READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D7.2
HTR Engine Based on HMMs P2

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	31.12.2017
Actual date of delivery	31.12.2017
Date of last update	31.12.2017
Deliverable number	D7.2
Deliverable title	HTR Engine Based on HMMs P2
Type	Demonstrator
Status & version	Final
Contributing WP(s)	WP7
Responsible beneficiary	UPVLC
Other contributors	UPVLC
Internal reviewers	Johannes Michael, Max Weidemann, Nathanael Philipp, Günter Mühlberger
Author(s)	Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Enrique Vidal
EC project officer	Martin Majek
Keywords	Handwritten Text Recognition, Hidden Markov models

Contents

1 Introduction 4

1.1 Task 7.1 - Hidden Markov Model-based HTR 4

2 Results on HMM-based HTR with DNN training approaches 5

2.1 HTR with the “Siglo de Oro” collection 5

2.2 HTR with the competition collections 7

2.2.1 The ICFHR-2014 Dataset 7

2.2.2 The ICFHR-2016 Dataset 8

2.3 HTR with the “Passau” collection 8

3 Results on HMM-based HTR with classical training approaches 9

3.1 HTR with the “Oficio de Hipotecas de Girona” collection 9

3.2 Retrieval information approaches based on HMMs applied to Handwritten documents 10

4 Plans for the next period 10

Executive summary

This report describes the research developed in the second period (P2) of the READ project on Handwritting Text Recognition based on Hidden Markov Models. A new tool has been developed that combines Hidden Markov Models and Deep Neural Networks. Several collections have been researched in this P2 and the current 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) [5, 4, 10, 1]. In recent years, pure NN-based methods have achieved impressive results in HTR [14, 13]. 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 P2 is that some parameters of the HMM are trained with Deep Neural Networks (DNN) techniques following [4] and [1]. The models involved in this expression are trained from examples.

Training $P(\mathbf{w})$ is currently easy since only plain text is necessary. This text can be obtained from the web or from linguistic resources. Usually, the more text in order to capture the frequency of word (or character) sequences adequately the better. 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. So there is no need for segmented words nor characters for training the optical models. In the last years, the emission parameters of the HMM are trained with DNN-based techniques that have allowed a 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 P2 we have focused on adapting the HMM-based techniques for training the HMM parameters with DNN techniques with a new software tool developed by UPVLC team that is known as *Laia*. This software is publicly available for research purposes [7].

The new results obtained in P2 with HMM-based HTR with HMM trained with DNN approaches are shown in Section 2. Section 3 describes the new obtained results in P2 with the classical method for training HMM.

2 Results on HMM-based HTR with DNN training approaches

The new tool developed by UPVLC team [7, 6] is a free software tool that has been tested on several collections and several experiments have been performed. The following sections show comparative HTR results with respect to P1 with this new tool.

2.1 HTR with the "Siglo de Oro" collection

During P2 we have been working with the documents written by Lope de Vega available at the Biblioteca Nacional de España (BNE)[1]and collaboration of the ProLope research group[2]. Both the BNE and ProLope are MOU partners.

Lope de Vega was a Spanish writer, poet and novelist. He was one of the key figures in the Spanish Golden Century of Baroque literature. Around $3,000$ sonnets, 3 novels, 4 novellas, 9 epic poems and about 500 plays are attributed to him. At the BNE there are around 250 registers whose authorship is Lope de Vega. However, not all of these documents are autographs, and there are documents written by several copyists. Figure 1 shows some samples.

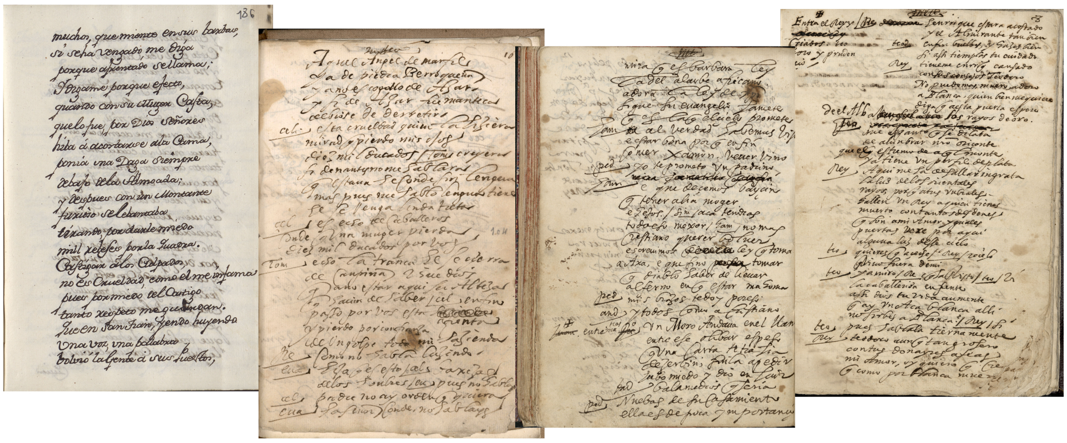

Figure 1: Pages of the Lope collection.

During P1 of the READ project, some HTR experiments with one document written by a copyist (Sanz de Pliegos) were carried out (see Deliverable D7.1). During this P2, we carried out comparative experiments with the new software developed in the UPVLC.

[1] http://www.bne.es/es/Inicio/index.html

[2] http://prolope.uab.cat/

The comedy used in these experiments is called *"El cuerdo loco"* and it is composed of 209 pages. These pages were manually annotated with the layout analysis of each page to indicate text blocks and lines and also they were completely transcribed line by line by an expert paleographer. The $3,995$ lines that compose the document were divided into 10 different partitions in order to carry out cross-validation experiments. Table 1 shows some statistics. Column "Average" is the average for the 10 partitions.

Table 1: Statistics of *"El cuerdo loco"* manuscript.

Number of:	Average	Total
Pages	20.9	209
Line	399.5	3,995
Run. words	2,287	22,873

The recognition errors obtained in this document are shown in Table 2. Using the traditional HTR system, the Word Error Rate (WER) is around 57% at word level and 26% at character level (CER) (see Deliverable D7.1). Using the new system the obtained results improved to 26% at word level and around 9% at character level.

Table 2: Results obtained in "El cuerdo loco" manuscript.

	WER	CER
P1. HMM-based system	56.8	26.3
P2. HMM-DNN-based system	26.2	8.7

On the other hand, during this P2, the work has been focused in the Lope's autographs. We chose two Lope's autographs called *"El cordobés valeroso, Pedro Carbonero"* and *"La prueba de los amigos"* and carried out experiments to test the HMM-DNN-based HTR approach. The selected documents were composed by around 400 pages ($6,783$ lines). These two documents are included in Transkribus.

The pages were manually annotated with the layout analysis of each page to indicate text blocks and lines and also they were completely transcribed line by line by an expert paleographer. We divided the pages between train and validation and we trained a HMM-DNN system. We also trained a 7-gram character language model using the transcripts of around 150 Lope comedies. The recognition errors are shown in Table 3.

Table 3: Results obtained two Lope's autographs.

	WER	CER
P2. HMM-DNN-based system	33.5	11.8

Finally, it was carried out the automatic segmentation, recognition and production of lattices for 14 Lope de Vega's autographs (for which no transcript is yet available). Thwaw lattices could be used for interactive transcription or for indexing. In period 3 (P3) we intend to exploit this collection for comprehensive experiments and a publication with the results will be prepared.

2.2 HTR with the competition collections

We tested also the new 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 and in the ICFHR 2016 conference. These datasets are publicly available (see Table 4).

Table 4: The datasets described in this section are publicly available for research purposes at the following web links.

Dataset	Web link
ICFHR-2014	`https://zenodo.org/record/44519#.WeIlkHV-qkA`
ICFHR-2016	`https://zenodo.org/record/218236#.WeImcXV-qkA`

We now summarize the results obtained with these datasets.

2.2.1 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 5 summarises the basic statistics of these pages. More details are provided in [12].

Table 5: 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 6 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 [12] and [2].

Table 6: Results obtained with the test set of the ICFHR-2014 dataset in the *Restricted track*.

Reference	WER	CER
[12]	14.6	5.0
[2]	14.1	5.0
P2. HMM-DNN-based system	9.0	5.0

2.2.2 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[3] composed of minutes of the council meetings 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 7 summarizes the basic statistics of these pages. More details are provided in [11].

Table 7: 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 8 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 [11].

Table 8: Results obtained with the test set of the ICFHR-2016 dataset in the *Restricted track*.

Reference	WER	CER
[11]	20.9	4.8
P2. HMM-DNN-based system	18.1	4.6

2.3 HTR with the “Passau” collection

This is a XVI-XVIII century collection of historical records, which involves around 26,000 images written in German. Most of the information contained in such images is handwritten

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

in tables, which should have to be previously detected by a layout analysis process. However, for the moment, the layout analysis of such tables is manually done including the text line detection within each of their cells. Once text line images were detected and extracted, they were transcribed using a HMM-DNN-based system.

So far the last obtained performance WER figure on a small subset of this collection is around 58% (200 pages for training and 90 for testing). It is remarkable that this collections is really difficult from an HTR point of view. See a example of this collection in Figure 2. Since the sentences are very short the contribution of the language model is not very significant.

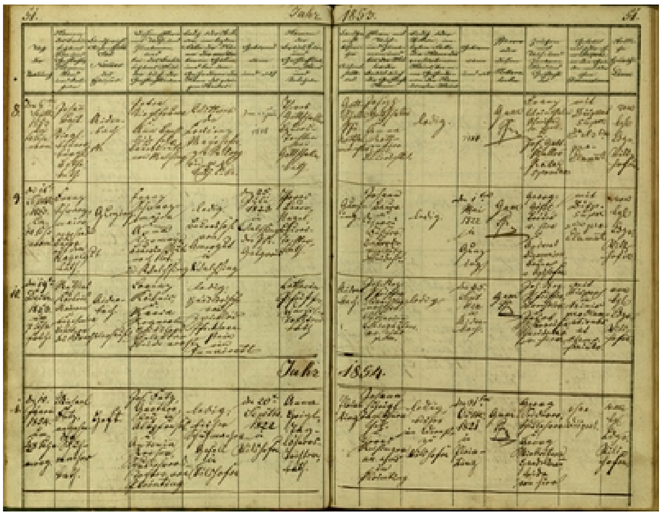

Figure 2: Pages of the Passau collection.

For P3 we intend to research on this collection for comprehensive experiments and a publication with the results will be prepared.

3 Results on HMM-based HTR with classical training approaches

During this P2 new HTR experiments with different READ collections have been carried out with classical approaches used for training the optical models. The following subsections summarize these results.

3.1 HTR with the “Oficio de Hipotecas de Girona” collection

This collection was provided by *Centre de Recerca d’História Rural* from the Universitat de Girona (MOU partner of the READ project). The transcription process based on HMM-based engine has been carried out in batches following the Interactive-Predicive approach (see Deliverables D7.4 and D7.5) of around 50 pages each, where previously transcribed batches are used to (re-)train system models in order to improve the accuracy of the models on each iteration. Besides the transcription of the collection, it is also pursued the semantic tagging of each recognized word (toponyms, anthroponyms, trades, registry typology, etc.), which was accomplished simultaneously with its recognition by the HTR engine. So far, by using the HMM-based system with Gaussian Mixture Models (GMM) as emission probabilities, the last obtained performance WER figure (which includes tagging) is around 27% (see Table 2 in Deliverables D7.5 for additional details).

For P3 we intend to research on this collection for comprehensive experiments and a publication with the results will be prepared.

3.2 Retrieval information approaches based on HMMs applied to Handwritten documents

In this case, it was proposed an information extraction approach for Handwritten Marriage Licenses Books using the MGGI methodology. These books follow a simple structure of the text in the records with a evolutionary vocabulary, mainly composed of proper names that change along the time. This distinct vocabulary makes automatic transcription and semantic information extraction difficult tasks (see Deliverable D7.1, page 9 and [8] for additional information).

In previous works [8], we studied the use of category-based language models and how a Grammatical Inference technique known as MGGI could improve the accuracy of these tasks. In this P2, we analyzed the main causes of the semantic errors observed in previous results and we applied a better implementation of the MGGI technique to solve these problems. Using the resulting language model, transcription and information extraction experiments have been carried out, and the results support the proposed approach [9]. Table 9 shows these results.

Table 9: Word Error Rate (WER), precision (π) and recall (ρ) obtained with the MGGI HTR systems (MGGI). The mean is computed for the absolute number of instances (I) and for categories (C). All results are percentages.

	WER	I-π	I-ρ	C-π	C-ρ
P1 [8]	10.1	85.3	76.2	78.3	72.2
P2 [9]	10.1	87.8	82.3	80.7	76.2

This approach has also been used in the baseline of the “ICDAR2017 Competition on Information Extraction in Historical Handwritten Records”, where we participated in the organization of this competition [3].

4 Plans for the next period

The main plans for the following period include:

- To transcribe the most “Siglo de Oro” collection and prepare word and/or character lattices for them being used in KWS.

- To complete experiments with all datasets used in the competitions and to improve the current results, specially at word level.

- To process more data from the "Passau" collection.
- To collaborate with *Centre de Recerca d'História Rural* for producing word and/or character lattices for them being used in KWS and interactive transcription.

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. Fornés, V. Romero, A. Baró, J.I. Toledo, J.A. Sánchez, E. Vidal, and J. Lladós. Icdar2017 competition on information extraction in historical handwritten records. In *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017)*, pages 1389–1394, 2017.

[4] 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.

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

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

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

[8] V. Romero, A. Fornés, J.A. Sánchez, and E. Vidal. Using the MGGI methodology for category-based language modeling in handwritten marriage licenses books. In *Proceedings of the 2016 International Conference on Frontiers in Handwriting Recognition*, pages 331–336, 2016.

[9] V. Romero, A. Fornés, E. Vidal, and J.A. Sánchez. Information extraction in handwritten marriage licenses books using the MGGI methodology. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 287–294, 2017.

[10] V. Romero, A.H. Toselli, and E. Vidal. *Multimodal Interactive Handwritten Text Transcription*. Series in Machine Perception and Artificial Intelligence (MPAI). World Scientific Publishing, 2012.

[11] 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.

[12] 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.

[13] 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.

[14] 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.