ARTICLE IN PRESS

Contents lists available at ScienceDirect

Pattern Recognition

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

Boosting the handwritten word spotting experience by including the user in the loop

Marçal Rusiñol*, Josep Lladós

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

ARTICLE INFO

Keywords:
Handwritten word spotting
Query by example
Relevance feedback
Query fusion
Multidimensional scaling

ABSTRACT

In this paper, we study the effect of taking the user into account in a query-by-example handwritten word spotting framework. Several off-the-shelf query fusion and relevance feedback strategies have been tested in the handwritten word spotting context. The increase in terms of precision when the user is included in the loop is assessed using two datasets of historical handwritten documents and two baseline word spotting approaches both based on the bag-of-visual-words model. We finally present two alternative ways of presenting the results to the user that might be more attractive and suitable to the user's needs than the classic ranked list.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Handwritten word spotting can be defined as the task of retrieving a set of locations from document images where a given word is likely to appear without explicitly transcribing all the handwritten words. Within the field of document image analysis, handwritten word spotting has received a lot of attention and is today a quite mature research topic. The kickoff word spotting approaches applied to handwritten document images were presented in the mid-90s [27,36]. Research in this topic has been mainly motivated by the huge amounts of cultural heritage assets that are still nowadays confined in digital libraries without any effective framework providing accessibility to those contents.

We can broadly define a taxonomy of handwritten word spotting methods that distinguish two main families. The first group consists of the word spotting methods that are aimed at detecting just a closed set of predefined words. These methods usually entail a training step in which a model for each of the possible words that the user wants to spot is built. Usually, these methods are preferred in multi-writer scenarios, where the user wants to assess whether a document contains one of the predefined keywords or not. Some examples of this family are the works proposed by Rodríguez-Serrano and Perronnin in [32], by Fischer et al. [11], by Choisy [6], by Edwards et al. in [9] or Chan et al. in [5] in which Hidden Markov Models (HMM) are used to model handwritten words, or the work proposed by Frinken et al. in [13] and in [14] in which Neural Networks (NN) are used to build the models. Such methods are usually known as *learning-based* methods since they entail the use of machine learning techniques.

One the other hand, there is another set of word spotting methods which are more retrieval-oriented. In that case, given a document collection which has been indexed off-line, the user casts a word query and he wants to retrieve from the image collection the similar instances of that word. In that case there is no training stage involved and the user can query whatever word he wants. Most of the early-days works on handwritten word spotting followed this paradigm, as the seminal publication of Manmatha et al. in [27] or the work of Syeda-Mahmood [36]. Such paradigm is often known as *query-by-example* methods, and they are based on matching the word provided by the user with the rest of words in the collection. Many recent handwritten word spotting methods that follow this paradigm have been proposed such as the works by Fornés et al. in [12], Lladós and Sánchez in [25], Zhang et al. in [39], Terasawa and Tanaka in [37] or Rusiñol et al. in [34]. We target our work in the query-by-example handwritten word spotting methods.

Query-by-example handwritten word spotting methods can be understood as a particular case of Content-Based Image Retrieval (CBIR), in which given an image collection (of handwritten words in our case) and a query image we want to retrieve the most similar image in terms of contents (in our case the actual textual contents). Although these word spotting methods are a particular application of the information retrieval (IR) field, very few works have taken advantage of common strategies that have been used within the IR community for long time. A clear example is the lack of word spotting methods that include the user in the loop. Just some works like the method by Bhardwaj et al. [3] or the one by Cao et al. [4] propose to include a relevance feedback step.

* Corresponding author. Tel.: +34 93 581 4090; fax:+34 93 581 1670.
E-mail addresses: marcal@cvc.uab.es, marcalrusi@gmail.com (M. Rusiñol), josep@cvc.uab.es (J. Lladós).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.07.008

Please cite this article as: M. Rusiñol, J. Lladós, Boosting the handwritten word spotting experience by including the user in the loop, Pattern Recognition (2013), http://dx.doi.org/10.1016/j.patcog.2013.07.008

They both use Rocchio's [31] well-known relevance feedback method and they both show significant improvements when including this feedback from the user. Similar conclusions were drawn in the case of typewritten word spotting in the work presented by Konidaris et al. [21] and Kesidis et al. [19].

We present in this paper a study on the effect of taking the user into account in a handwritten word spotting framework. We test in this paper the two different approaches, namely, query fusion and relevance feedback. The former consists of asking to the user to cast several queries instead of a single one and somehow combine the results. The latter consists of retrieving the similar words from the dataset and asking to the user to provide some feedback about which results were correct and which were incorrect. This relevance feedback allows to provide an enhanced result list in a subsequent iteration. Several off-the-shelf IR methods are applied in the word spotting context. The increase in terms of precision is assessed using two datasets of historical handwritten documents and two baseline word spotting approaches both based on a bag-of-visual-words model. This paper is an extension of a previous conference version [35]. We have substantially extended its contents by proposing a new baseline method, adding four additional score normalization strategies and by finally introducing two different alternative ways of visualizing the spotting results.

The remainder of this paper is organized as follows. We overview in Section 2 the baseline handwritten word spotting methods. Section 3 is focused on the query fusion experiments whereas Section 4 deals with relevance feedback. In Section 5 we present the document image datasets and the evaluation measures. We then provide in Section 6 the experimental results. In Section 7 we propose the two alternative results visualization options. We conclude and present some discussion on Section 8.

2. Baseline bag-of-visual-words methods

In this section, we give the details of our word spotting baseline methods. Here, we assume that the words in the document pages have been previously segmented by a layout analysis step. Both the queries and the items in the database are thus segmented word snippets. The way we describe those word images is based on the bag-of-visual-words (BoVW) model powered by either SIFT [26] or Shape Context [2] descriptors. We start with a clustering of the descriptors to build a codebook. Once we have the codebook, word images are encoded by the BoVW model. In a last step, in order to produce more robust word descriptors, we add some coarse spatial information to the orderless BoVW model. Let us first detail the baseline system using SIFT features and subsequently the one using the shape context descriptor.

2.1. SIFT features

The first baseline consisting of a BoVW model powered by SIFT features was proposed in [34], and the exact parametrization we use here has been compared against a number of alternate handwritten word representations in [24]. We refer the interested reader to [24] for an exhaustive description of the representation method.

For each word image in the reference set, we densely calculate the SIFT descriptors over a regular grid by using the method presented by Fulkerson et al. in [15]. Three different SIFT descriptor scales are considered. The grid and scale parameters are dependent on the word sizes, and in our case have been experimentally set. We can see in Fig. 1 an example of dense SIFT features extracted from a word image. Because the descriptors are densely sampled, some SIFT descriptors calculated in low textured regions are unreliable. Therefore, descriptors having a low gradient magnitude before normalization are directly discarded.

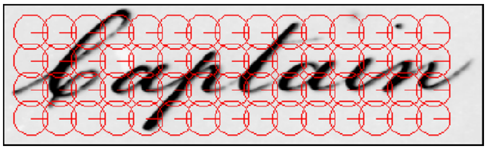

Fig. 1. Dense SIFT features extracted from a word image.

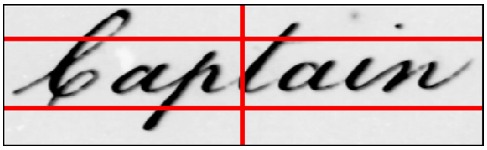

Fig. 2. Second level of the proposed SPM configuration. Ascenders and descenders information and right and left parts of the words are captured.

Once the SIFT descriptors are calculated by clustering the descriptor feature space into k clusters we obtain the codebook that quantizes SIFT feature vectors into visual words. We use the k-means algorithm to perform the clustering of the feature vectors. In this work, we use a codebook with dimensionality of $k=20{,}000$ visual words.

For each of the word images, we extract the SIFT descriptors, and we quantize them into visual words with the codebook. Then, the visual word associated to a descriptor corresponds to the index of the cluster that each descriptor belongs to. The BoVW feature vector for a given word snippet is then computed by counting the occurrences of each of the visual words in the image.

However, one of the main limitations of the bag-of-words-based models is that they do not take into account the spatial distribution of the features. In order to add spatial information to the orderless BoVW model, Lazebnik et al. [23] proposed the Spatial Pyramid Matching (SPM) method. This method roughly takes into account the word distribution over the image by creating a pyramid of spatial bins.

This pyramid is recursively constructed by splitting the images in spatial bins following the vertical and horizontal axis. At each spatial bin, a different BoVW histogram is extracted. The resulting descriptor is obtained by concatenating all the BoVW histograms. Therefore, the final dimensionality of the descriptor is determined by the number of levels used to build the pyramid.

In our experiments, we have adapted the idea of SPM to be used in the context of handwritten word representation. We use the SPM configuration presented in Fig. 2 where the two different levels are used. The first level is the whole word image and in the second level we divide it in its right and left parts and its upper, central and lower parts. With this configuration we aim to capture information about the ascenders and descenders of the words as well as information about the right and left parts of the words. Since we used a two level SPM with 7 spatial bins, we therefore obtain a final descriptor of 140,000 dimensions for each word image.

2.2. Shape context features

As a second baseline system we propose to build the BoVW model in terms of shape context features [2]. The idea of aggregate shape context descriptors into a bag-of-words representation was originally proposed by Mori et al. in [28]. Shape context descriptors have also been proven to yield good results to represent words [25].

Download English Version:

https://daneshyari.com/en/article/10360385

Download Persian Version:

https://daneshyari.com/article/10360385

Daneshyari.com