RECOGNITION & ENRICHMENT OF ARCHIVAL DOCUMENTS

D3.9
ScriptNet:Competition P3

Research competition

Giorgos Sfikas, Markus Diem, Florian Kleber, Basilis Gatos, George Louloudis, Nikolaos Stamatopoulos, Stavros Perantonis

NCSR 'Demokritos'

Distribution: http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	READ
Project full title	Recognition and Enrichment of Archival Documents
Instrument	H2020-EINFRA-2015-1
Thematic priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date/duration	01 January 2016 / 42 Months

Distribution	Public
Contract. date of delivery	30.06.2019
Actual date of delivery	30.06.2019
Date of last update	30.06.2019
Deliverable number	D3.9
Deliverable title	ScriptNet:Competition P3
Type	other
Status & version	1.0
Contributing WP(s)	WP3
Responsible beneficiary	NCSR
Other contributors	NCSR, UPVLC
Internal reviewers	Joan-Andreu Sanchez
Author(s)	Giorgos Sfikas, Markus Diem, Florian Kleber, Basilis Gatos, George Louloudis, Nikolaos Stamatopoulos, Stavros Perantonis
EC project officer	Martin Majek
Keywords	research competition platform, ScriptNet

Contents

1 Executive summary

This deliverable reports on the research competitions organised by the READ consortium, as well as the status of the ScriptNet competitions platform at the end of the third year of the READ project.

2 Introduction

The goal of this task is the organisation of open research competitions, throughout the duration of the project, that will be promoted among the computer science community. Research competitions are scheduled to be organised and promoted as part of important document processing conferences every year of the project. In the third year of the project (2018), several competitions have been organised as part of the International Conference on Frontiers in Handwriting Recognition (ICFHR) conference [1] . ICFHR, along with ICDAR, are the major document image processing events of each year, both held interchangeably and on a biennial basis.

Starting from the previous year, the majority of the READ-organised research competitions were integrated with the *Scriptnet platform*, a platform/site specifically developed by the READ consortium. The ICFHR HTR 2016 competition was already integrated with the Scriptnet platform and remained open for submissions all through the duration of 2017, 2018 and 2019. Furthermore, all Scriptnet-integrated competitions organized in conjuction with ICDAR 2017 have remained open for submissions during 2018 and 2019 as well. Scriptnet is therefore proposed as a unified platform where competition organisers can create and customise their competition, and competition participants can register, follow and submit their results.

3 Scriptnet platform technical developments

The Scriptnet platform has been developed as a site written on and running on Django, a well-known and robust web-based framework [1]. In year one of the project, Scriptnet had commenced development from scratch, as well as beta testing. In years two and three, the Scriptnet platform has proven ready to face real-world conditions, with which it has coped with success. In year three (2018) one new competition has been integrated in the Scriptnet platform, receiving in total and along with the previous competitions that have remained open, more than 560 result file submissions and processing them automatically with success (nearly tripling the total submissions compared with year two).

The developed code is always available in public at Github [2]. The public Github repository now contains more than 455 code commits, while a total of 74 issues and 27 pull requests have been succesfully addressed and closed. The latest stable release of the platform is running at `https://scriptnet.iit.demokritos.gr/competitions`.

[1] `http://icfhr2018.org/competitions.html`

[2] `https://github.com/Transkribus/competitions`

In order to increase the safety of the submitted results, and as a measure against unexpected events that may jeopardise normal Scriptnet server execution, we have setup a separate, private git server that includes all commits that correspond to database submissions. Regular git commits are setup automatically on a 24-hour basis, to ensure that a very detailed account of the Scriptnet platform submission history is saved.

This year, the competition scoreboard has been slightly tweaked to handle differentiating between submissions before the competition deadline, and late submissions, i.e. after the competition deadline. Submissions before the competition deadline now are placed on the competition scoreboard, but late submissions, while possible, do not affect the scoreboard.

4 Reception of the Scriptnet platform competitions

This year's research competitions have validated the usefulness of the Scriptnet platform, continuing on the positive trend set on the previous year. Overall, we can say that the platform succesfully realized the expectations of its developers as well as to those of its users.

As of the end of year two of the READ project, 168 users are registered on the Scriptnet platform (110 in year two) These users correspond to a total of 90 different participant affiliations (68 in year two). Participant affiliations are typically universities and research centers, as well as libraries and other institutions from all over the globe.

As of mid-2019, one new competition for this year has already been integrated with the Scriptnet platform ("ICDAR2019 Competition on BAseline Detection and Page Segmentation").

In 2018, a new competition was integrated with the platform, namely "ICFHR 2018 Automated Text Recognition on a READ dataset (ICFHR2018 ATR)". This competition has met with a very positive reception, with a total of 38 users following the competition, and creating more than 160 submissions by the end of the current year.

The total number of result files submitted in total, taking into account all Scriptnet/READ competitions, surpassed 560 submissions (222 in year two). All of these submissions have been succesfully processed automatically upon submission, with the developed Django backend.

5 Organisation of competitions in international conferences

5.1 ICFHR 2018 Automated Text Recognition on a READ dataset (ICFHR2018 ATR)

Automated Text Recognition (ATR) has made huge progress within the last few years. Even for complex historical documents, character error rates (CER) below 10% can be achieved. In according competitions (e.g. the recent ICDAR2017 HTR), the training data usually was taken from the same document as the test data.

In contrast, for practical applications, typically there is no ground truth available for the specific document to be transcribed. Consequently, in order to train an ATR system that then produces reasonably low error rates, a certain amount (possibly up to some hundreds) of pages would have to be transcribed in good quality. But due to the essential human effort for creating ground truth, this is both expensive and time consuming.

On the other hand, many text corpora have already been transcribed and published. This raises the question to what extend such (more or less public) datasets could be used to pre-train a rather universal ATR system such that one can minimize the amount of additional training data, which is necessary for proper subsequent document specific applicability. Moreover, the dependency between the amount of available specific training data and the CER gain is of apparent practical interest.

In order to encourage further research towards robust ATR systems, i.e. which can properly deal with distinct scripts, this competition targets on emulating such application scenarios by providing a new, rather heterogeneous dataset containing various documents from different writers, time periods and languages. The documents were taken from the advanced transcription platform Transkribus that currently is under further development in the EU Horizon-2020 project READ.

The competition was integrated in the ScriptNet platform and it is still open to allow new systems to be tested.

Contest web page: `https://scriptnet.iit.demokritos.gr/competitions/10/`

5.2 cBAD: ICDAR2019 Competition on BAseline Detection and Page Segmentation

This is a competition that benchmarks two aspects of layout analysis: text extraction and page segmentation. The objective is to automatically locate and correctly label regions in document page images. A challenging dataset that contains 2700 manually annotated document images will be published together with this competition. The dataset consists of documents that were collected from seven European archives and has document pages that originate from different locations and times. The competition will be organized using ScriptNet with well-known evaluation schemes that benchmark the methods of participating research teams. By these means, we try to facilitate page segmentation research in the document analysis community.

Contest web page: https://scriptnet.iit.demokritos.gr/competitions/11/

References

[1] *Django: The Web framework for perfectionists with deadlines* `https://www.djangoproject.com/`

[2] A.Fornés, V.Romero, A.Baró, J.I.Toledo, J.A.Sánchez, E.Vidal, J.Lladós: *"ICDAR 2017 competition on Information Extraction in Historical Handwritten Records"*, In proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017), November 2017

[3] I.Pratikakis, K.Zagoris, G.Barlas, B.Gatos. *"ICDAR 2017 competition on Information Extraction in Historical Handwritten Records"*, In proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017), November 2017

[4] J.A.Sánchez, V.Romero, A.H.Toselli, M.Villegas, E.Vidal, *"ICDAR 2017 Competition on Handwritten Text Recognition on the READ Dataset"*, In proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017), November 2017

[5] S.Fiel, F.Kleber, M.Diem, V.Christlein, G.Louloudis, N.Stamatopoulos, B.Gatos, *"ICDAR 2017 Competition on Historical Document Writer Identification (Historical-WI)"*, In proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017), November 2017

[6] S.Fiel, F.Kleber, M.Diem, B.Gatos, T.Grüning, *"cBAD: ICDAR2017 Competition on Baseline Detection"*, In proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017), November 2017

[7] I. Pratikakis, K. Zagoris, B. Gatos, J. Puigcerver, A. H. Toselli and E. Vidal., *ICFHR2016 Handwritten Keyword Spotting Competition (H-KWS 2016).* In "Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR 2016)". Pages 613618, Shenzhen, China (October 2016). Published by IEEE Computer Society, ISBN-13: 978-1-5090-0981-7.

[8] J.A. Sánchez, V. Romero, A. H. Toselli, E. Vidal, *ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset*, In "Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR 2016)". pages. 630–635. Shenzhen, China (October 2016). Published by IEEE Computer Society, ISBN-13: 978-1-5090-0981-7.

[9] I. Pratikakis, K. Zagoris, B. Gatos, G. Louloudis, N. Stamatopoulos, *ICFHR 2014 competition on handwritten keyword spotting (H-KWS 2014)*, In "Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR 2014)" (pp. 814-819).