Data Gathering and Analysis for Arabic Historical Documents

Authors : Ali Dulla

Abstract : This paper introduces a new dataset (and the methodology used to generate it) based on a wide range of historical Arabic documents containing clean data simple and homogeneous-page layouts. The experiments are implemented on printed and handwritten documents obtained respectively from some important libraries such as Qatar Digital Library, the British Library and the Library of Congress. We have gathered and commented on 150 archival document images from different locations and time periods. It is based on different documents from the 17th-19th century. The dataset comprises differing page layouts and degradations that challenge text line segmentation methods. Ground truth is produced using the Aletheia tool by PRImA and stored in an XML representation, in the PAGE (Page Analysis and Ground truth Elements) format. The dataset presented will be easily available to researchers world-wide for research into the obstacles facing various historical Arabic documents such as geometric correction of historical Arabic documents.

Keywords : dataset production, ground truth production, historical documents, arbitrary warping, geometric correction
Conference Title : ICCAIP 2018 : International Conference on Computer Analysis of Images and Patterns
Conference Location : Paris, France
Conference Dates : May 17-18, 2018