Pattern Recognition 68 (2017) 310–332

Contents lists available at ScienceDirect

Pattern Recognition

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

# A survey of document image word spotting techniques

Angelos P. Giotis[a,b,*], Giorgos Sfikas[b], Basilis Gatos[b], Christophoros Nikou[a]

[a] Department of Computer Science and Engineering, University of Ioannina, Greece
[b] Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", GR-15310 Athens, Greece

ARTICLE INFO

*Article history:*
Received 26 February 2016
Revised 5 November 2016
Accepted 21 February 2017
Available online 27 February 2017

*Keywords:*
Word spotting
Retrieval
Document indexing
Features
Representation
Relevance feedback

ABSTRACT

Vast collections of documents available in image format need to be indexed for information retrieval purposes. In this framework, word spotting is an alternative solution to optical character recognition (OCR), which is rather inefficient for recognizing text of degraded quality and unknown fonts usually appearing in printed text, or writing style variations in handwritten documents. Over the past decade there has been a growing interest in addressing document indexing using word spotting which is reflected by the continuously increasing number of approaches. However, there exist very few comprehensive studies which analyze the various aspects of a word spotting system. This work aims to review the recent approaches as well as fill the gaps in several topics with respect to the related works. The nature of texts and inherent challenges addressed by word spotting methods are thoroughly examined. After presenting the core steps which compose a word spotting system, we investigate the use of retrieval enhancement techniques based on relevance feedback which improve the retrieved results. Finally, we present the datasets which are widely used for word spotting, we describe the evaluation standards and measures applied for performance assessment and discuss the results achieved by the state of the art.

## 1. Introduction

A great amount of information in libraries and cultural institutions exist all over the world and need to be digitized so as to preserve it and protect it from frequent handling. Among others, Google has put an effort to digitize books on a large scale [1,2], thereby providing support to the document understanding research community. In order to create digital libraries which allow efficient searching and browsing for future users, thousands of digitized documents have to be transcribed or at least indexed at a certain degree. However, the automatic recognition of poor quality printed text and especially, handwritten text, is not feasible by traditional OCR approaches which mainly suffice for modern printed documents with simple layouts and known fonts. Most of the constraints encountered by recognition systems stem from difficulties in segmenting characters or words, the variability of the handwriting and the open vocabulary. For this reason, more flexible information retrieval and image analysis techniques are required.

### *1.1. Document indexing using image retrieval methods*

The actual problem behind building digital libraries lies on the retrieval of digitized documents in terms of reliable extraction and access to specific information. While a document image processing system analyzes different text regions so as to convert them to machine-readable text using OCR, a document image retrieval system searches whether a document image contains particular words of interest, without the need for correct character recognition, but by directly characterizing image features at character, word, line or even document level.

On one hand, *recognition-based* retrieval relies on the complete recognition of documents either at character level using OCR, or at word level using *word recognition* methods. In the latter case, the goal is to correctly classify a query word into a labeled class, or else, obtain its transcription. Most methods of this type require prior transcription of text-lines, words or characters to train character or word models. During the search phase, a text dictionary or lexicon is used and only words from that lexicon can be used as candidate transcriptions in the recognition task. These methods usually rely on hidden Markov models (HMMs) [3,4], conditional random fields (CRFs) [5], neural networks (NNs) [6,7] or they might follow a hybrid approach by combining different classifiers, such as support vector machines (SVMs) with HMMs [8,9] or HMMs with NNs [10]. An obvious drawback of these approaches is that they

[*] Corresponding author at: Department of Computer Science and Engineering, University of Ioannina, Greece.

*E-mail addresses:* agiotis@cs.uoi.gr (A.P. Giotis), sfikas@iit.demokritos.gr (G. Sfikas), bgat@iit.demokritos.gr (B. Gatos), cnikou@cs.uoi.gr (C. Nikou).

http://dx.doi.org/10.1016/j.patcog.2017.02.023
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

have to deal with the inherent handwriting variability and handle a large number of word and character models. Nevertheless, the scope of this work does not focus on recognition-based retrieval methods and thus, we only briefly refer to them.

On the other hand, the *recognition-free* retrieval which is also known in the literature as *word spotting* or *keyword spotting* is the main subject of this study. The goal here is to retrieve all instances of user queries in a set of document images which may be segmented at text lines or words. Actually, the user formulates a query and the system evaluates its similarity with the stored documents and returns as output a ranked list of results which are most similar to the query. The process is totally based on matching between common representations of features, such as color, texture, geometric shape or textual features, while conversion of whole documents into machine readable format and recognition do not take place at all. Therefore, the selection and use of proper features and robust matching techniques are the most important aspects of a word spotting system.

Word spotting methods may be divided into multiple categories according to various factors. Depending on how the input is specified by the user we can distinguish *query-by-example* (QBE) from *query-by-string* (QBS) methods. In the QBE scenario, the user selects an image of the word to be searched in the document collection, whereas in the QBS paradigm, the user provides an arbitrary text string as input to the system. Another way to categorize word spotting methods depends on whether training data are used offline, either to learn character and word models or tune the parameters of the system. This way we can distinguish *learning-based* from *learning-free* approaches. Finally, word spotting methods which can be directly applied to whole document pages are considered as *segmentation-free*, in contrast with *segmentation-based* methods, where a segmentation step has to be applied at line or word level during preprocessing.

Word spotting was initially proposed in the speech recognition community [11]. Its application was adopted later on for printed [12,13] and handwritten [14] document indexing. While early approaches were based on raw features extracted directly from image pixels [14,15], the state of the art is to characterize document images with more complex features based on gradient information, shape structure, texture, etc. (see Section 4.1).

### 1.2. Applications

There are a variety of applications of word spotting for document indexing and retrieval including the following:

- retrieval of documents with a given word in company files,
- searching online in cultural heritage collections stored in libraries all over the world,
- automatic sorting of handwritten mail containing significant words (e.g. “urgent”, “cancelation”, “complain”) [16],
- identification of figures and their corresponding captions [17],
- keyword retrieval in pre-hospital care reports (PCR forms) [18],
- word spotting in graphical documents such as maps [19],
- retrieval of cuneiform structures from ancient clay tablets [20],
- assisting human transcribers in identifying words in degraded documents, especially those appearing for the first time.

Although word spotting and word recognition belong to two separate retrieval paradigms, they sometimes interact by assisting one another. For instance, the authors in [21] propose a keyword spotting approach relying on a NN-based recognition system. On the contrary, in [22], word spotting contributes as a means of bootstrapping a handwriting recognition system, in terms of selecting new elements from the retrieved results. These elements can be used to augment the training set through a semi-supervised procedure, thus increasing the final recognition accuracy while at the same time avoiding the costly manual annotation process.

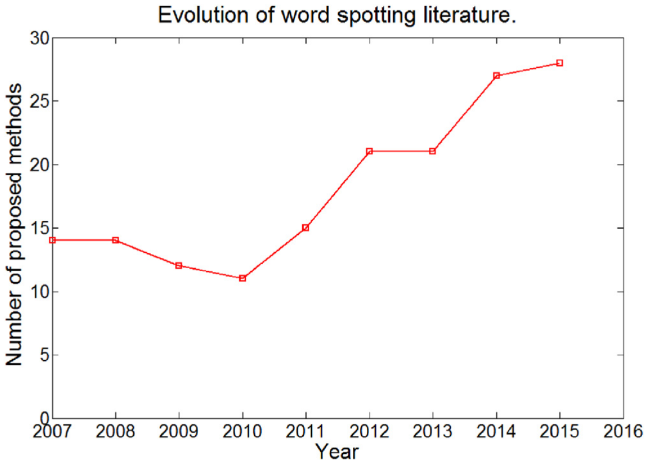

**Fig. 1.** Word spotting approaches published over the last decade.

### 1.3. Evolution of the related works

In order to track the recent literature, we present some statistics related to the evolution of word spotting methods over the last decade. The research community concentrates on indexing historical documents on a grand scale using word spotting and thus, we consider that the whole process remains an open problem. To the best of our knowledge, Fig. 1 provides a concise view of the various word spotting approaches for offline, handwritten or printed documents, which were published in conferences and journals since 2007. As it can be seen in Fig. 1, there is an increased number of papers over the past few years which confirms the growing interest of the community in word spotting.

### 1.4. Contributions and outline of the paper

Apart from the proposed methods, there also exist a number of surveys for word spotting, either for a specific script, or a particular domain (machine-printed, handwritten), or even for a variety of applications. Murugappan et al. [23] present a study for word spotting in printed documents. The authors divide the word spotting methods according to a character-based and a word-based representation depending on the features used in each case. Their work implies that character-based approaches provide satisfactory results if character segmentation is easy to obtain, whereas word-based approaches can deal with touching characters efficiently and analyze the shapes of the words without explicit character recognition. In addition, a comparative study for segmentation and word spotting methods is presented in [24] for handwritten and printed text in Arabic documents. The segmentation techniques rely on horizontal and vertical profile features and scale space segmentation. The features under comparison are geometrical moments and word profiles, whereas the similarity computation is carried out using the cosine metric and dynamic time warping (DTW). An explicit view of the various aspects of a word spotting system is presented by Marinai et al. [25]. Therein, the different features used for each technique are categorized according to the layer at which the similarity computation is applied (pixel/column features, connected components, word level features etc.). Image representations (i.e. feature vectors) with respect to the specific feature types are also analyzed along with the respective similarity measures. Finally, the work of Tan et al. [26] underlines the necessity for content-based image retrieval as an economical alternative to OCR,