International Journal on Document Analysis and Recognition (IJDAR) (2019) 22:189–191
https://doi.org/10.1007/s10032-019-00342-z

EDITORIAL

Editorial for special issue on "Advanced Topics in Document Analysis and Recognition"

Cheng-Lin Liu[1] · Andreas Dengel[2] · Rafael Dueire Lins[3]

Published online: 20 August 2019

The technology of Document Analysis and Recognition, as a subfield of pattern recognition, faces many application needs in the real world, such as the digitization of books, newspapers and archives, invoices and business documents, postal mail sorting, license plate recognition, information retrieval of Web documents, text-based communication and interaction. Huge progress has been achieved in the past 50 years in research and applications. Particularly, the fast development and wide application of deep learning techniques have led to significant performance upgrading in recent years. The improvement of technology makes existing applications more promising and new applications realizable. The extension of applications, e.g., from simple documents to complex documents, from scanned paper documents to camera-captured natural scene documents, in turn raises many research problems.

This special issue is aimed to report the latest advances of document analysis and recognition on complex documents using advanced techniques. By November 2018, 30 submissions were received for consideration. Each submission was assigned to a guest editor or an associate editor of IJDAR, who invited at least two reviewers to review the submission. After one to three rounds of review by July 2019, 11 articles were decided to be accepted for publication in this issue. The 11 articles can be divided into two categories, with the contributions outlined below.

The first six articles address text detection and document analysis applications. The first article "An Anchor-Free Region Proposal Network for Faster R-CNN based Text Detection Approaches," by Zhuoyao Zhong, Lei Sun and Qiang Huo, proposes a novel anchor-free region proposal network (AF-RPN) to replace the original anchor-based RPN in the faster R-CNN framework to address the variability of scale, aspect ratio and orientation of text bounding box. Compared with the anchor-based region proposal generation approaches, AF-RPN can get rid of complicated anchor design and achieves higher recall rate on both horizontal and multi-oriented text detection benchmark tasks.

The article "A Two-Stage Method for Text Line Detection in Historical Documents," by Tobias Grüning, Gundram Leifert, Tobias Strauß, Johannes Michael and Roger Labahn, presents a two-stage text line detection method for historical documents. In the first stage, a deep neural network called ARU-Net labels pixels to belong to one of the three classes: baseline, separator and other. The predictions are used as input for the second stage which performs a bottom-up clustering to build baselines of text lines. The method is capable of handling complex layouts as well as curved and arbitrarily oriented text lines.

The article "Coarse-to-fine Document Localization in Natural Scene Image with Regional Attention and Recursive Corner Refinement," by Anna Zhu, Chen Zhang, Zhi Li and Shengwu Xiong, proposes a coarse-to-fine document localization approach to detect the four corner points of the document in natural scene images. First, the four corners are roughly predicted through a deep neural network-based Joint Corner Detector (JCD) with an attention mechanism, and then, the predicted corners are refined by a corner-specific CNN-based refiner. The promise of the method was shown for document localization in natural scene images with complex background.

The article "Comic MTL: multi-task model for comic book image analysis," by Nhu-Van Nguyen, Christophe Rigaud and Jean-Christophe Burie, proposes one model that can learn multiple tasks called Comic MTL instead of using one model per task, to reduce the processing time for comic book image analysis. A task for relation analysis for bal-

✉ Cheng-Lin Liu
liucl@nlpr.ia.ac.cn

Andreas Dengel
andreas.dengel@dfki.de

Rafael Dueire Lins
rdl@ufpe.br

[1] Institute of Automation of Chinese Academy of Sciences, Beijing 100190, People's Republic of China

[2] DFKI, Kaiserslautern, Germany

[3] Universidade Federal de Pernambuco, Recife, Brazil

loons and characters is also integrated. Experimental results show that the Comic MTL model can detect the associations between balloons and their speakers (comic characters) and handle multiple tasks like panel and character detection and balloons segmentation.

The article "Generalized Framework for Summarization of Fixed-Camera Lecture Videos by Detecting and Binarizing Handwritten Content," by Bhargava Urala Kota, Kenny Davila, Alexander Stone, Srirangaraj Setlur and Venu Govindaraju, reports a system for extracting and summarizing contents in lecture videos. First, a deep learning pipeline is proposed for detecting handwritten text, formulae and sketches and binarizing the extracted content in video frames. The spatiotemporal structure of the binarized detections is exploited to compute associativity information of content across all video frames. This information is later used to segment the video. And finally, summarization is performed to produce temporal splits of the video minimizing the number of conflicts present on each video segment.

The article "A comparison of local features for camera-based document image retrieval and spotting," by Quoc Bao Dang, Mickal Coustaty, Muhammad Muzzamil Luqman and Jean-Marc Ogier, presents a comprehensive comparison of robustness of local features for camera-based document image retrieval and spotting system. After a literature review of the state of the art of local features extraction including keypoint detectors and descriptors, a dataset and evaluation protocol for camera-based document image retrieval and spotting systems are presented. Then, performance measurements and detailed evaluation of local features from the literature are given.

The next five articles address character and text recognition. In "Boosting Scene Character Recognition by Learning Canonical Forms of Glyphs," Yizhi Wang, Zhouhui Lian, Yingmin Tang and Jianguo Xiao propose a novel methodology for boosting scene character recognition by learning canonical forms of glyphs, based on the fact that characters appearing in scene images are all derived from their corresponding canonical forms. They design a GAN-based model to make the learned deep feature of a given scene character capable of reconstructing corresponding glyphs in a number of standard font styles. This results in deep features that are more discriminative in recognition and less sensitive against image disturbing factors.

In "Are 2D-LSTM really dead for offline text recognition?", Bastien Moysset and Ronaldo Messina present a fair comparison between 2D-LSTM and competing models on complex datasets that are more representative of challenging "real-world" data, compared to "academic" datasets that are more restricted in their complexity. They aim at determining when and why the 1D and 2D recurrent models have different results. They also use a language model to assess the effects of linguistic constraints on different networks. The results show that for challenging datasets, 2D-LSTM networks can provide the highest performances.

In "Handwritten Arabic Text Recognition Using Multi-Stage Sub-Core Shape HMMs," Irfan Ahmad and Gernot Fink present a multi-stage HMM-based text recognition system for handwritten Arabic. This system employs a novel way of representing Arabic characters by separating the core shapes from the diacritics and then representing these core shapes by smaller units called as sub-core shapes. Contextual HMM modeling utilizing these sub-core shapes is presented to get significantly compact recognizer. Furthermore, multi-stream contextual sub-core shape HMMs are presented. Experimental results show that the presented system outperforms the standard character-shape system.

In "Dynamic Temporal Residual Network for Sequence Modeling," Ruijie Yan, Liangrui Peng, Shanyu Xiao, Michael T. Johnson and Shengjin Wang propose a dynamic temporal residual network (DTRN) by incorporating residual learning into an LSTM network along the temporal dimension, so as to better model the dynamic temporal dependencies in sequential data. Experiments on three commonly used public handwriting recognition datasets (IFN/ENIT, IAM and Rimes) and one speech recognition dataset (TIMIT) show that the proposed method outperforms previous related sequence modeling methods.

The last article "On optimal stopping strategies for text recognition in a video stream as an application of a monotone sequential decision model," by Konstantin Bulatov, Nikita Razumnyi and Vladimir V. Arlazarov, addresses the novel problem of stopping text field recognition process in a video stream, which is particularly relevant to real-time mobile document recognition systems. On providing a decision-theoretic framework for this problem, and exploring similarities with existing stopping rule problems, a strategy is proposed based on thresholding the estimation of the expected difference between consequent recognition results. The method was shown to outperform previously published methods based on identical results cluster size thresholding.

The guest editors thank all the authors who shared their invaluable work and all reviewers for their insightful comments in considering the submissions for this special issue. We also want to thank the Editors-in-Chief of "International Journal on Document Analysis and Recognition" for giving us an opportunity to guest-edit this special issue and their helps in managing the reviewing. Finally, we want to recognize the hard work of journal assistants Karthika Navukkarasu and Priya Verma for their ongoing support and assistance in this tremendous effort.

Guest Editors:
Cheng-Lin Liu
Andreas Dengel
Rafael Lins