Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 1563–1571
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Offline recognition of omnifont Arabic text using the HMM ToolKit (HTK)

M.S. Khorsheed *

King AbdulAziz City for Science and Technology (KACST), P.O. Box 6086, Riyadh 11442, Saudi Arabia

Received 28 November 2005; received in revised form 19 March 2007
Available online 31 March 2007

Communicated by O. Siohan

Abstract

This paper presents a cursive Arabic text recognition system. The system decomposes the document image into text line images and extracts a set of simple statistical features from a narrow window which is sliding a long that text line. It then injects the resulting feature vectors to the Hidden Markov Model Toolkit (HTK). HTK is a portable toolkit for speech recognition system. The proposed system is applied to a data corpus which includes Arabic text of more than 600 A4-size sheets typewritten in multiple computer-generated fonts.

Keywords: Document analysis; Pattern analysis and recognition; Machine vision; Arabic OCR; HTK

1. Introduction

Among the branches of pattern recognition is the automatic reading of a text, namely, text recognition. The objective is to imitate the human ability to read printed text with human accuracy, but at a higher speed.

Most optical character recognition methods assume that individual characters can be isolated, and such techniques, although successful when presented with Latin typewritten or typeset text, cannot be applied reliably to cursive script, such as Arabic. Arabic language provides a rich source of technical challenges for recognition algorithms. The most obvious characteristics of the Arabic language is that Arabic scripts are inherently cursive; writing isolated characters in 'block letters' is an unacceptable and unused writing style. The shape of the letter is context sensitive, depending on its location within a word. For example a letter as 'ه' has four different shapes: isolated 'ه' as in 'عبده', beginning 'هـ' as in 'هند', middle 'ـهـ' as in 'بهر' and end 'ـه' as in 'منه'. Certain character combinations form new ligature shapes which are often font dependent. Some ligatures involve vertical stacking of characters, see Fig. 1. Since not all letters connect, word boundary location becomes an interesting problem, as spacing may separate not only words but also certain characters within a word.

Cursiveness, besides other characteristics of Arabic language, has obliged researchers to examine some obstacles which have only recently been addressed by researchers of other languages. These obstacles have played an important role in delaying character recognition systems for Arabic language compared to other languages such as Latin and Chinese. Moreover, the absence of support utilities such as a language corpus and electronic dictionaries implies that each researcher or a researching group has to build a separate lexicon. This makes it not possible to set a comparison between various algorithms since each one of them is implemented on a different data set.

There are two approaches to tackling the problem of cursiveness of Arabic script: the global approach and the analytical approach. The global approach treats the word as a whole. Features are extracted from the unsegmented word and compared to a model (Khorsheed and Clocksin,

* Tel.: +966 50 299 6 299; fax: +966 4813728.
E-mail address: mkhorshd@kacst.edu.sa

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.03.014

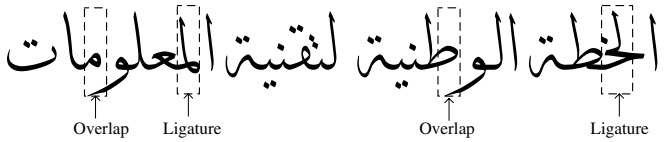

Fig. 1. A sample of Arabic text showing ligatures and overlaps.

2000). The analytical approach decomposes the word into smaller units (Hassibi, 1994), which may or may not correspond to characters. Previous research on Arabic text recognition has confirmed the difficulties in attempting to segment Arabic words into individual characters (Khorsheed, 2002).

Hidden Markov Models (HMMs) (Rabiner and Juang, 1993) are among other classification systems that are used to recognise character, word or text. They are statistical models which have been found extremely efficient for a wide spectrum of applications, especially speech processing. This success has motivated recent attempts to implement HMMs in character recognition whether on-line (Kim et al., 1997) or off-line (Kim and Park, 1996). The HMM provides an explicit representation for time-varying patterns and probabilistic interpretations that can tolerate variations in these patterns.

In off-line recognition systems, the general idea is to transform the word image into a sequence of observations. The observations produced by the training samples are used to tune the model parameters whereas those produced by the testing samples are used to investigate the system performance.

Some researchers made a critical assumption that the word image is already segmented and thus their research dealt with segmented characters (Kundu et al., 1989). In this case, each state in the model represents a letter in the alphabet set, and each feature vector is equivalent to one observation. Others (Dehghan et al., 2001) segmented the input word image into a sequence of segments in which an individual segment might be a complete character, a partial character, or joint characters. The HMM parameters were estimated from the lexicon and the training image segments. This segmentation may also be applied to the contour (Elyacoubi et al., 1999) to find segmentation points, then to extract the features from these segments and transfer the feature vectors into an observation sequence. Segmentation can be avoided if the skeleton of the word is decomposed into small strokes of which each is transformed into a feature vector and then an observation. Processing the original word image directly is another alternative (Mohamed and Gader, 1996). This system combined segmentation-free and segmentation-based techniques. The segmentation-free technique constructed a continuous density HMM for each lexicon string. The segmentation-based technique used dynamic programming to match word image and strings.

Hidden Markov modelling is suitable for 1D time sequential signals such as speech. The image signal has two dimensions. The 2D nature of the OCR problem is a fundamental difference between text and speech recognition problems. This justifies the conclusion that extending a 1D-HMM to a 2D-HMM can achieve greater advantages. Fully connected 2D-HMMs would lead to a recognition algorithm of exponential complexity (Agazzi and Kuo, 1993). To remedy this, the connectivity of the network can be reduced. This results in pseudo 2D-HMMs which are successfully implemented to recognise typewritten characters (Agazzi and Kuo, 1993) and handwritten characters (Park and Lee, 1998). 2DHMM was also implemented to recognise typewritten and handwritten Arabic words (Miled and Amra, 2001).

Abdelazim and Hashish (1989) applied HMMs to recognise Hindu numerals that are used with Arabic text; '٠١٢٣٤٥٦٧٨٩ – 0123456789'. Each numeral was represented with a separate HMM. The observation sequence was passed to all the 10 models and assigned to the numeral with the highest model probability of the observation sequence $P(O|\lambda)$. Amin and Mari (1989) implemented HMM to enhance the recognition rate at post processing stage. In (Allam, 1995), the contour of the word image was segmented after baseline estimation. This resulted in a sequence of labels, the latter of which was classified by finding the HMM which gave the highest probability. To reduce the computation time and enhance the recognition rate, a segment was not compared to the whole set of models it was rather compared to a selected group according to the position of that segment within the word. Alma'adeed et al. (2002) developed a handwritten word recognition system that is based on a model discriminant HMM. They created about 10,000 Arabic word database were each word in the lexicon is represented by a distinct model. Other ongoing research (Bazzi et al., 1999) depends on the estimation of character models, a lexicon, and grammar from training samples. The training phase takes scanned lines of text coupled with the ground truth, the text equivalent of the text image, as input. Then, each line is divided into narrow overlapping vertical windows from which feature vectors are extracted. The character modelling component takes the feature vectors and the corresponding ground truth and estimates the character models. The recognition phase follows the same step to extract the feature vectors which are used with different knowledge sources estimated in the training phase to find the character sequence with the highest likelihood $P(O|\lambda)$.

This paper presents a HMM-based system to recognise cursive Arabic text offline. Statistical features are extracted from the text line image and fed to the recogniser. The system is built on the Hidden Markov Models Toolkit (HTK) (Young et al., 2001). This is primarily designed for building HMM-based speech processing tools in particular recognisers. The proposed system depends on the technique of character models and grammar from training samples. It is lexicon free; uses the basic units during recognition. This offers an open vocabulary recognition, however, this may come at the expense of increased error.