A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models

Sameh M. Awaidah, Sabri A. Mahmoud *

King Fahd University of Petroleum and Minerals, College of Computer Sciences and Engineering, Dhahran 31261, Saudi Arabia

ARTICLE INFO

ABSTRACT

This paper describes a technique for the recognition of optical off-line handwritten Arabic (Indian) numerals using hidden Markov models (HMM). Features that measure the image characteristics at local, intermediate, and large scales were applied. Gradient, structural, and concavity features at the sub-regions level are extracted and used as the features for the Arabic (Indian) numeral. Several experiments were conducted for estimating the suitable number of image divisions, and the best combination of features using the HMM classifier. A number of experiments were conducted to estimate the best number of states and codebook sizes in terms of the highest recognition rate possible. In this work, we did not follow the general trend of using the sliding window technique with HMM. Instead, a multi-resolution feature extraction approach was implemented on the whole digit.

A database of 44 writers, with 48 samples per digit resulting in a database of 21 120 samples was used. The achieved average recognition rate is 99%. The classification errors were analysed and attributed to bad data, different writing styles of some digits, errors between digit pairs, and genuine errors. The presented technique, which is writer independent, proved to be effective in the automatic recognition of Arabic (Indian) numerals.

1. Introduction

Handwritten digit recognition is a vital component in many applications; office automation, check verification, and a large variety of banking, business, postal address reading, sorting and reading handwritten and printed postal codes, and data entry applications are few examples.

The recognition of handwritten text (characters and numeral digits) is a more difficult task due to the different handwriting styles of the writers that are subject to inter-, and intra-writer variations. Arabic handwriting, unlike Latin, has many handwriting styles like Naskh, Kofi, and others. In several instances, writers mix between these writing styles [1]. This makes the recognition problem more difficult hence requiring more sophisticated and very advanced feature extraction and recognition techniques.

Arabic text recognition (ATR) has not been researched as thoroughly as Latin, Japanese, or Chinese. The lag of research on Arabic text recognition ATR compared with other languages (e.g. Latin or Chinese) may be attributed, in part, to lack of adequate support including lack of benchmarking databases. The calligraphic nature of the Arabic set is distinguished from other languages in several ways. For example, Arabic text is written from right to left, with the Arabic having 28 basic characters, of which 16 have from one to three dots. Those dots differentiate between the otherwise similar characters. Within a word, some characters connect to the preceding and/or following characters, and some do not connect. The shape of an

---