RECOGNITION & ENRICHMENT OF ARCHIVAL DOCUMENTS

D7.3
HTR Engine Based on HMMs P3

Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Enrique Vidal

UPVLC

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	31.12.2018
Actual date of delivery	31.12.2018
Date of last update	31.12.2018
Deliverable number	D7.3
Deliverable title	HTR Engine Based on HMMs P3
Type	Demonstrator
Status & version	Final
Contributing WP(s)	WP7
Responsible beneficiary	UPVLC
Other contributors	UPVLC
Internal reviewers	Max Weidemann
Author(s)	Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Enrique Vidal
EC project officer	Christophe DOIN
Keywords	Handwritten Text Recognition, Hidden Markov models

Contents

Executive summary

This report describes the research developed in the third period (P3) of the READ project on Handwritting Text Recognition based on Hidden Markov Models. The laia tool that was developed in P2 has been tested on several collections. This tool combines Hidden Markov Models and Deep Neural Networks. Several collections have been researched in this period and the obtained results are described here.

1 Introduction

Classical Handwritten Text Recognition (HTR) borrows concepts and methods from the field of Automatic Speech Recognition, such as Hidden Markov Models (HMM), n-grams and Neural Networks (NN) [4, 1]. In recent years, pure NN-based methods have achieved impressive results in HTR [11, 9, 10]. But for taking profit of these advantages, NN in combination with HMM allows a seamless integration of language models for decoding. Furthermore, HMM-based techniques have very attractive characteristics that make them very convenient for several problems: there exist efficient techniques for dealing with lattice-based techniques (as it happens in Task 2.2 and Task 2.5 of READ), and the decoding problem is well known for HMM-based HTR.

1.1 Task 7.1 - Hidden Markov Model-based HTR

The problem of HTR can be stated formally as follows:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{x}) = \arg\max_{\mathbf{w}} P(\mathbf{x} \mid \mathbf{w}) P(\mathbf{w}) \tag{1}$$

where $\hat{\mathbf{w}}$ is the best transcript for the line image $\mathbf{x}$ among all possible transcripts $\mathbf{w}$. $P(\mathbf{x} \mid \mathbf{w})$ represents the optical modelling that is approximated with HMM in this task and $P(\mathbf{w})$ is the language model (LM) that is approximated with n-grams. The relevant contribution in this P3 is that the *Laia Toolkit*[1] [5], developed in P2 has been extensively tested in many collections. This tool trains some parameters of the HMM with Deep Neural Networks (DNN) techniques following [3] and [1].

Training $P(\mathbf{w})$ is currently easy since only plain text is necessary. Language model training is mainly researched in Task 7.4-Language modelling.

Training HMM for computing $P(\mathbf{x} \mid \mathbf{w})$ is more difficult since it is necessary to have line images and their corresponding diplomatic transcripts, each line with its corresponding transcript. The emission parameters of the HMM are trained with DNN-based techniques that have allowed very impressive improvements.

Task 7.1 in READ is related with research on HMM-based techniques for HTR. This means that both training techniques and decoding techniques are researched and developed. In P3 we have focused on training and testing the HMM-based techniques using the *Laia Toolkit*

The new results obtained in P3 with HMM-based HTR with HMM trained with DNN approaches are shown in Section 2.

[1] `https://github.com/jpuigcerver/Laia`

2 Results on HMM-based HTR with DNN training approaches

The tool developed by the UPVLC team [6, 5] is a free software tool that is now in the UPVLC workflow for HTR. The following sections show comparative HTR results with respect to other research groups on public datasets that have been prepared in the READ project.

2.1 HTR with the competition collections

We tested the HMM-DNN-based tool developed by UPVLC with some of the datasets used in the HTR competitions that have been held in the ICFHR 2014 conference, the ICDAR 2015 conference, the ICFHR 2016 conference, and the ICDAR 2017 conference. These datasets are publicly available (see Table 1).

Table 1: The datasets described in this section are publicly available for research purposes at the following web links.

Dataset	Web link
ICFHR-2014	http://doi.org/10.5281/zenodo.44519
ICDAR-2015	http://doi.org/10.5281/zenodo.248733
ICFHR-2016	http://doi.org/10.5281/zenodo.1164045
ICDAR-2017	http://doi.org/10.5281/zenodo.835489

Results on the ICFHR-2014 dataset and on the ICFHR-2016 dataset were reported in P2 report. We also report these results for completeness for ICFHR-2014 dataset and new results for the ICFHR-2016 dataset. We now summarize the results obtained with these datasets.

2.1.1 The ICFHR-2014 Dataset

The data was taken from a large set of manuscripts with about $80,000$ documents written by the renowned English philosopher and reformer Jeremy Bentham (1748-1832).

The dataset for this competition was composed of 433 page image, each encompasing of a single text block in most cases. These 433 pages contained $11,537$ lines with nearly $110,000$ running words and a vocabulary of more than $9,500$ different words. The last column in Table 2 summarises the basic statistics of these pages. More details are provided in [8].

Table 2: The Bentham dataset used in the ICFHR-2014 competition.

Number of:	Training	Validation	Test	Total
Pages	350	50	33	433
Lines	9,198	1,415	860	11,473
Running words	86,075	12,962	7,868	106,905

Two tracks were planned in this competition: i) *Restricted track*: participants were allowed to use just the data provided by the organisers for training and tuning their systems; ii) *Unrestricted track*: participants were allowed to use any data of their choice.

Table 3 shows a summary of the most relevant results obtained in the *Restricted track*. We observe clearly better results with respect to previously published papers. HMM-DNN-based systems were also used in [8] and [2].

Table 3: Results obtained with the test set of the ICFHR-2014 dataset in the *Restricted track*.

Reference	WER	CER
[2]	14.1	5.0
P2. HMM-DNN-based system	**9.7**	**5.0**

2.1.2 The ICDAR-2015 Dataset

The ICDAR-2015 dataset contains more difficult pages from a layout analysis point of view, drawn again from the Bentham collection. The images have marginal notes, fainted writing, stamps, skewed images, lines with different slope in the same page, variable slanted writing, inter-line text, etc.

The dataset was divided into four subsets as shown in Table 4: *Train-B1*, with line images aligned with their line transcripts, was intended for training (this is the whole ICFHR-2014 dataset); *Train-B2*, also intended for training, was provided only with page-level transcripts, i.e., without alignment of line images with line transcripts; *Test*, to be used for evaluating the HTR results.

Table 4: Main statistics of the ICDAR-2015 dataset.

Number of:	Train-B1	Train-B2	Training	Test	Total
Pages	433	313	746	50	796
Lines	11,473	8,947	20,420	1,332	21,752
Running words	106,905	70,447	177,352	9,440	186,792

The same tracks defined in the ICFHR 2014 competition were defined in this dataset: a *Restricted track* and an *Unrestricted track*.

Table 5 shows a summary of the most relevant results obtained in the *Restricted track*. We observe clearly better results with respect to previously published papers.

Table 5: Test set CER & WER for the ICDAR-2015 dataset in the *restricted track*.

References	WER	CER
[12, 11]	30.2	15.5
P3. HMM-DNN-based system	**30.0**	**12.8**

2.1.3 The ICFHR-2016 Dataset

In this edition, German was chosen for the contest. The proposed dataset consisted of a subset of documents from the Ratsprotokolle collection[2] composed of minutes of the council meet-

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

ings held from 1470 to 1805 (about 30.000 pages), which is used in the READ project. This dataset is written in Early Modern German.

The dataset for this competition was composed of 450 page images, each encompassing of a single text block in most cases, but also with many marginal notes and added interlines. These 450 pages contained $10,550$ lines with nearly $43,500$ running words and a vocabulary of more than $8,000$ different words. The last column in Table 6 summarizes the basic statistics of these pages. More details are provided in [7].

Table 6: The Ratsprotokolle dataset used in the HTR contest.

Number of:	Train	Validation	Test	Total
Pages	350	50	50	450
Lines	8,367	1,043	1,140	10,550
Running words	35,169	3,994	4,297	43,460

Two tracks were planned in this competition: i) *Restricted track*: participants were allowed to use just the data provided by the organizers for training and tuning their systems; ii) *Unrestricted track*: participants were allowed to use any data of their choice.

Table 7 shows a summary of the most relevant results obtained in the restricted track. We observe better results at word level with respect to previously published papers. A HMM-DNN-based system was also used in [7].

Table 7: Results obtained with the test set of the ICFHR-2016 dataset in the *Restricted track*.

Reference	WER	CER
[7]	20.9	4.8
P2. HMM-DNN-based system	18.1	4.6
P3. HMM-DNN-based system	**17.5**	**4.5**

2.1.4 The ICDAR-2017 Dataset

Most of the images used in the ICDAR-2017 benchmark were taken from the Alfred Escher Letter Collection (AEC)[3] collection, but handwritten text images from other German collections of the same period were also included. Many of these extra images are of poor quality and/or low resolution. Overall, the text considered in this benchmark has been written by several hands. The dataset encompasses $10\,172$ page images, divided into four subsets: two for training (Train-A and Train-B) and two for testing (Test-A and Test-B2) (see Table 8).

Train-A consists of 50 page images, each including one or more text blocks, making a total of about $1,000$ lines. GT was produced semi-automatically and manually reviewed at line level. The second training subset (Train-B) has $10,000$ images with around $200,000$ lines. In this subset, no geometric information about the location of the text lines in the images is provided, but the corresponding transcripts do have correct line breaks according to how lines appear in the images. Note that this information is relevant since it can be exploited to

[3] `https://www.briefedition.alfred-escher.ch/`

Table 8: Main statistics of the ICDAR-2017 dataset.

Number of:	Train-A	Train-B	Total Train	Test-A	Test-B2
Pages	50	10,000	10,050	65	57
Lines	1,386	204,775	206,161	1,573	1,412
Running words	15,169	1,754,026	1,769,195	14,880	14,460

improve line detection which, in turn, can help to automatically obtain more transcript-aligned line images for training.

Finally, the subsets Test-A and Test-B respectively contain 65 and 57 page images. Images in the first subset are annotated with baselines, while those in second include only rough geometry of regions where lines may be detected and recognized.

Two tracks are defined for this benchmark being one of the a *traditional* track. Table 9 shows a summary of the most relevant results obtained with this dataset in ICDAR 2017.

Table 9: CER and WER obtained for Test-A in the *Traditional challenge* of ICDAR-2017.

References	WER	CER
[10]	19.1	7.0
P3. HMM-DNN-based system	**17.6**	**5.8**

References

[1] T. Bluche. *Deep Neural Networks for Large Vocabulary Handwritten Text Recognition*. PhD thesis, Ecole Doctorale Informatique de Paris-Sud - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, may 2015. Discipline : Informatique.

[2] T. Bluche. *Deep Neural Networks for Large Vocabulary Handwritten Text Recognition*. PhD thesis, Université Paris Sud - Paris XI, May 2015.

[3] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Tr. PAMI*, 31(5):855–868, 2009.

[4] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[5] J. Puigcerver. Are multidimensional recurrent layers really necessary for handwritten text recognition? In *Proc. ICDAR*, pages 67–72, 2017.

[6] J. Puigcerver, D. Martin-Albo, and M. Villegas. Laia: A deep learning toolkit for htr. `https://github.com/jpuigcerver/Laia`, 2016. GitHub repository.

[7] J.A. Sánchez and U. Pal. Hanwrittent text recognition for bengali. In *Proceedings of the 2016 International Conference on Frontiers in Handwriting Recognition*, pages 542–547, 2016.

[8] J.A. Sánchez, V. Romero, A.H. Toselli, and E. Vidal. ICFHR2014 competition on handwritten text recognition on transcriptorium datasets (HTRtS). In *ICFHR*, pages 181–186, 2014.

[9] J.A. Sánchez, V. Romero, A.H. Toselli, and E. Vidal. ICFHR2016 competition on handwritten text recognition on the READ dataset. In *Proceedings of the 2016 International Conference on Frontiers in Handwriting Recognition*, pages 630–635, 2016.

[10] J.A. Sánchez, V. Romero, A.H. Toselli, M. Villegas, and E. Vidal. Icdar2017 competition on handwritten text recognition on the read dataset. In *Proc. International Conference on Document Analysis*, pages 1383–1388, 2017.

[11] J.A. Sánchez, A.H. Toselli, V. Romero, and E. Vidal. ICDAR 2015 competition HTRtS: Handwritten text recognition on the tranScriptorium dataset. In *13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015.

[12] T. Strauß. *Decoding the Output of Neural Networks. A Discriminative Approach*. PhD thesis, Universität Rostock, June 2017.