Pattern Recognition 47 (2014) 1039–1050

Contents lists available at ScienceDirect

Pattern Recognition

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

Statistical script independent word spotting in offline handwritten documents

Safwan Wshah *, Gaurav Kumar, Venu Govindaraju

Department of Computer Science and Engineering, University at Buffalo, 113 Davis Hall, Amherst, NY 14260-2500, United States

ARTICLE INFO

Available online 10 October 2013

Keywords:
Script independent
Keyword spotting
Hidden Markov models

ABSTRACT

We propose a statistical script independent line based word spotting framework for offline handwritten documents based on Hidden Markov Models. We propose and compare an exhaustive study of filler models and background models for better representation of background or non-keyword text. The candidate keywords are pruned in a two stage spotting framework using the character based and lexicon based background models. The system deals with large vocabulary without the need for word or character segmentation. The script independent word spotting system is evaluated on a mixed corpus of public dataset from several scripts such as IAM for English, AMA for Arabic and LAW for Devanagari.

1. Introduction

The recognition of unconstrained offline handwritten documents has been a major area of research during last decades. Due to large variability in writing styles and huge vocabulary, the problem is still far from being completely solved [22,34]. As a result, word spotting has been proposed as an alternative of full transcription to retrieve keywords from document images [27]. The inputs to a word-spotting system are document sets or databases and an element denoted as query, the output is a set of images or sub-images from the database that are relevant to the query, making it similar to the classical information retrieval system. Word spotting finds its application in many areas such as information retrieval and indexing of handwritten document that are made available for searching and browsing [9]. An extensive number of multilingual handwritten documents and forms are sent every day to companies for processing [6]. An efficient retrieval system for these documents has the advantage of saving these companies time and money. As another example, recently, majority of libraries around the world have digitized their valuable handwritten books transcribed in many scripts ranging from old ancient to modern ages. An application of word spotting is to make these books searchable.

The optimum trend in word spotting systems is to propose methods that show high accuracy, high speed and work on any language with minimum preprocessing steps such as preparing the query format or word segmentation. Our goal in this work is to develop approaches to improve the word spotting performance in handwritten documents by simulating any keyword, even those unseen in the training corpus, and effectively deal with large background vocabulary without the need for word or character segmentation, and to make it scalable over many languages such as English, Arabic, and Devanagari. We will also evaluate the performance of the system and compare it to current approaches. In the proposed method, we assume minimal preprocessing during training and validation. Our method does not require segmentation or lines into words because it works on the line as one unit. The required keyword and non-keyword models are generated at run time to look for certain key word with minimal preprocessing step.

In this work we elaborate on our work [36] where we have introduced the script independent word-segmentation free keyword spotting framework based on Hidden Markov Models (HMMs). This framework is scalable across multiple scripts. We learn HMMs of trained characters and combine them to simulate any keyword, even those unseen in the training corpus. We use filler models for better representation of non-keyword image regions avoiding the limitations of line-based keyword spotting technique, which largely relies on lexicon free score normalization and white space separation. Our system is capable of dealing with large background vocabulary without the need for word or character segmentation and is scalable over many languages such as English, Arabic and Devanagari. The main characteristic of the proposed approach is utilizing script independent methods for feature extraction, training and recognition.

This work is a detailed illustration in terms of framework setup and detailed evaluation of the proposed technique. Some of the key attributes such as feature extraction and different filler and background models proposed in this work are evaluated extensively along with an analysis of system complexity. In the

* Corresponding author. Tel.: +1 716 587 1594.
E-mail addresses: safwan.weshah@gmail.com, srwshah@buffalo.edu (S. Wshah).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.09.019

experimental evaluation section, the system has been evaluated on individual and mixed public datasets of English, Arabic and Devanagari manuscripts using different sizes of the keyword list. The proper modeling of the filler and background models has been investigated and the system results compared with the model presented by Fischer et al. [13] on English, Arabic and Devanagari showing better performance over all the languages.

The rest of the paper is structured as follows. We present related work and categorize existing approaches in Section 2. In Section 3 we propose our approach including image preprocessing, feature extraction and spotting framework. We discuss the complexity of our system in Section 4 and experimental evaluation in Section 5.

2. Related work

One of the first word spotting approaches for document images was proposed by [19]. Since then, many word spotting approaches have been proposed. Mainly, word spotting approaches are divided into two types based on the query input: query-by-example and query-by-string [25]. The query-by-example or template based approach requires images that are hard to prepare and may not exist in the training set. In the template based approach, input image is matched to a set of template keyword images and the outputs are the images most similar to the query image. The image is represented as a sequence of features and usually compared with dynamic time warping (DTW) technique [24,19,32]. The main advantage of this approach is that there is minimum learning involved. However, there are limitations of dealing with wide variety of unknown writers [13]. There are certain segmentation free techniques such as Rusinol et al. [26], Leydier et al. [18] that work at document level detecting interest points using gradient or scale-invariant transform features. The query by example is not focus of our work.

The query-by-string refers to the word-spotting techniques where the input is the string that needs to be located [7,5]. The query-by-string is more complicated than query-by-example. In the query-by-string the keyword models need to be created even though no samples of that keyword exist in the training set. We further categorize the query by string techniques based on processing level as below.

2.1. Word recognition based spotting

In word based spotting such as Rodrguez-Serrano and Perronnin [25], Saabni and El-Sana [27], the HMM model for each keyword is trained separately. The score of each HMM is normalized with respect to the score of the same topology HMM trained for all non-keywords. This approach relies heavily on perfect word segmentation and requires several samples for each keyword in the training set. In a similar way [8,33], use word segmentation free technique to train character models to build the keywords as well as non-keywords. The drawback of their approach is the confidence measure with respect to a general non-keyword model that represents everything but keywords as well as the dependence on the white space model to segment the words.

2.2. Line recognition based spotting

In the line based approach, the word or character segmentation step is done during the spotting process. Chan et al. [4], Edwards et al. [7] train character HMMs from manually segmented templates assuming small variation in data. Fischer et al. [13] proposed a line level approach using HMM character models under the assumption that no more than one keyword can be spotted in a given line. Their approach outperformed the template based methods for single writer with few training samples and multi-writers with many training samples. A major drawback in their approach is the dependency on the white space to separate keywords from rest of the text. This not only has a large influence on the spotting results but also prevents the system from being scalable over other languages such as Arabic in which the space could be within or between the words revealing little information about the word boundaries [4]. Besides, the lexicon free approach to model the non-keyword has large negative effect on their system performance as well.

Frinken et al. [15] proposed a neural network-based spotting system. It parses the line to recognize the sequence of the characters and maps each character's position and its probability. It then takes the sequence of the character probabilities, a dictionary, and a language model and computes a likely sequence of words. The drawback of this approach is the dependency on the recognition system. In addition, increasing the number of the keywords increases the accuracy due to the use of an efficient language model based on a big dictionary, making it more like a

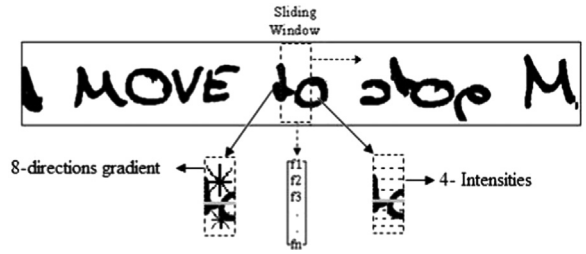

Fig. 1. Feature extraction using a sliding window.

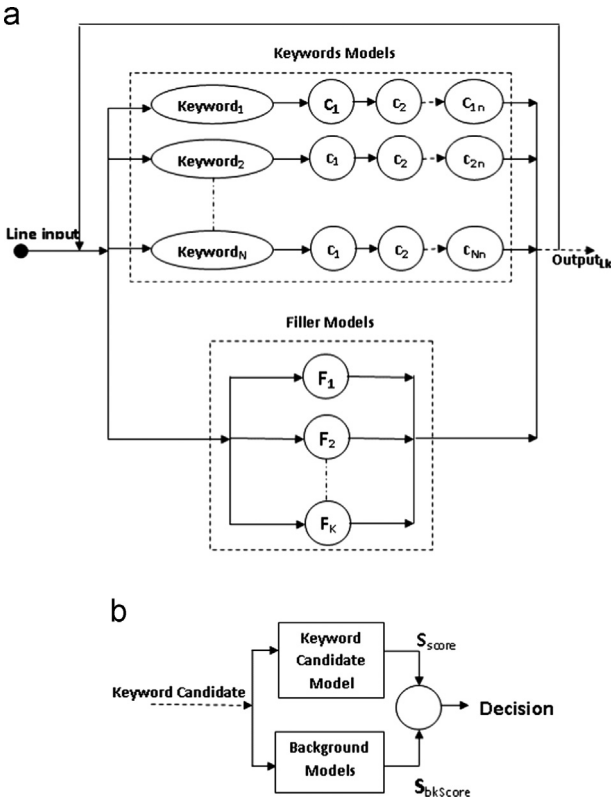

Fig. 2. Main model, (a) keywords and filler models, (b) score normalization using background models.