Available online at www.sciencedirect.com

SCIENCE DIRECT®

Engineering Applications of Artificial Intelligence 19 (2006) 29–39

Engineering Applications of
ARTIFICIAL INTELLIGENCE

www.elsevier.com/locate/engappai

Classifiers combination and syntax analysis for Arabic literal amount recognition

Nadir Farah*, Labiba Souici, Mokhtar Sellami

Laboratoire de recherche en Informatique: LRI Laboratory University, Badji Mokhtar BP12, 23200 Annaba, Algérie

Received 30 March 2004; accepted 11 May 2005
Available online 29 June 2005

Abstract

Automatic handwriting recognition has a variety of applications in real world problems, such as mail sorting and check processing. Recently, it has been demonstrated that combining the decisions of several classifiers and integrating multiple information sources can lead to better recognition results. This article presents an approach for recognizing handwritten Arabic literal (legal) amounts. The proposed system uses a set of holistic structural features to describe the words. These features are presented to three classifiers: multilayer neural network, k nearest neighbor, and fuzzy k nearest neighbor. The classification results are then combined using several schemes; we retained the score summation one for this work. A syntactic post-classification process is then carried out to find the best match among the candidate words. The performance of this approach is superior to the system which ignores all contextual information and simply relies on the recognition scores of the recognizers.

Keywords: Arabic word recognition; Holistic approach; Multiclassifiers; Sum combination; Syntactic analysis

1. Introduction

Handwritten recognition is a very active research domain that led to several works, e.g. for the Latin writing (Madhvanath and Govindaraju, 2001; Steinherz et al., 1999; Suen, 1998). The current systems tendency is oriented toward the classifiers combination (Zouari et al., 2002; Xu et al., 1992), and the integration of multiple information sources.

Many systems attempted handwritten recognition; some of them have based their attempts on a special type of writing where characters are separated. Nowadays, the research is oriented toward the cursive writing that is more complex to process.

These last years, a number of papers which analyze the work done on Arabic characters/words recognition have appeared (Amin, 1998; Al Badr and Mahmoud, 1995; Essoukhri Ben Amara, 2002). Some obstacles have played an important role in delaying character/word recognition systems for Arabic language compared to other languages such as Latin and Chinese. Among these obstacles we can find the special morphology of Arabic writing, and the lack of communication between researchers in this field.

In this article we are interested in off-line handwritten Arabic word recognition, using a limited lexicon. In this direction some works moved toward Markov models (Pechwitz and Maergner, 2003), other toward neural models (Snoussi et al., 2002) or toward the neuro-symbolic systems (Farah et al., 2004; Souici-Meslati and Sellami, 2004).

Our approach is inspired by the human reading process that considers the global word shapes (Madhvanath and Govindaraju, 2001) and uses contextual knowledge based on the considered document syntax.

*Corresponding author. Tel.: +213 38 87 29 91; fax: +213 38 87 29 04.
E-mail address: farahnadir@hotmail.com (N. Farah).

0952-1976/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2005.05.005

Our work leads to the realization of a handwritten Arabic literal amount recognition system, based on a global approach, using structural high-level features (ascenders, descenders, loops, etc.).

The recognition is performed by a multiclassifiers system (Ruta and Gabrys, 2000), composed of three modules, a multilayer perceptron, a *k* nearest neighbor classifier, and a fuzzy *k* nearest neighbor one.

A combination module is used on the outcome results by classifiers to compensate for individual classifier weaknesses, and a post classification step permits to validate the combiner propositions.

The remainder of this paper is organized as follows. In Section 2, Arabic writing characteristics are presented and some works on Arabic word recognition are summarized in Section 3, then a brief overview of the system architecture is done in Section 4. Sections 5 and 6 present preprocessing and features extraction. The three individual classification systems are described in Section 7 and their results in Section 8. A combination approach of classifiers is introduced in Section 9, then the post-classification in Section 10. The paper concludes with discussion of the results and an outlook to future work.

2. Arabic writing characteristics

The Arabic language is very rich and difficult, by its structure and possibilities. Arabic script is written from right to left. It starts from the right-most position of the page toward the left in a cursive way. The Arabic alphabet consists of 28 basic characters.

The shape of the character is context sensitive, depending on its location within a word. A letter can have four different shapes: isolated, at the beginning, in the middle, at the end. Some Arabic characteristics are particular, we can find for example:

- 10 of them have one dot
  ز,ض,ب,ن,ظ,ذ,ف,غ,خ,ج
- 03 of them have two dots
  يـ,تـ,قـ
- 02 have three dots
  شـ,ثـ
- Several characters present loops
  و,ة,ـعـ,ق,ف,م,ص ...

The diacritical dots of a character can be located above or below it but not the two simultaneously.

Most of the characters can be connected from both sides, the right and the left one, however, there are six letters that impose a space after (ز,ر,د,ذ,و,ا), they can be connected from only the right side, it is for this reason that Arabic language is said to be semi-cursive. This characteristic implies that each word may be composed of one unit or more (sub-words).

Certain character combinations form new ligature shapes which are often font dependant. Some ligatures involve vertical stacking of characters, this characteristic complicates the problem of segmentation (known as analytic approach) (Al Badr and Mahmoud, 1995; Essoukhri Ben Amara, 2002).

In this work we deal with a restricted vocabulary containing 48 words (Table 1). These words are those used by Arabic writers while filling the literal amount field of their checks.

Table 1
Arabic literal amounts vocabulary

احد	تسعة	ستون	اربعمائة	ألفا	مليار ان
اثنان	عشر	سبعون	خمسمائة	الفان	ملايير
ثلاثة	عشرة	ثمانون	ستمائة	مليون	سنتيم
اربعة	اثنا	تسعون	سبعمائة	ملايين	و
خمسة	عشرون	مائة	ثمانمائة	مليونا	دينار
ستة	ثلاثون	مائتا	تسعمائة	مليونان	دنانير
سبعة	اربعون	مائتان	ألف	مليار	سنتيمات
ثمانية	خمسون	ثلاثمائة	الاف	مليارا	جزائري

3. Related work on Arabic words recognition

Pechwitz and Maergner (2003) proposed a system based on semi-continuous one-dimensional hidden Markov models (HMM) using pixel values as basic features. The tests were performed using Tunisian town/village names and achieve maximal recognition rates of about 89%. Snoussi-Maddouri et al. (2002), describe a system based on a transparent neural network (TNN) which proceeds by a global vision of structural descriptors during propagation step and local vision by normalized Fourier descriptors during retro-propagation step. An evaluation was done on literal amount of Arabic checks and Tunisian town/village names, the recognition rates were around 90%.

In our previous works in this field, we first used perceptual high-level features and structural holistic classifier applied to Arabic literal amounts, the system accuracy was about 83.55% when it was tested on a set of 240 words (Souici et al., 1999). We were also interested by hybrid classifiers and we proposed a knowledge-based neural network for the recognition of Algerian city names (Farah et al., 2004) and literal amounts (Souici-Meslati and Sellami, 2004). In these works, we started with words perceptual features analysis in order to construct a hierarchical knowledge