IJDAR (2011) 14:1–2
DOI 10.1007/s10032-010-0145-1

EDITORIAL

Editorial

Liu Wenyin · Ernest Valveny

Published online: 6 January 2011

Performance evaluation of algorithms has become more and more important in pattern recognition in general, and document analysis and recognition, in particular. More and more people are interested in those algorithms/methods/systems which are thoroughly evaluated in terms of commonly accepted performance metrics on publicly accessible ground-truth data. Thus, there is a need for standard datasets and protocols to compare and evaluate existing methods. The goal of an evaluation protocol should be to establish a solid knowledge of the state of the art in a given research problem, i.e., to determine the weaknesses and strengths of the proposed methods on a common and general set of input data. Performance evaluation should allow the selection of the best-suited method for a given application of the methodology under evaluation.

There are usually two main tasks in the definition of any framework for performance evaluation: definition of the datasets (including collection or automatic generation of data, ground-truthing, definition of training and test datasets) and performance characterization, which determines the metrics and the protocol to match the results of a given method with the ground-truth in order to give a measure of the performance. These general principles have been applied in the past to the evaluation of different processes related to document analysis, including thinning, OCR, page segmentation, table understanding, vectorization, or symbol recognition. As a result, there has been an increasing interest and participation for the series of competitions organized in the framework of the International Conference on Document Analysis in Recognition. In the last ICDAR'09 nine competitions were proposed with more than one hundred participating algorithms (methods/systems/configurations).

This special Issue of IJDAR on Performance Evaluation of Document Analysis and Recognition Algorithms aims at promoting further the research and practice in this area by selecting best extended papers describing the protocol and the results of ICDAR competitions, as well as other publicly submitted papers describing innovative protocols, technologies, methodologies, tools, and applications for performance evaluation of any algorithm in the area of document analysis and recognition, making a clear and useful contribution in any of the aspects of a performance evaluation framework (including ground-truth data, metrics, evaluation protocol, comparison with state-of-the-art methods).

The special issue includes five papers describing ICDAR'09 competitions. Three of them have to do with the evaluation of different tasks related to handwriting. In particular, the first paper by El Abed and Märgner presents the results of the third competition on Arabic handwriting recognition, following those of 2005 and 2007. The paper includes a description of the competition framework as well as a summary of each system participating in the competition. Being the third edition of this series of competitions, the authors analyze the results in comparison to past editions and propose some actions for future editions. The second paper, by El Abed et al. describes the first competition on online Arabic handwriting recognition, including details about the data and how the ground-truth was obtained, the evaluation protocol, a description of participating systems and the analysis of the results. The third paper related to handwriting, by Gatos et al., is devoted to the evaluation of off-line handwritten text at the level of text lines and words. The competition

L. Wenyin (✉)
City University of Hong Kong, Hong Kong, China
e-mail: csliuwy@cityu.edu.hk

E. Valveny
Computer Vision Center - Universitat Autònoma de Barcelona,
Edifici O, Campus UAB, 08193 Bellaterra, Spain
e-mail: Ernest.Valveny@uab.cat

follows that of ICDAR'07, though new data have been added for testing. The paper contains a description of the dataset, the ground-truth, the evaluation criteria and the winning algorithms. Results are presented, analyzed and compared with those obtained using two state-of-the-art methods. The fourth paper related to ICDAR competitions, by Gatos et al., explains the work done for the binarization contest, including evaluation measures and results of the 43 participating systems. The top 5 systems are also summarized and results are compared with state-of-the-art methods. Finally, the paper by Doucet et al. introduces the framework used for the competition on structure extraction from OCR-ed books, where the goal was to build hyperlinked tables of contents from the results of OCR. The framework includes a collaborative ground-truth protocol to annotate the dataset and the definition of evaluation measures based on precision and recall of different types of entries in the table of contents. The analysis of the results and a brief description of the methods complete the paper.

In addition to papers related to ICDAR competitions, five more papers have been selected for this special issue. These papers present contributions in two of the elements of a framework for performance evaluation: dataset creation (including ground-truthing) and evaluation metrics. The paper by Jin et al. explains the work done to build a big database of online handwritten Chinese text with 11 different datasets and more than 3.6 million characters. It describes a set of rules defined to ensure a representative dataset, the process of ground-truthing and labeling, and the properties of the 11 datasets. The database is tested using the results obtained using one state-of-the-art method. The paper by MacLean et al. presents an approach to using a grammar for creating ground-truth for sketches. The grammar is used to generate templates according to the rules of the specific sketch domain. These templates are manually reproduced and a matching algorithm automatically puts in correspondence the manual sketch with the ground-truth generated by the grammar. The paper shows how this approach has been used to create a dataset of 4500 sketches corresponding to mathematical expressions. Following with mathematical expressions, the paper by Sain et al. makes a proposal of a performance measure for the evaluation of recognition of mathematical expressions based on tree matching between the trees of the ground-truth and the output of the recognizer. It gives not only a global measure, but also information on the particular errors introduced by the recognizer. It has been tested using existing recognition methods on both scanned and handwritten documents. The paper by Visani et al. proposes a protocol to evaluate shape descriptors in terms of descriptive power (uniqueness, distinctiveness and robustness towards noise) of a given descriptor and complementarity between multiple descriptors in order to help selecting the best combination. The protocol is independent of the final application and it has been tested using two well-known shape descriptors, ART and shape context on the GREC'03 symbol database. Finally, the paper by Silva proposes two new evaluation metrics, completeness and purity (C&P), and illustrates using the examples of table location and segmentation that they are more relevant for division/aggregation tasks in document analysis than the traditionally used precision and recall metrics.

The guest editors would like to thank all the authors that have submitted papers to this special issue and all the reviewers for their contribution. We would also like to thank the Editors-in-Chief, Profs. D. Doermann, K. Tombre and S. Marinai, the Editorial Board, and Ms. G. Balasubramanian for making it possible to organize this special issue.