International Journal on Document Analysis and Recognition (IJDAR) (2021) 24:1–2
https://doi.org/10.1007/s10032-021-00372-6

EDITORIAL

Deep learning for graphics recognition: document understanding and beyond

Jean-Christophe Burie[1] · Alicia Fornés[2] · K. C. Santosh[3] · Muhammad Muzzamil Luqman[1]

Published online: 22 May 2021

Graphics Recognition is the subfield of pattern recognition that deals with graphic entities. Graphical cues (including graphical languages) are often better to describe complex ideas as compared to text. The recognition of graphical elements (e.g., graphical notations) in heterogeneous documents is useful in understanding the contents and user intentions, or to identify the application domain. Since the 1980s, the analysis, and interpretations of graphical documents (e.g., electrical circuit diagrams, engineering drawings, etc.), handwritten and printed graphical elements/cues (e.g., logos, stamps, etc.), graphics-based information retrieval and sketches, to name a few, have been challenging research topics.

In recent years, due to the ability to exploit big data and its superior representation and prediction performance, deep learning (DL) has demonstrated great successes in various applications of pattern recognition and artificial intelligence, including character and text recognition, image segmentation, object detection and recognition. In certain cases, some issues can even be considered solved when using DL methods. In pattern recognition, the convolutional neural network (CNN) and the recurrent neural network (RNN) with long short-term memory (LSTM) have been successfully applied. The performance improvement is effective when DL models are considered, which require relatively large data within the supervised framework. How can graphics recognition (GREC) benefit from DL models? Can common DL models be used to tackle GREC problems and advance document understanding overall? Unlike work in conventional pattern recognition and machine learning, in this special issue, we focus on high-level DL architectures and/or models for graphics recognition. In this special issue, we acknowledge new advances in document analysis and recognition (DAR) that use DL models to analyze graphic-rich documents.

The guest editors received eight full submissions for this special issue. The topics ranged from document image segmentation, layout analysis, text and object localization to character and text recognition, language modeling and handwritten mathematics recognition. They included signature verification, document retrieval and document understanding. The guest editors created a strict, peer-review process and invited guest reviewers to consider all submissions. At least two reviewers reviewed each paper, and most accepted papers underwent second-round reviews. Finally, the guest editors and reviewers accepted four papers for publication in this special issue. An outline of the contents follows:

The article "Arrow R-CNN for Handwritten Diagram Recognition'' by Bernhard Schäfer, Margret Keuper and Heiner Stuckenschmidt addresses the problem of offline handwritten diagram recognition. The authors propose Arrow R-CNN, a deep learning system for joint symbol and structure recognition in handwritten diagrams. Arrow R-CNN extends the Faster R-CNN object detector with an arrowhead and tail key point predictor and a diagram-aware post-processing method. A network architecture and data augmentation methods, targeted at small diagram datasets, are proposed. The diagram-aware postprocessing method addresses the insufficiencies of standard Faster R-CNN post-processing. It reconstructs a diagram from a set of symbol detections and arrow key points. Arrow R-CNN substantially

✉ Jean-Christophe Burie
jean-christophe.burie@univ-lr.fr

Alicia Fornés
afornes@cvc.uab.es

K. C. Santosh
santosh.kc@usd.edu

Muhammad Muzzamil Luqman
muhammad_muzzamil.luqman@univ-lr.fr

[1] Laboratoire L3i, La Rochelle Université, La Rochelle, France

[2] Computer Vision Centre, Computer Science Department, Universitat Autónoma de Barcelona, Barcelona, Spain

[3] KC's PAMI Research Lab - Computer Science, The University of South Dakota, Vermillion, SD 57069, USA

improves the diagram recognition in comparison with state-of-the-art methods.

The article "Knowledge driven Description Synthesis for Floor Plan Interpretation" by Shreya Goyal, Chiranjoy Chattopadhyay and Gaurav Bhatnagar addresses the problem of caption generation from floor plan images. The authors propose two models, Description Synthesis from Image Cue (DSIC) and Transformer-Based Description Generation (TBDG) for text generation from floor plan images. These two models take advantage of modern deep neural networks for visual feature extraction and text generation. The difference between the models is in the way they take input from the floor plan image. The DSIC model takes only visual features automatically extracted by a deep neural network, while the TBDG model learns textual captions extracted from input floor plan images with paragraphs. Experiments are carried out on a large scale publicly available dataset and are compared with state-of-the-art techniques to show the effectiveness of the proposed models.

The article "Cross-modal Photo-Caricature Face Recognition Based on Dynamic Multi-task Learning" by Zuheng Ming, Jean-Christophe Burie, Muhammad Muzzamil Luqman addresses the problem of face recognition in caricatures. Contrary to real face recognition, the recognition of caricatures implies dealing with exaggerated facial features. In this work, the authors propose to approach this problem through multi-task learning, making use of both caricatures and real faces. The first branch in the system architecture is used for caricature identification, the second one for caricature verification, and the third branch for visual identification. Instead of using fixed weights, they propose to dynamically learn the weights of these tasks depending on the importance of each task. This proposed dynamic multi-task learning for cross-modal caricature-visual face recognition has been tested using two caricature datasets. The experimental results are compared with state-of-the-art methods, demonstrating the superiority of multi-task learning with dynamic weights for this problem.

The article "CNN based segmentation of speech balloons and narrative text boxes from comic book page images" by Arpita Dutta, Samit Biswas and Amit Kumar Das addresses the problem of segmentation of speech balloons and narrative text boxes in comics. The correct segmentation, retrieval and relation between these elements allows for the tracking of conversations among characters, facilitating the further processing and understanding of comics. For this task, the authors propose a shape-aware dual stream architecture based on CNNs. The CNN model concatenates both edge and semantic information for a more efficient detection of narrative text boxes and speech balloons. The proposed method is tested on several public comic datasets and compared to some state-of-the-art methods, demonstrating the suitability of their approach. Finally, the authors have also created a new comic dataset that uses the Bangla language together with the development of a semi-automatic approach for generating ground truth, which can significantly benefit the research community.

We would like to thank the authors who contributed to this special issue and express our gratitude to the reviewers who did a tremendous job within tight deadlines. Finally, we are indebted to Springer support staff as well as the "International Journal on Document Analysis and Recognition (IJDAR)" Editors-in-Chief—Prof. Koichi Kise, Prof. Daniel Lopresti, and Prof. Simone Marinai—for their constant support and understanding in completing this special issue.