Pattern Recognition 60 (2016) 596–612

Contents lists available at ScienceDirect

Pattern Recognition

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

Flexible Sequence Matching technique: An effective learning-free approach for word spotting

Tanmoy Mondal [a,*], Nicolas Ragot [a,*], Jean-Yves Ramel [a], Umapada Pal [b]

[a] Université François Rabelais, Laboratoire d'Informatique (LI EA 6300), Tours, France
[b] Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India

ARTICLE INFO

Article history:
Received 1 September 2015
Received in revised form
14 April 2016
Accepted 2 May 2016
Available online 24 May 2016

Keywords:
Flexible Sequence Matching (FSM)
Dynamic Time Warping (DTW)
Minimal Variance Matching (MVM)
Subsequence DTW (SSDTW)
Continuous Dynamic Programming (CDP)
Word spotting
Historical documents
Handwritten documents
Printed documents
George Washington dataset

ABSTRACT

In this paper, a robust method is presented to perform word spotting in degraded handwritten and printed document images. A new sequence matching technique, called the Flexible Sequence Matching (FSM) algorithm, is introduced for this word spotting task. The FSM algorithm was specially designed to incorporate crucial characteristics of other sequence matching algorithms (especially Dynamic Time Warping (DTW), Subsequence DTW (SSDTW), Minimal Variance Matching (MVM) and Continuous Dynamic Programming (CDP)). Along with the characteristics of multiple matching (many-to-one and one-to-many), FSM is strongly capable of skipping existing outliers or noisy elements, regardless of their positions in the target signal. More precisely, in the domain of word spotting, FSM has the ability to retrieve complete words or words that contain only a part of the query. Furthermore, due to its adaptable skipping capability, FSM is less sensitive to local variation in the spelling of words and to local degradation effects within the word image. The multiple matching capability (many-to-one, one-to-many) of FSM helps it addressing the stretching effects of query and/or target images. Moreover, FSM is designed in such a way that with little modification, its architecture can be changed into the architecture of DTW, MVM, and SSDTW and to CDP-like techniques. To illustrate these possibilities for FSM applied to specific cases of word spotting, such as incorrect word segmentation and word-level local variations, we performed experiments on historical handwritten documents and also on historical printed document images. To demonstrate the capabilities of sub-sequence matching, of noise skipping, as well as the ability to work in a multilingual paradigm with local spelling variations, we have considered properly segmented lines of historical handwritten documents in different languages and improperly as well as properly segmented words in printed and handwritten historical documents. From the comparative experimental results shown in this paper, it can be clearly seen that FSM can be equivalent or better than most DTW-based word spotting techniques in the literature while providing at the same time more meaningful correspondences between elements.

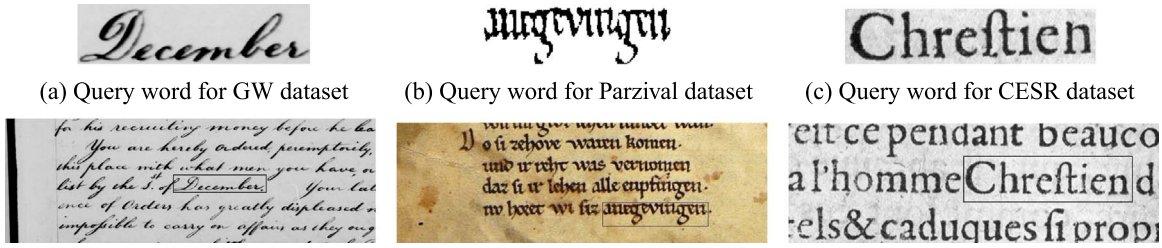
1. Introduction

Today's world of high quality document digitization has provided a stirring alternative to preserving precious ancient manuscripts. It has provided easy, hassle-free access of these ancient manuscripts for historians and researchers. Retrieving information from these knowledge resources is useful for interpreting and understanding history in various domains and for knowing our cultural as well as societal heritage. However, digitization alone cannot be very helpful until these collections of manuscripts can be indexed and made searchable. The performance of the available OCR engines highly dependent on the burdensome process of learning. Moreover, the writing and font style variability, linguistics and script dependencies and poor document quality caused by high degradation effects are the bottlenecks of such systems. The process of manual or semi-automatic transcription of the entire text of handwritten or printed documents for searching any specific word is a tedious and costly job. For this reason research has been emphasized on word spotting. This technique can be defined as the: "*localization of words of interest in the dataset without actually interpreting the content*" [1], and the result of such a search could look like the result shown in Fig. 1 (without transcription). These figures (Fig. 1) demonstrates a layman's view of the word spotting outcome of the system.[1]

* Corresponding authors.
E-mail addresses: tanmoy.mondal@etu.univ-tours.fr (T. Mondal), nicolas.ragot@univ-tours.fr (N. Ragot), jean-yves.ramel@univ-tours.fr (J.-Y. Ramel), umapada@isical.ac.in (U. Pal).

[1] For a detailed description of the datasets, please see the experimental evaluation Section 4. For a detailed description of the Parzival dataset, please see: ⟨http://www.iam.unibe.ch/fki/databases/iam-historical-document-database/parzival-database⟩.

http://dx.doi.org/10.1016/j.patcog.2016.05.011
0031-3203/



(a) Query word for GW dataset (b) Query word for Parzival dataset (c) Query word for CESR dataset

(d) Sample page of GW dataset (e) Sample page of Parzival dataset (f) Sample page of CESR dataset

Fig. 1. Example of word spotting outputs (marked by rectangular boxes in (d), (e), (f)) corresponding to queries ((a), (b), (c)) from document images extracted from 3 different datasets (see Section 4). The spotted query words on the complete document page are marked by a rectangular box. (a) Query word for GW dataset. (b) Query word for Parzival dataset. (c) Query word for CESR dataset. (d) Sample page of GW dataset. (e) Sample page of Parzival dataset. (f) Sample page of CESR dataset.

A popular way to categorize word spotting techniques is to consider those that are based on *query-by-example* and those that are based on *query-by-string*. In the former category, a region of a document is defined by the user, and the system should return all of the regions that contain the same text region, that is the same as the region defined by the user. These are often achieved by learning-free, image-matching-based approaches. For approaches that belong to the *query-by-string* category, queries of arbitrary character combinations can be searched. These approaches require a model for every character. Consequently, they are often achieved by learning-based approaches, such as HMM [2,3] or a Bidirectional Long Short-Term Memory (BLSTM) neural network [4]. These approaches allow us to obtain very good performance when the learning set is representative of the writing/font styles that are found in the document to be searched. The well-known drawback of learning-based approaches is the requirement of a set (most often enormous) of transcribed text line images for training, which could be costly to obtain for some of the historical datasets. Only very few approaches appear to be able to work with a low level of training data [5,6]. Moreover, the training (transcription of the learning set and learning of models) could have to be re-performed for new documents, depending on the variability of the writing/font styles. Thus, if neither the language nor the alphabet of a historical document are known or if creating a new learning set and retraining the system is necessary but not possible, a learning-free approach to word spotting might be the only available option. Consequently, a fair comparison between the two approaches is difficult to perform without including these criteria and we decided in this study to focus on learning-free approaches. These approaches can be further categorized depending on the level of segmentation.

1.1. Segmentation-based word spotting methods

The concept of word spotting as the task of detecting words in document images without actually understanding or transcribing the content, was initially the subject of experimentation by Manmatha et al. [1]. This approach relies on the segmentation of full document images into word images. A general and highly applicable approach for comparing word images is to represent them by a sequence of features, which are extracted by using a sliding window. These word images can be thought of as a 2D signal, which can be matched using dynamic programming [1,7] based approaches. Some methods that were oriented toward bitwise comparison of images were also investigated [8], as well as holistic approaches that describe a full image of words [9,10]. An approach based on low-dimensional, fixed-length representations of word images, which is fast to compute and fast to compare, is proposed in [11]. Based on the topological and morphological information of handwriting, a skeleton based graph matching technique is used in [12], for performing word spotting in handwritten historical documents. There have also been some attempts to spot words on segmented lines to avoid the problems of word segmentation. Indeed, depending on the document quality, line segmentation could be comparatively easier than word segmentation. The partial sequence matching property of CDP [13] is one possibility. Using over segmentation is also an alternative as in [14], where the comparison of sequences of primitives obtained by segmentation and clustering (corresponding to similar characters or pieces of characters) is investigated.

The necessity of proper word segmentation (or line segmentation in some cases) and the high computational complexity of matching are critical bottlenecks of most of the techniques in this category. Moreover, these techniques are prone to the usual degradation noise that is found in historical document images. Most of them cannot spot out-of-vocabulary words.

1.2. Segmentation free word spotting methods

In [15], the authors proposed another type of matching technique based on differential features to match only the informative parts of the words, which are detected by patches. A common approach for segmentation-free word spotting is to consider the task as an image retrieval task for an input shape, which represents the query image [16]. For example, a HOG descriptors based sliding window is used in [17] to locate the document regions that are the most similar to the query. In [18], by treating the query as a compact shape, a pixel-based dissimilarity is calculated between the query image and the full document page (using the Squared Sum Distance) for locating words. A heat kernel signature (HKS) based technique is proposed in [19]. By detecting SIFT based key points on the document pages and the query image, HKS descriptors are extracted from a local patch that is centered at the key points. Then, a method is proposed to locate the local zones that contain a sufficient number of matching key points that corresponds to the query image. Bag of visual words (BoVW) based approaches were also used to identify the zones of the image that share common characteristics with the query word. In [20], the Longest Weighted Profile based zone filtering technique is used from BoVW to identify the location of the query words in the document image. In [21], local patches powered by SIFT descriptors are described by a BoVW. By projecting the patch descriptors to a topic space with a latent semantic analysis technique and compressing the descriptors with a product quantization method, the approach can efficiently index the document information both in terms of memory and time. Overall, segmentation-free approaches can overcome the curse of the segmentation problems, but they have comparatively low accuracy (in comparison with segmentation-based and learning-based approaches) and a high computational burden, considering the full image