Pattern Recognition 47 (2014) 1096–1112

Contents lists available at ScienceDirect

Pattern Recognition

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

KHATT: An open Arabic offline handwritten text database

Sabri A. Mahmoud [a,*], Irfan Ahmad [a], Wasfi G. Al-Khatib [a], Mohammad Alshayeb [a], Mohammad Tanvir Parvez [b], Volker Märgner [c], Gernot A. Fink [d]

[a] King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
[b] Qassim University, Qassim 51477, Saudi Arabia
[c] Technische Universität Braunschweig, 38092 Braunschweig, Germany
[d] Technische Universität Dortmund, 44227 Dortmund, Germany

ARTICLE INFO

Available online 13 September 2013

Keywords:
Arabic handwritten text database
Arabic OCR
Document *analysis*
Form processing

ABSTRACT

A comprehensive Arabic handwritten text database is an essential resource for Arabic handwritten text recognition research. This is especially true due to the lack of such database for Arabic handwritten text. In this paper, we report our comprehensive Arabic offline Handwritten Text database (KHATT) consisting of 1000 handwritten forms written by 1000 distinct writers from different countries. The forms were scanned at 200, 300, and 600 dpi resolutions. The database contains 2000 randomly selected paragraphs from 46 sources, 2000 minimal text paragraph covering all the shapes of Arabic characters, and optionally written paragraphs on open subjects. The 2000 random text paragraphs consist of 9327 lines. The database forms were randomly divided into 70%, 15%, and 15% sets for training, testing, and verification, respectively. This enables researchers to use the database and compare their results. A formal verification procedure is implemented to align the handwritten text with its ground truth at the form, paragraph and line levels. The verified ground truth database contains meta-data describing the written text at the page, paragraph, and line levels in text and XML formats. Tools to extract paragraphs from pages and segment paragraphs into lines are developed. In addition we are presenting our experimental results on the database using two classifiers, viz. Hidden Markov Models (HMM) and our novel syntactic classifier.

The database is made freely available to researchers world-wide for research in various handwritten-related problems such as text recognition, writer identification and verification, forms analysis, pre-processing, segmentation. Several international research groups/researchers acquired the database for use in their research so far.

* Corresponding author. Tel.: +966 3 860 1117; fax: +966 3 860 2174.
E-mail addresses: smasaad@kfupm.edu.sa, smasaad@gmail.com (S.A. Mahmoud), irfanics@kfupm.edu.sa (I. Ahmad), wasfi@kfupm.edu.sa (W.G. Al-Khatib), alshayeb@kfupm.edu.sa (M. Alshayeb), m.parvez@qu.edu.sa (M. Tanvir Parvez), v.maergner@tu-bs.de (V. Märgner), Gernot.Fink@tu-dortmund.de (G.A. Fink).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.08.009