READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D3.8
ScriptNet:Competition P2

Research competition

Giorgos Sfikas, Basilis Gatos, Verónica Romero Gómez, 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	31.12.2017
Actual date of delivery	28.12.2017
Date of last update	22.11.2017
Deliverable number	D3.8
Deliverable title	ScriptNet:Competition P2
Type	other
Status & version	1.0
Contributing WP(s)	WP3
Responsible beneficiary	NCSR
Other contributors	NCSR, UPVLC
Internal reviewers	TB completed
Author(s)	Giorgos Sfikas, Basilis Gatos, Verónica Romero Gómez, 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 second 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. This year, several competitions have been organised as part of the International Conference on Document Analysis and Recognition (ICDAR 2017) conference [1] . ICDAR is the main document image processing event of the year, held on a biennial basis.

Starting from this year, the majority of the READ-organised research competitions were integrated with the *Scriptnet platform*, a platform/site specifically developed by the READ consortium. Furthermore, the ICFHR HTR 2016 competition was integrated with the Scriptnet platform and remained open for submissions all through the duration of 2017. 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 year two, the Scriptnet platform has proven ready to face real-world conditions, with which it has coped with success. Four (4) new competitions have been integrated in the Scriptnet platform, receiving in total more than 200 result file submissions and processing them automatically with success.

The developed code is always available in public at Github [2]. The public Github repository now contains more than 431 code commits, while a total of 72 issues and 21 pull requests have been succesfully addressed and closed. The latest stable release of the platform is running at `https://scriptnet.iit.demokritos.gr/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.

[1] `http://u-pat.org/ICDAR2017/index.php`
[2] `https://github.com/Transkribus/competitions`

Competition	Followers	Submitters
Baseline detection competition 2017	13(6)	6(4)
Writer identification competition 2017	7(5)	5(4)
Keyword spotting competition 2017	13(6)	1(1)
Handwritten text recognition 2017	22(18)	4(3)

Table 1: Participation statistics for Scriptnet-integrated competitions held in 2017. Figures indicate number of individual participants/submitting groups. The number of participants and groups not affiliated with READ is indicated in parenthesis.

The option to *follow* one or more competitions has been added to the platform. Any registered individual may choose to follow the competitions of his/her choice; competition organisers can then notify the competition followers with a mailed message if need be, or may choose to allow certain files/data to be downloaded only by registered followers. Furthermore, the 'follower' functionality allows competition organisers to better monitor the number of possible participants to their competition. The experience of this year has been that the number of followers has always been greater than the number of actual participating/submitting parties, meaning that many participants eventually did not submit any methods. Apparently, these teams were interested in following the results of the other teams' submissions. This is not unnatural, and is consistent with the experience of non-Scriptnet competitions.

Competition organisers can easily check the list of followers by clicking on the 'My Scriptnet' tab after having logged in. Furthermore, they can check the number and details of each submission at the corresponding competition tab. The competition submission list is normally private/non-public by default, unless the organisers choose otherwise.

We have furthermore added a custom time limit for consecutive submissions to a competition. Competition organisers could hence choose for how long a participant should wait before performing an additional submission. This option has been originally added as a measure of protection against participants that would potentially try to overfit their method to the test data. Competition organisers have also the option to disable this measure.

4 Reception of the Scriptnet platform competitions

This year's research competitions were effectively the first test of the Scriptnet platform under real-world conditions. 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, 110 users are registered on the Scriptnet platform. These users correspond to a total of 62 different participant affiliations. Participant affiliations are typically universities and research centers, as well as libraries and other institutions from all over the globe.

We can see information about followers and submitters of the 2017 competitions in table 4. The number of followers (non-members of the READ consortium) ranged from a minimum of 5 (Writer identification competition 2017) to a maximum of 18 (HTR competition 2017). Concerning the number of groups that submitted their methods, with the exception of the Keyword spotting competition that only had 1 submitting group, the other competitions had 3-4 groups competing.

The total number of result files submitted throughout 2017, taking into account all Scriptnet/READ competitions, surpassed 222 submissions. 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 ICDAR 2017 competition on baseline detection (cBAD)

The cBAD competition aims at benchmarking state-of-the-art baseline detection algorithms. It is in line with previous competitions such as the ICDAR 2013 Handwriting Segmentation Contest. A new, challenging, dataset was created to test the behavior of state-of-the-art systems on real world data. Since traditional evaluation schemes are not applicable to the size and modality of this dataset, we introduced a new one that includes baselines to measure performance. We received submissions from five different teams for both tracks.

Contest web page: `https://scriptnet.iit.demokritos.gr/competitions/5/`

Paper: [6]

5.2 ICDAR 2017 Competition on Historical Document Writer Identification (Historical-WI)

Following the successful organization of ICDAR 2011 Writer Identification Contest, ICFHR 2012 Writer Identification Contest Challenge 1: Latin Documents and ICDAR 2013 Competition on Writer Identification, a competition for writer identification techniques was organized in the framework of ICDAR2017, in order to record recent advances in the field of writer identification. This competition deals with the identification of writers in historical handwritten documents. The task is to generate a ranking of the images stored in the database according to the similarity of the handwriting. The ranking for each page (most similar on the first place) is then submitted to the competition website and the identification rate and mean average precision is then calculated.

Contest web page: `https://scriptnet.iit.demokritos.gr/competitions/6/`

Paper: [5]

5.3 ICDAR 2017 Competition on Handwritten Text Recognition on the READ Dataset (ICDAR2017 HTR)

The "ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset" competition organised in the framework of the ICDAR 2017 aims at introducing a usual scenario for some collections in which there exist transcripts at page level for many pages, but these transcripts are not aligned with text lines in the document images. The problem is then to align automatically these transcripts with the corresponding line images for subsequently training an HTR system. In this scenario, it is feasible to annotate accurately/manually line images with their transcripts only for a few pages. This competition provides the researchers researchers working on off-line Handwritten Text Recognition (HTR) a suitable benchmark to compare their techniques. Most of the dataset was taken from the Alfred Escher Letter Collection (AEC) 5 which is written in German but it also has pages in French and Italian. The selected dataset also included handwritten images drawn from other German collections, and thus characterized by being written by several hands. 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/8/`

Paper: [4]

5.4 ICDAR 2017 Handwritten Keyword Spotting Competition (ICDAR2017 KWS)

The H-KWS 2017, organized in the context of the ICDAR 2017 conference, was proposed for benchmarking the two main KWS settings, Query by Example (QbE) and the Query by String (QbS), under unified criteria. The evaluation will focus on assessing KWS capabilities needed for large-scale applications of text retrieval in document images. In line with the requirements of large-scale indexing and retrieval, evaluation will not be based on the geometric accuracy of the spotted words. Instead, larger image regions such as lines, or even full page images will be considered as search targets.

This competition took into account the fact that it is common to have a relatively moderate amount of transcribed material available when indexing large collections of handwritten documents. The cost of producing these transcripts is often negligible with rerspect to the overall cost of the indexing project. Therefore, in this competition a moderate amount of training data, in the form of transcribed page images, was provided to all the entrants. The competition was integrated in the ScriptNet platform.

Contest web page: `https://scriptnet.iit.demokritos.gr/competitions/7/`

This competition was eventually cancelled, as only one entrant participated. However it is still open in the ScriptNet platform.

5.5 ICDAR 2017 Handwritten Document Image Binarization competition

DIBCO 2017 is the international Competition on Document Image Binarization organized in conjunction with the ICDAR 2017 conference. The general objective of the con-

test is to identify current advances in document image binarization of machine-printed and handwritten document images using performance evaluation measures that are motivated by document image analysis and recognition requirements. Eighteen (18) research groups have participated in the competition with twenty six (26) distinct algorithms. The DIBCO 2017 testing dataset consists of 10 machine-printed and 10 handwritten document images for which the associated ground truth was built manually for the evaluation. The selection of the images in the dataset was made so that representative degradations appear. The machine-printed documents of the dataset originate from collections that belong to the IMPACT project, while the handwritten document images originate from collections that belong to READ project partners. The testing dataset along with the associated ground truth as well as the evaluation software are publicly available at: http://vc.ee.duth.gr/dibco2017/benchmark.

Contest web page: `https://vc.ee.duth.gr/dibco2017/`

Paper: [3]

5.6 ICDAR 2017 Information Extraction competition

The ICDAR2017 Competition on Information Extraction in Historical Handwritten Records aims to foster the research in the information extraction field and offer a benchmark for the research community. The extraction of relevant information from historical handwritten documents is one of the key steps in order to make the manuscripts accessible and searchable. In this competition, the goal was to detect the named entities and assign each of them a semantic category, and therefore, to simulate the filling in of a knowledge database. The proposed dataset consists of historical handwritten marriages records from the Archives of the Cathedral of Barcelona. The pages used were written in old Catalan by one single writer in the 17^{th} century.

Contest web page: `http://www.cvc.uab.es/5cofm/competition/`

Paper: [2]

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