READ

RECOGNITION & ENRICHMENT OF ARCHIVAL DOCUMENTS

D3.7
ScriptNet:Competition P1

Research competition

Giorgos Sfikas, Basilis Gatos, Verónica Romero Gómez
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	31.12.2016
Actual date of delivery	28.12.2016
Date of last update	30.11.2016
Deliverable number	D3.7
Deliverable title	ScriptNet:Competition P1
Type	other
Status & version	1.0
Contributing WP(s)	WP3
Responsible beneficiary	NCSR
Other contributors	NCSR, UPVLC
Internal reviewers	Louise Seaward, Joan Andreu Sánchez, Vili Haukkovaara
Author(s)	Giorgos Sfikas, Basilis Gatos, Verónica Romero Gómez
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 first 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. This will comprise mainly the organisation of the competitions as part of a well-known conference (e.g. ICDAR, ICFHR, etc.) and the announcement of the competition via dedicated lists and conferences.

We believe that research competitions are an important part of the research process in any applied field, and this is the case also with research in document processing. Research competitions are a good opportunity for labs to showcase the results of their newest algorithms. Past competitions, the announcement of the competitions related with this work package as well as the announcement of the common ScriptNet competitions platform have been greeted with positive remarks by related researchers.

In ScriptNet, the focus is on having handwritten document collections of historical significance as the testbed for the various document processing techniques. A typical example of a ScriptNet collection is the "Ratsprotokolle collection" [1], used in the ICFHR Handwriting Recognition competition (HTR 2016). This dataset consists of 450 handwritten pages written by multiple writers. Research groups have been invited to submit handwriting recognition methods; their methods have been automatically numerically evaluated (e.g. by measuring the number of correct words on the result) and published by the ScriptNet platform back-end [2]. Other research competitions of interest to ScriptNet include, but are not limited to : keyword spotting, document binarization, writer identification, baseline detection, word segmentation.

Concerning the related datasets, organizers have an option to have part of the datasets fully public, and part of the datasets held private. This is in the sense of availability of "ground-truth" information. For example, for HTR ground-truth would mean a perfect transcription of all word images. Concerning the test set, ground-truth is typically necessary to calculate numerical results, but not necessary to be publicly available (see also section 4 on this topic).

The research competitions that are organised by the READ consortium are scheduled to be integrated with the *ScriptNet platform*, a common competitions platform/site, under which the competitions run already, or are scheduled to run in the near future as on-going competitions.

[1] `http://stadtarchiv-archiviostorico.gemeinde.bozen.it/bohisto/Archiv/Handschrift/detail/14492`

[2] `https://scriptnet.iit.demokritos.gr/competitions/4/1/viewresults/`

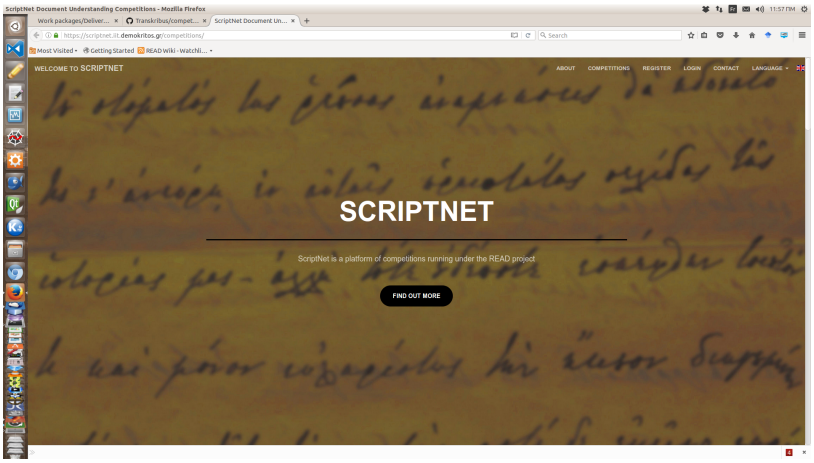
3 Overview of the ScriptNet platform

We have developed the ScriptNet platform in Django, a very popular and robust web-based framework [1]. All of the code is publicly available at github, at `https://github.com/Transkribus/competitions`. In the first year of the project, we have uploaded more than 300 commits for the Scriptnet platform and closed more than 30 issues and 8 pull requests on github.

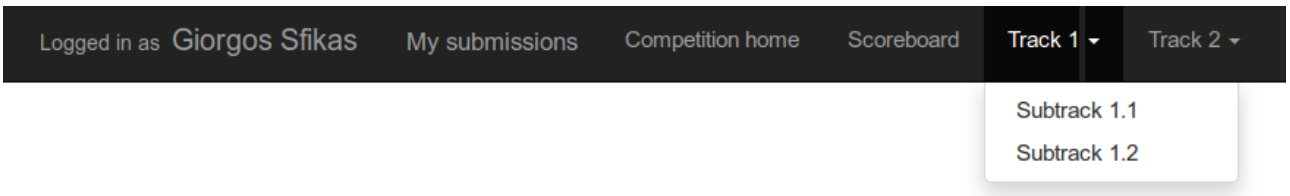
Competitors (research groups, autonomous researchers, scholars) can register on the platform, select an active competition, and submit their method. *Competition organizers* (typically research groups, members or non-members of the READ consortium) can create and customize a new competition through a point-and-click interface. Competition organizers can also upload their custom benchmarks/metrics.

The main features of the ScriptNet platform are as follows.

Authenticated users can log-in and submit the results of their methods through a simple and intuitive Bootstrap-based interface:



Each competition follows an hierarchy of competition - track - subtrack levels, that is easily manipulated by organisers, as well as easily accessible from competition participants:



Results are **automatically evaluated and presented** on the front-end:

POG	Retsinas et al. "Projections of Oriented Gradients"	Giorgos Retsinas	NCSR Demokritos / IIT / CIL	0.7026	0.8471	0.7486	0.6597	0.8624	0.8626

These can be compared versus already existing submissions:

All submissions and results for subtrack [1.1]Segmentation-based/Bentham database

Name	Method Info	Submitter	Affiliation	Result is public	Map	P@5	P@10	R-Precision	Ndcg-Binary	Ndcg
G1_SIMPLE	Kovalchuk, Alon, Lior Wolf, and Nachum Dershowitz. "A simple and fast word spotting method.", ICFHR 2014	Alon Kovalchuk	The Blavatnik School of Computer Science, Tel-Aviv University, Israel	✓	0.524	0.7381	0.6027	0.5024	0.742	0.7433
G2_ATTRIBUTES	Almazán, Jon, et al. "Word spotting and recognition with embedded attributes." IEEE Transactions on Pattern Analysis and Machine Intelligence 36.12 (2014): 2552-2566.	Jon Almazán	Computer Vision Center, Barcelona, Spain	✓	0.5126	0.7244	0.5777	0.4868	0.7442	0.7447
G3_INKBALL	Howe, Nicholas R. "Part-structured inkball models for one-shot handwritten word spotting.", ICDAR 2013	Nicholas R. Howe	Smith College Department of Computer Science	✓	0.4628	0.7181	0.5625	0.4591	0.6382	0.6418
POG	Retsinas et al. "Projections of Oriented Gradients"	Giorgos Retsinas	NCSR Demokritos / IIT / CIL	✓	0.7026	0.8471	0.7486	0.6597	0.8624	0.8626

A scoreboard of per-track ranking is also **automatically populated**:

Ranking for Track 1: Segmentation-based

Name	Method Info	Submitter	Affiliation	Result is public	Score
G2_ATTRIBUTES	Almazán, Jon, et al. "Word spotting and recognition with embedded attributes." IEEE Transactions on Pattern Analysis and Machine Intelligence 36.12 (2014): 2552-2566.	Jon Almazán	Computer Vision Center, Barcelona, Spain	–	15
G1_SIMPLE	Kovalchuk, Alon, Lior Wolf, and Nachum Dershowitz. "A simple and fast word spotting method.", ICFHR 2014	Alon Kovalchuk	The Blavatnik School of Computer Science, Tel-Aviv University, Israel	–	22
POG	Retsinas et al. "Projections of Oriented Gradients"	Giorgos Retsinas	NCSR Demokritos / IIT / CIL	–	11
G3_INKBALL	Howe, Nicholas R. "Part-structured inkball models for one-shot handwritten word spotting.", ICDAR 2013	Nicholas R. Howe	Smith College Department of Computer Science	–	32

Note also that the platform is internationalisation-ready, currently supporting three languages (English, French, Greek). The latest stable release of the platform is running at `https://scriptnet.iit.demokritos.gr/competitions`.

4 Timeline of integration with the ScriptNet platform

READ competitions are expected to be integrated with the ScriptNet platform, either from the date of their announcement or at least at some point of their lifetime.

An example of the timeline of integration of a READ competition, is as follows. The use case is the HTR 2016 competition:

- Beginning 2016: Competition announcement of HTR 2016.
- At ICFHR 2016: Competition results and paper are presented, winner(s) is(are) announced.
- After ICFHR 2016: HTR 2016 is turned into an on-going competition, integrated with scriptnet. No ground truth data for the test set is released at this point. Competitors can login and submit methods(results) to the platform, and check how their method fares related to other submissions.

- Beginning 2017: A new HTR competition is announced ("HTR 2017"), which will be related with ICDAR 2017. This will use a different dataset (training/test) than HTR 2016.
- End of 2017: Ground truth data for the test set is released for HTR 2016. The HTR 2016 competition is not on-going anymore. It will still be available at the scriptnet platform, but the scoreboard will be 'locked'. This means that new submissions will not change the publicly shown scoreboard, and new submissions will not be shown in public. Meanwhile, HTR 2017 will have become an on-going competition. The same cycle continues HTR 2017, and so on.

The reason that ground truth data for the test set is held back and released only towards the end of the lifetime of a competition is that, once GT is fully available it becomes trivial for anyone to simply submit the ground truth file and get a perfect score – in other words, cheat. While we do believe that the vast majority of our participants are not interested in any form of cheating, it is necessary for us to do whatever is possible to discourage and/or make cheating impossible.

5 Related datasets and DOI

All datasets used with the research competitions, either as training sets or ground truth test sets, are expected to have an assigned DOI. To this end we are using Zenodo as a dataset repository (`https://zenodo.org`).

6 Test competition integrated with the ScriptNet platform

In order to test the ScriptNet platform functionality, we have integrated a replica of the H-KWS 2014 competition [5] with our platform. We have recreated the track-subtrack structure of the competition, and uploaded all the result files of the original submissions to our platform. Results for various benchmarks (the same as the original benchmarks, i.e. MAP, Precision at 5, etc.) were automatically evaluated by the platform back-end mechanism. We have also uploaded a few new submissions to further test the platform; the platform ran smoothly, replicating exactly all numerical results and score-board rankings. The test competition is available as part of the github code (develop branch, `https://github.com/Transkribus/competitions/tree/develop`.

7 Organization of competitions in international conferences

7.1 ICFHR 2016 Handwritten Document Image Binarization competition

In H-DIBCO 2016 [6], the general objective is to record recent advances in document image binarization using established evaluation performance measures. The benchmarking dataset that is used in the contest augments the existing dataset of the DIBCO series containing handwritten document images that are representative of the potential problems which are challenging in the binarization process.

Contest web page: `https://vc.ee.duth.gr/h-dibco2016/`

7.2 ICFHR 2016 Handwritten Text Recognition competition

The "ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset" competition [4] organized in the framework of the ICFHR 2016 aims to bring together researchers working on off-line Handwritten Text Recognition (HTR) and provide them a suitable benchmark to compare their techniques on the task of transcribing typical historical handwritten documents. The proposed dataset consists of a subset of documents composed of minutes of the council meetings held from 1470 to 1805 (about 30.000 pages), which will be used in the READ project. This dataset is written in Early Modern German. The number of writers is unknown. Handwriting in this collection is complex enough to challenge the HTR software.

Contest web page: `https://scriptnet.iit.demokritos.gr/competitions/4/,http://transcriptorium.eu/~htrcontest/`

Dataset at Zenodo (doi:10.5281/zenodo.218236): `https://zenodo.org/record/218236`

7.3 ICFHR 2016 Keyword Spotting competition

The H-KWS 2016 [3], organized in the context of the ICFHR 2016 conference aims at setting up an evaluation framework for benchmarking handwritten keyword spotting (KWS) examining both the Query by Example (QbE) and the Query by String (QbS) approaches. Both KWS approaches were hosted into two different tracks, which in turn were split into two distinct challenges, namely, a segmentation-based and a segmentation-free to accommodate different perspectives adopted by researchers in the KWS field. In addition, the competition aims to evaluate the submitted training-based methods under different amounts of training data. The data used in the competition consisted of historical German and English documents with their own characteristics and complexities. For more details refer to:

Contest web page: `https://www.prhlt.upv.es/contests/icfhr2016-kws/index.html`

7.4 CLEF 2016 Keyword Spotting competition

In the context of the CLEF 2016 conference evaluation labs, a competition related to keyword spotting was organized [2]. Several novelties were introduced in comparison to other related contests, specifically: multiple word queries, finding local blocks of text, results in transition between consecutive pages, handling words broken between lines, words unseen in training and queries with zero relevant results. Four groups participated, one of which (URO) obtained very competitive results for the novel challenges of broken words and words unseen in training. Even though the competition has passed, it will be included in ScriptNet soon so that other groups can submit their new developments. The dataset prepared for this competition has been left publicly available at Zenodo.

Contest web page: `http://imageclef.org/2016/handwritten`

Dataset at Zenodo (doi:10.5281/zenodo.52994): `https://zenodo.org/record/52994`

Overview presentation: `http://imageclef.org/system/files/Villegas16_CLEF_Handwritten-Overview_presentation.pdf`

Overview paper: `http://ceur-ws.org/Vol-1609/16090233.pdf`

8 Contests integrated in ScriptNet

Before the end of the first year of the project, two competitions have been integrated with the Scriptnet competitions platform, running at `https://scriptnet.iit.demokritos.gr/competitions/` . These are:

- The "ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset" competition
- The new "ICDAR 2017 Competition on Baseline Detection" competition

References

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

[2] M. Villegas, J. Puigcerver, A.H. Toselli, J.A. Sánchez, E. Vidal, *Overview of the ImageCLEF 2016 Handwritten Scanned Document Retrieval Task.* In: CLEF2016 Working Notes. CEUR Workshop Proceedings, vol. 1609, pp. 233–253. CEUR-WS.org, Évora, Portugal (September 5-8 2016)

[3] 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.

[4] 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.

[5] 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).

[6] I. Pratikakis, K. Zagoris, G. Barlas, B. Gatos, *ICFHR 2016 competition on Handwritten Document Image Binarization (H-DIBCO 2016)* In "Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR 2016)"