Pattern Recognition 45 (2012) 4225–4236

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

A synthesised word approach to word retrieval in handwritten documents

Y. Liang, M.C. Fairhurst, R.M. Guest *

School of Engineering and Digital Art, University of Kent, CT2 7NT, UK

ARTICLE INFO

Article history:
Received 20 December 2011
Received in revised form
25 May 2012
Accepted 29 May 2012
Available online 15 June 2012

Keywords:
Handwriting analysis
Digital archives
Handwritten word retrieval
Word spotting
Information retrieval
Handwriting recognition
Historical manuscript analysis

ABSTRACT

Recent technological advances have enhanced the computer-based indexing and searching of digitised printed books. The performance now achievable in this domain, however, does not at present extend to handwritten texts which inherently contain more significant letter-based variation within their content. Furthermore, in most studies that address the handwritten text retrieval problem, a large training dataset is required which, very often, influences the context and search lexicon. In this paper a novel method is described to overcome the training data problem using a character-based modelling (termed *grapheme spectrum*) approach and a word modelling technique (termed *synthesised word*) enabling the retrieval of keywords that have not explicitly been seen in the training set. When tested on an illustrative historical manuscript the performance of the proposed word retrieval technique shows a clear advantage over existing methods.

1. Introduction

The past decade has seen a growth in the digitisation of archive books, texts and manuscripts as a means of preserving such documents, conserving historical heritage and enhancing access to often rare and fragile objects. Storage of manuscripts in electronic form also enables an enhanced ability to search and index content, the accuracy of which relies on the ability to automatically identify content within a document. Compared to optical character recognition applied to printed text, recognition of handwritten text is impaired by the (often considerable) human behavioural variation found in handwriting (both from the same writer and between writers) [1].

The terms *Word Retrieval* and *Word spotting* [2] are used interchangeably in literature referring to a collection of techniques which aim to provide solutions for retrieving keywords in handwritten document images. The distinction between the two terms is made in [3]: in a word spotting application, the query is an instance of the word extracted from the image of the target document, whereas in word retrieval, the query is initially in the form of an ASCII representation of the word and is subsequently translated to a feature-based model.

In word spotting approaches [4–6], the process usually relies on stored instances of the query word, and searches for best-matching word segmentations within the page images. Using this type of approach, although encouraging retrieval rates are shown in much of the available work reported, the fact that training samples (and usually many such samples) of the query word must be obtained is a significant constraint that reduces the viability and usability of the technique. As an alternative to stored instances of entire words, an instance of the query word segmented from the target document by the user [7], or a handwritten instance of the query word provided by the user imitating the target document's handwriting style [8] is accepted as the template in other studies.

With word retrieval techniques [3,9–12], a feature-based representation of the query word is generated using character level training samples. The algorithm thus searches for segmentations within the page images that best match this representation in feature-space. Approaches in this category have the potential to retrieve "out-of-vocabulary" (OOV) words of which samples have not been included in the training process or provided as templates of the query.

Word retrieval being the general goal of our current study, aiming to address the viability issue, we further define a specific context within which the technique is employed. We observe that for a word retrieval tool to be useful for users, in addition to its accuracy rate, it must fulfil other usability conditions, such as the following:

(1) It should require a minimum amount of training data, and we hypothesise that tens of samples for each character is within the tolerance of usability, because a training dataset of this size can usually be collected from one to two pages of a manuscript.
(2) It is capable of searching for words that are unseen in the training dataset. This condition indicates that the user does

* Corresponding author. Tel.: +44 0 1227 823717; fax: +44 0 1227 456084.
E-mail address: r.m.guest@kent.ac.uk (R.M. Guest).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.05.024

not need to find an instance of the query word in order to carry out the search.

Within this well-defined context, a novel method is proposed for the retrieval of keywords not explicitly appearing in the training set using a character-based modelling (termed *grapheme spectrum*) approach and a word modelling technique (termed *synthesised word*).

In this paper, we will first present a review of previous work on related topics in Section 2. The proposed method will be described in Sections 3 and 4, and the experiments carried out and results obtained will be described in Section 5. Some conclusions are drawn and future research discussed in Section 6.

2. Related work

According to [3], whether or not an image of a query word is specifically required for the training process distinguishes between word-spotting and word-retrieval techniques. The majority of related work belong to the word-spotting category, including [4–6,13].

The work reported in [7,8] are similar to the approaches mentioned above in the sense that an image of the query word must be provided prior to performing the retrieval function. A notable difference is, however, a training process with labelled data is not required. The matching algorithm employed in [7] compares the feature-based representation of the provided instance of the query word against those of the zones-of-interest within the page images. In [8] the page images are decomposed into connected-components (CCs) which are clustered to form a library, and the query word, represented by an image, is compared directly with the entries in the library of CCs. Both methods perform very well in their context. When tested on the GW20 dataset (containing 20 pages taken from George Washington's manuscripts) using 15 keywords, a precision of 60% was achieved in [7] where an equal recall rate is found. In [8] the word spotting rate using a sample of the word from the original manuscript is reported to be 95% when only the top matching cluster of BCCs (A BCC, basic connected component, is the representative image of all the CCs in a cluster) is considered. The performance based on a user-provided handwritten query is an F-measure [14] of 52% assessed under the same condition. F-measure is a common metric in Information Retrieval that assesses both the precision and the recall [14], resulting in a value varying between 0% and 100%. A higher value of F-measure is associated with a more precise performance of the system. The performance reported in [8] is encouraging. However, the result is possibly over-optimistic because a word is considered "found" as long as the target BCC is found within the best matching cluster, which in fact consists of a number of BCCs which in turn may correspond to different words. An advantage of this work is that it eliminates the need for word segmentation because the connected components in the testing dataset are not grouped into possible words. However, in this method the user must find at least one instance of the query word or produce an imitation of the query word in order to retrieve other instances. In the context of our research, we emphasise the advantage that the user does not need to present an exact sample of the query word before retrieving it. Also, the work in [8] implicitly assumed that the handwritten words are separated (no overlapping ink) which is not necessarily true with handwritten documents.

In the word retrieval approaches reported in [9–12], the recognition model is trained using character images, or even components below the complete character level (fragments into which word images are decomposed according to a set of heuristic rules), and the class in the recognition problem is at a character level. This type of system can potentially recognise OOV words, which have not been seen in the training dataset. In [10,11] character recognition models are established using a publicly available English character database. The word recognition model in [10,11] consists of the probability of individual connected components belonging to the characters in the query word. Although the model can potentially retrieve OOV words, the main goal of this work is to integrate a word segmentation probabilistic model into the word recognition model, and the reported performance is an average precision of 30%.

In the work reported in [9], the training process comprises of two stages. In the first stage, a dataset consisting of 32 samples for each character class is collected through manual character segmentation. The character recogniser generated as a result of stage one is employed in stage two for automated character segmentation for the purpose of boosting the character data samples. The individual character images as a result of the automatic segmentation are then presented to the training of a HMM word recogniser. When tested with 20 pages of George Washington's letters (GW20 database [6]) using 20-fold cross-validation, this approach has achieved a recognition rate of 84% for words within the lexicon of training samples, and 32% for OOV words.

The work reported in [12] represents a significant advantage over the other work falling into this category: The word retrieval method in [12] is based on a handwriting recognition model developed in the authors' previous work [15,16]. The system takes images of text lines alongside an associated transcription as an input to the training process. Therefore, using this method, no manual word or character segmentation is required. In the feature extraction stage, nine geometric measurements are taken from a vertical one-pixel-wide window which is slid across the text line. At each window position, the probability that a window corresponds to each character is evaluated, using not only the features extracted from current window but also those from adjacent windows. The recognition of a text line is the result of maximising the combined probabilities of all windows within the text line. Following a training phase, the system is able to perform recognition of handwritten text within each text line image. The system was initially trained on 1539 multi-author pages from the IAM database [17]. When performing handwriting recognition tasks on manuscripts written by a specific hand, the initial model is retrained on writing samples of the target author. When retrained and tested on the GW20 database using a four-fold cross validation, the system achieved an average precision of 86%. Although the method is capable of retrieving OOV words, the chosen keywords all have samples in the training subset.

Another important contribution to the field of word retrieval is reported in [3], where a document model is defined by a combination of an alphabet, a glyph book and an (optional) grammar. Each symbol in the alphabet represents a character or a set of characters, and is attached with a varying number of graphemes defined in the glyph book, along with a number of allographic and linguistic features. The contribution made in [3], compared to the authors' previous work [7] is primarily the formulation of the query word using the alphabet model, which enables the user to retrieve a word without providing an instance extracted from the target document. Testing on a Latin manuscript (MS14) [3,7], the system achieved a precision of 54% where an equal recall was reached, and 68% on an Arabic manuscript (MS6191) [3].

Whilst sharing the same objective of word retrieval specifically in terms of the retrieval of OOV words with the studies highlighted above [3,9–12], our work differs in its attempt to investigate the possibility of establishing a word retrieval system using a very small dataset for the training process.