Distorted Document Images Dataset for Text Detection and Recognition

Authors : Ilia Zharikov, Philipp Nikitin, Ilia Vasiliev, Vladimir Dokholyan

Abstract : With the increasing popularity of document analysis and recognition systems, text detection (TD) and optical character recognition (OCR) in document images become challenging tasks. However, according to our best knowledge, no publicly available datasets for these particular problems exist. In this paper, we introduce a Distorted Document Images dataset (DDI-100) and provide a detailed analysis of the DDI-100 in its current state. To create the dataset we collected 7000 unique document pages, and extend it by applying different types of distortions and geometric transformations. In total, DDI-100 contains more than 100,000 document images together with binary text masks, text and character locations in terms of bounding boxes. We also present an analysis of several state-of-the-art TD and OCR approaches on the presented dataset. Lastly, we demonstrate the usefulness of DDI-100 to improve accuracy and stability of the considered TD and OCR models.

Keywords : document analysis, open dataset, optical character recognition, text detection
Conference Title : ICMV 2019 : International Conference on Machine Vision
Conference Location : Berlin, Germany
Conference Dates : May 21-22, 2019

Open Science Index, Computer and Information Engineering Vol:13, No:05, 2019 publications.waset.org/abstracts/106148.pdf