SAMRA: Dataset in Al-Soudani Arabic Maghrebi Script for Recognition of Arabic Ancient Words Handwritten

Authors : Sidi Ahmed Maouloud, Cheikh Ba

Abstract : Much of West Africa's cultural heritage is written in the Al-Soudani Arabic script, which was widely used in West Africa before the time of European colonization. This Al-Soudani Arabic script is an African version of the Maghrebi script, in particular, the Al-Mebssout script. However, the local African qualities were incorporated into the Al-Soudani script in a way that gave it a unique African diversity and character. Despite the existence of several Arabic datasets in Oriental script, allowing for the analysis, layout, and recognition of texts written in these calligraphies, many Arabic scripts and written traditions remain understudied. In this paper, we present a dataset of words from Al-Soudani calligraphy scripts. This dataset consists of 100 images selected from three different manuscripts written in Al-Soudani Arabic script by different copyists. The primary source for this database was the libraries of Boston University and Cambridge University. This dataset highlights the unique characteristics of the Al-Soudani Arabic script as well as the new challenges it presents in terms of automatic word recognition of Arabic manuscripts. An HTR system based on a hybrid ANN (CRNN-CTC) is also proposed to test this dataset. SAMRA is a dataset of annotated Arabic manuscript words in the Al-Soudani script that can help researchers automatically recognize and analyze manuscript words written in this script.

Keywords : dataset, CRNN-CTC, handwritten words recognition, Al-Soudani Arabic script, HTR, manuscripts

Conference Title : ICDAR 2023 : International Conference on Document Analysis and Recognition

Conference Location : Bali, Indonesia

Conference Dates : January 09-10, 2023