts. However, handwriting recognition often do not perform satisfactorily enough in the context of historical documents. Documents presenting severe degradations or using ancient glyphs might difficult the task of recognizing individual characters, and the lexicon definition and language modeling steps are not straightforwardly solved in such a context. *Keyword spotting* has become a crucial tool to provide accessibility to historical collection's contents. Keyword spotting can be defined as the pattern recognition task aimed at locating and retrieving a particular keyword from a document image collection without explicitly transcribing the whole corpus.

Two different families of keyword spotting methods can be found in the document image analysis literature. On the one hand, *learning-based* methods, such as [2–4], use supervised machine learning techniques to train models of the words that the user wants to spot. Those models are then used to classify whether an incoming document image contains or not one of the sought words. On the other hand, *example-based* methods, such as [5–7], receive as input an instance of the keyword that the user wants to retrieve from a previously indexed document image collection. Learning-based methods are preferred for applications where the keywords to spot are a priori known and fixed. If the training set is large enough they are usually able to deal with multiple writers. However, the cost of having a useful amount of annotated data available might be unbearable in most scenarios. In that sense methods running with few or none training data are preferred. It is the case of example-based methods, which are specially interesting when it is not feasible to obtain labeled data. They also present the advantage that the user is free to cast whatever query keyword he wants and is not restricted to the set of modeled words.

However, one of the main drawbacks of keyword spotting methods, either learning or example-based, is that they usually need a layout analysis step that segments the document images into words [8–11] or text lines [4,6]. But this segmentation step is not always straightforward and might be error prone. In fact, although word and text line segmentation is a quite mature research topic, it is far from being a solved problem in critical scenarios dealing with handwritten text and highly degraded documents [12,13]. Any segmentation errors affect the subsequent word representations and matching steps. This dependence on a good word segmentation motivated the researchers of the keyword spotting domain to recently move towards complete segmentation-free methods. In [14,15], Leydier et al. proposed

* Corresponding author. Tel.: +34 93 581 23 01; fax: +34 93 581 16 70.
E-mail addresses: marcal@cvc.uab.cat (M. Rusiñol), aldavert@cvc.uab.cat (D. Aldavert), ricard@cvc.uab.cat (R. Toledo), josep@cvc.uab.cat (J. Lladós).

http://dx.doi.org/10.1016/j.patcog.2014.08.021
0031-3203/© 2014 Elsevier Ltd. All rights reserved.

a word spotting methodology based on local keypoints. For a given query image, interest points are extracted and encoded by a simple descriptor based on gradient information. The word spotting is then performed by trying to locate zones of the document images with similar interest points. This retrieved zones are then filtered and only the ones sharing the same spatial configuration than the query model are returned. A similar approach using SIFT keypoints for spotting both words and graphical symbols in line drawings was presented in [16] by Rusiñol and Lladós. However, directly matching local keypoints might be too computationally expensive when dealing with large datasets, and thus researchers started to apply the bag-of-visual words (BoVW) paradigm for keyword spotting purposes. For instance, Roy et al. proposed in [17] to cluster local image features into a codebook of representative character primitives for typewritten keyword spotting. Another segmentation-free word spotting method is presented in [18] by Gatos and Pratikakis. In that case, the authors propose to use a sliding-window approach with a patch descriptor that encodes pixel densities. The hypothetic locations where the queried word is likely to appear are found by a template matching strategy. The method proposed by Almazán et al. [7] presents another sliding-window approach where local patches are represented by gradient-based descriptors and the retrieval step is performed by using an exemplar support vector machine framework. In [19], Rothacker et al. combined the use of a patch based BoVW representation with HMMs to efficiently and accurately spot keywords in handwritten documents. Finally, the recent work by Howe [20] presents a multi-writer keyword spotting method that models the possible stroke distortions by inferring a generative word appearance model. The literature dealing with segmentation-free keyword spotting methods is rather scarce since it is a relatively new and unexplored research topic. However, we strongly believe that bypassing the segmentation step is a must in the context of historical document collections where achieving a perfect word or text line segmentation is unfeasible. So, architectures that dismiss the segmentation step present a clear asset in the context of historical documents.

In addition, quite often, keyword spotting methods rely on computing expensive distances exhaustively between the query and the words in the collection such as DTW [5] or learning and applying complex models such as HMMs [2,3,19] or neural networks [4]. In that sense, in large-scale scenarios, the complexity issue should to be taken into account by proposing efficient and scalable methods both in terms of memory usage and response time.

In this paper we present an efficient segmentation-free keyword spotting method based on a BoVW model powered by SIFT descriptors in a patch-based framework. Since an explicit word segmentation is avoided, the proposed method can be applied in scenarios where word segmentation might be problematic such as documents that do not follow a classical Manhattan layout, or even be used to spot handwritten annotations that do not follow a regular text line structure. Other preprocessing steps such as binarization and slant correction are also avoided, directly processing the raw image. The proposed architecture follows the query-by-example paradigm and does not involve any supervised learning method, thus not relying on any previous content transcription. Our proposal adapts techniques that have been successfully applied in other computer vision problems to the historical documents' context. By using such general representations instead of relying on hand-crafted features, both handwritten and typewritten documents are handled indifferently.

This work is a significantly extended version of our previous conference paper [21] that introduced our proposed methodology. Specifically, we have enhanced our preliminary version by including an indexation scheme aimed to scale the proposed method to handle large datasets. A multi-length patch representation is also introduced, which increases the retrieval performance by taking into account the different possible lengths of the query words. A thorough analysis and evaluation of all involved parameters of the method is presented in order to assess the configuration maximizing the retrieval performance. Finally, a performance comparison with the recent state-of-the-art literature in keyword spotting is also presented.

The remainder of this paper is organized as follows. In Section 2, we present how the document corpora are constructed and organized. We detail the feature extraction from document pages and the encoding system used in order to efficiently query the collection. Section 3 details the retrieval stage. We show how queries are treated and how regions of interest are determined within document pages. Experimental results are presented in Section 4. We study the influence of the method's parameters and compare our performance against a number of state-of-the-art keyword spotting approaches. Finally, conclusions and further research lines are drawn in Section 5.

2. Off-line corpus representation

The word spotting problem is addressed by dividing the original document images into a set of densely sampled local patches. These local patches are the basic structure used to spot the words within the document: once a query image is given, the local patches are used to determine the page locations where the query keyword has a greater likelihood to appear. With such a procedure having an explicit word segmentation is avoided as well as any other word pre-processing steps (i.e. binarization and slant correction). These local patches must roughly match the size of the text in the document. More precisely, the height H of the local patches should roughly match the height of the text in the document. This height parameter H can be either set automatically, by for instance using a projection profile algorithm, or it can by manually set by the user.

Then, for a given height H, four different widths W_ℓ are defined in order to cope with queries of different lengths. Specifically, the geometry of the patches has been set to $H \times H$, $2H \times H$, $3H \times H$ and $4H \times H$ and is densely sampled using a regular grid of $(H/3) \times (H/3)$ pixels. The most convenient patch width will be determined at query time. This setup guarantees that there is enough overlapping between the local patches and the document words so that each word in the document is covered by at least a patch. Although a salient patch detection strategy will effectively reduce the amount of patches to be processed [18], by densely sampling them no assumption has been made on which portions of the documents are important to the final user.

2.1. Local patch descriptor

Local patches are described using the BoVW signature so that first visual words are extracted from the document images. The visual words are obtained by densely sampling SIFT descriptors over the image by using the method proposed by Fulkerson et al. in [22]. The SIFT descriptors are sampled over a regular grid of 5×5 pixels at three different scales: $H/2$, $3H/4$ and H. This multiscale representation is used to capture from fine to coarse characteristics from the word characters. The finer scale characterizes sub-parts of a character while the coarser scale characterizes whole characters and their surroundings.

The performance of the BoVW model depends on the amount of visual words extracted from the image. In the related literature it has been noted that the larger is the amount of descriptors extracted from an image, the better the performance is [23]. Therefore, a dense sampling strategy has a clear advantage over