ICFHR2014 Competition on Handwritten Text Recognition on tranScriptorium Datasets (HTRtS)

Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Enrique Vidal
Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València
València 46022 (Spain)
{*jandreu,vromero,atoselli,evidal*}*@prhlt.upv.es*

***Abstract*—A contest on Handwritten Text Recognition organised in the context of the ICFHR 2014 conference is described. Two tracks with increased freedom on the use of training data were proposed and three research groups participated in these two tracks. The handwritten images for this contest were drawn from an English data set which is currently being considered in the tranScriptorium project. The the goal of this project is to develop innovative, efficient and cost-effective solutions for the transcription of historical handwritten document images, focusing on four languages: English, Spanish, German and Dutch. For the English language, the so-called "Bentham collection" is being considered in tranScriptorium. It encompasses a large set of manuscripts written by the renowned English philosopher and reformer Jeremy Bentham (1748-1832). A small subset of this collection has been chosen for the present HTR competition. The selected subset has been written by several hands (Bentham himself and his secretaries) and entails significant varibilities and difficulties regarding the quality of text images and writting styles. Training and test data were provided in the form of carefully segmented line images, along with the corresponding transcripts. The three participants achieved very good results, with transcription word error rates ranging from 15.0% down to 8.6%.**

***Keywords*-Handwritten Text Recognition**

I. Introduction

tranScriptorium [1][1] is a three years project which aims to develop innovative, efficient and cost-effective solutions for the indexing, search and full transcription of historical handwritten document images.

Currently, huge amounts of these documents are being published by on-line digital libraries worldwide. For these raw digital images to be really useful, they need be transcribed and/or indexed. For typical handwritten text images of historical documents, traditional Optical Character Recognition (OCR) is simply not usable since characters can not be isolated automatically in these images. Therefore, holistic, segmentation-free Handwritting Text Recognition (HTR) techniques are needed which not require any explicit character or word segmentation. Current technology for HTR borrows concepts and methods from the field of Automatic Speech Recognition, such as Hidden Markov Models (HMMs), Neural Networks and N-grams [2], [3], [4]. These models are trained from samples by using efficient techniques. A remarkable characteristic of these techniques is that few human resources are needed to develop useful transcription systems.

To this end, tranScriptorium aims to research on modern, holistic Handwritten Text Recognition (HTR) technology. It is focusing on four languages: English, Spanish, German and Dutch. For the English language, a large set of manuscripts written by the renowned English philosopher and reformer Jeremy Bentham (1748-1832)[5] has been chosen. The digitised material include legal reform, punishment, the constitution, religion, and his panopticon prison scheme. The Bentham Papers include manuscripts written by Bentham himself over a period of sixty years, as well as fair copies written by Bentham's secretarial staff. The transcription of this collection is currently being carried out by amateur volunteers participating in the award-winning crowd-sourcing initiative known as "Transcribe Bentham"[2].

Page images of the Bentham collection generally entail important layout analysis difficulties (see Fig. 1), like marginal notes, fainted writing, stamps, skewed images, lines with different slope in the same page, slanted lines, inter-line sentences, etc.

It is also difficult from the HTR point of view. It is written by several hands, it has many crossed-out and hyphenated words, etc. Portions of the collection are written in French, and Bentham occasionally used Latin and ancient Greek in his writings. Preliminary HTR results on a small set of 53 pages of the Bentham collection were reported in [6] by using some of the HTR techniques previously mentioned.

The Bentham collection has more than 80,000 documents, most of them digitised. From the digitised documents, more than 6,000 have been transcribed with the crowd-sourcing platform previously mentioned. The transcripts are recorded in TEI-compliant XML format. Given the nature of the transcription process, the transcripts produced by the amateur volunteers are not completely consistent, and therefore the transcripts are finally reviewed by expert transcribers. It is

[1] http://www.transcriptorium.eu/

[2] http://blogs.ucl.ac.uk/transcribe-bentham/

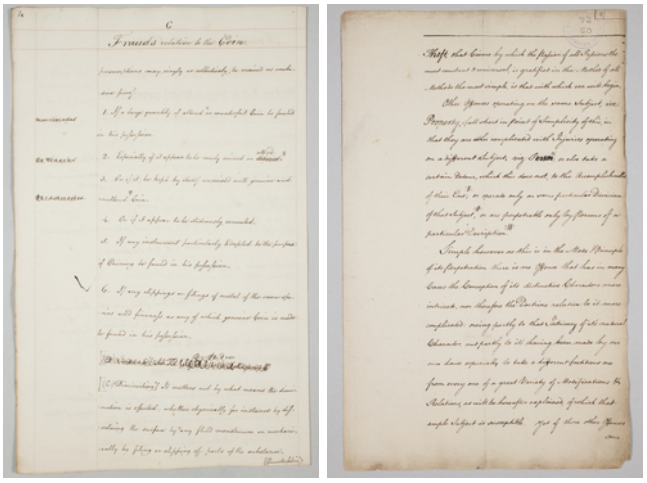

Figure 1. Document samples of the Bentham dataset to be processed in TRANSCRIPTORIUM.

worth noting that no geometric information is registered through the crowd sourcing annotation process. Therefore one of the first tasks in TRANSCRIPTORIUM was to perform line segmentation and to pair the extracted lines and their corresponding transcripts [6]. Line detection and text alignment is the only preprocessing step strictly necessary for training HTR models.

This paper describes the HTRtS contest that was carried out with part of this dataset in the context of the International Conference on Frontier in Handwriting Recognition (ICFHR) 2014. HTRtS was organised by members of the Pattern Recognition and Human Language Technology research center that participate in TRANSCRIPTORIUM, with the help of other members of the consortium. In this first edition of the competition, seven research groups were registered and finally three participants actually tested their systems and submitted official results.

Section II describes the dataset in more detail. Section III describes how the competition was organised. The main characteristics of the participant systems are described in Section IV and their official results are reported in Section V.

II. DATA DESCRIPTION

The dataset for this competition was composed of 433 page images[3], each encompasing of a single text block in most cases. These pages entailed several line detection and transcription difficulties and the corresponding ground truth was produced semi-automatically.

On the one hand, the transcripts for these pages were available since they were transcribed with the Transcription Desk tool of the "Transcribe Bentham" project[4]. On the other hand automatically obtained line regions [6] were manually revised with the Aletheia tool [7] (see examples of extracted lines in Fig. 2). The following step was to pair each line transcript with its corresponding line image. Given the large amount of transcribed data, it was not feasible to perform these pairings manually, and therefore a semi-automatic procedure was carried out [6]. The ground truth information was registered in PAGE format [8]. TEI[5] marks were removed and ignored for the contest.

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 I summarises the basic statistics of these pages.

Table I
THE BENTHAM DATASET USED IN THE HTRTS CONTEST.

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
Lexicon	8,658	2,709	1,946	9,716
Running OOV	-	857	417	-
OOV Lexicon	-	681	377	-
Character set size	86	86	86	86
Run. Characters	442,336	67,400	40,938	550,674

The data set were divided into three subsets for training, validation and test, respectively encompassing 350, 50 and 33 images. Since it was not possible to accurately identify the writers in all cases, the pages were shuffled before distributing them over these three subsets. This means that some writers can appear both in the training and in the test sets. Table I contains basic statistics of these partitions. The rows "Running word" and "Running OOV" show the total number of words and Out-Of-Vocabulary (OOV) words, respectively. The OOV words in the Validation column are words that do not appear in the training set, while the those in the Test column are words that do not appear neither in the training nor in the validation sets. The row "OOV Lexicon" shows the number of *different* running OOV words.

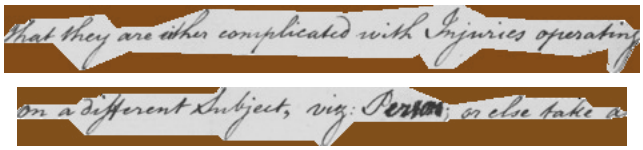

Figure 2. Sample lines such as they were provided for experimentation.

A noticeable aspect of this dataset is the existence of many short lines that are mainly, numbers, section headers, subsection item symbols, and added words that in this dataset were considered as separated lines (see an example of added word above line 5 in left image in Fig 1). Fig 3 shows the sentence length histogram of the training and validation sets. Note that there are many short lines; for example, 9.5% of the lines have just a word. It should be emphatized that word n-gram language models are of little help to capture context in this kind of lines.

[3]The dataset is free available at the TRANSCRIPTORIUM web page.
[4]http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham
[5]http://www.tei-c.org/index.xml

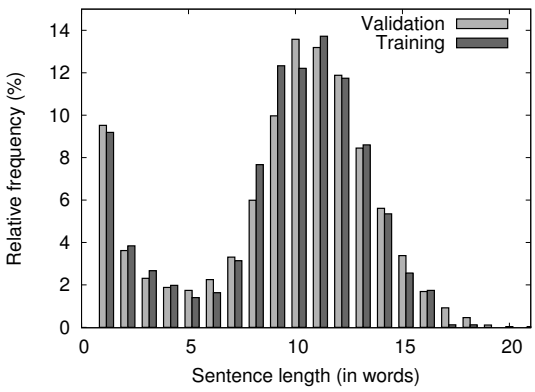

Figure 3. Normalised sentence length histograms.

III. Competition Protocol

The training and validation sets described in the previous section were provided to the participants as soon as the competition became open, while the test part was kept hidden and released in due time just to obtain the results to be evaluated and compared. The training and validation data available for the participants consisted of:

- The original images of all the training and validation pages.
- The PAGE file corresponding to each page image. For each text line in this image, the PAGE file contained a bounding polygon and the corresponding correct transcript.
- The preprocessed and extracted line images for all the lines of the training and validation sets in grayscale (see examples in Fig. 2).
- The corresponding transcripts of each of these lines. These transcripts had the punctuation symbols separated from words, and the final results on the test set had to be submitted in the same way.

The first pair of items was redundant with the second and was provided for those who wished to try improving results by using specific image pre-processing and line extraction tools. Note that the purpose of providing the lines correctly detected was to focus the contest on HTR and not in layout analysis and line detection and extraction. The test images, with the transcript fields empty in the PAGE file, were eventually provided in the same (redundant) formats for evaluation purposes. The results had to be provide with capital letters correctly detected.

A baseline system based on hidden Markov models trained with the Hidden Markov Model Toolkit[6] (HTK) and n-gram models trained with the SRILM[7] toolkit was provided, including a set of scripts to perform a basic training and test experiment. The participants could use this baseline system as an initial approach. They were allowed to improve this baseline by changing one or several of the following processes:

- page-level pre-processing and line extraction.
- line pre-processing and normalisation.
- feature extraction.
- recognition system and/or approach.
- types of character and/or language models.
- etc.

The participants had to send the output transcripts for the test set. Several results per participant were allowed corresponding to different runs of their own systems and all the results were considered for the final decision.

The evaluation metric was the Word Error Rate (WER) between the reference transcripts of the test set and the recognition results. The winner would be the participant that obtained the least WER. A web-based platform was available for the participants to submit their recognition results.

Two tracks were planned in this competition:

- Restricted track: participants were allowed to use just the data provided by the organisers for training and tuning their systems.
- Unrestricted track: participants were allowed to use any data of their choice.

The purpose of defining two tracks was to have the possibility of comparing techniques with respect to the amount of training data used.

The competition was planned in such a way that the participants had seven weeks for preparing their systems before the test set was provided. Then, they had one week for sending their transcription results on the test set. Along that week, the participants did not receive any feedback about their results on the test data. When the competition closed, the competitors were informed only about their own results and they were asked for submitting a description of the system for which they obtained their best results. These descriptions are summarised in Section IV.

IV. Systems Description

Seven research groups registered at the contest and finally three of them submitted official results. Two participants submitted results of several systems to just one track and one participant submitted results of several systems to the two tracks described in Section III. The three research groups, listed in the same order they registered were:

- Artificial Intelligence and Image Analysis (A2IA)[8].
- Computational Intelligence Technology Laboratory (CITlab)[9].
- Spoken Language Processing Group at Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI).[10]

[6]http://htk.eng.cam.ac.uk

[7]http://www.speech.sri.com/projects/srilm/

[8]http://www.a2ia.com/en
http://www.a2ialab.com/doku.php

[9]http://www.citlab.uni-rostock.de

[10]http://www.limsi.fr/Scientifique/tlp/

Table II includes the identifiers used to identify the best results that each partner submitted to each track.

Table II
IDENTIFIERS OF THE BEST RESULTS OBTAINED BY EACH PARTNER IN EACH TRACK.

	Restricted track	Unrestricted track
A2IA	-	A2IA-Un
CITlab	CITlab-Re	-
LIMSI	LIMSI-Re	LIMSI-Un

A2IA submitted results just to the unrestricted track, CITlab submitted results just to the restricted track and LIMSI submitted results both to the restricted and the unrestricted tracks.

- **A2IA-Un:** In the system developed by A2IA, the only pre-processing that was carried out on the images was to convert them to gray-scale and to normalise them. The Optical Model (OM) was a Multi-directional Long Short-Term Memory (MDLSTM) [9] Neural Network (NN), as described in [10]. The output are the word-tokens recognised by constraining the character predictions from the OM by a Language Model (LM). OM produced predictions for each character in the charset. The number of predictions depended on the width of the image. The OM was trained with multiple corpora [11], [12], [13] which included an undocumented database with approximately 153K lines of an historic digitised database[11] and then re-trained with the new Bentham data. The system used an hybrid word/character LM that comprised a top-level LM (3-gram on words/punctuation) for the most frequent words (30k) and a secondary-level LM (10-gram on characters) that dealt with OOV words [14]. Connectionist Temporal Classification (CTC) [15] and stochastic gradient descent, with a fixed learning ratio of 10e-3, were used to train the Recurrent NN (RNN). The LM was trained with and additional Bentham dataset that was gathered through the web[12]. This dataset amounted to 5.86M words, with a vocabulary of 57,324 word-tokens but vocabulary was limited to the 30k most frequent words in the training data. With this additional data the lexicon OOV decreased from 377 (see Table I) to 275, and the running OOV decreased from 417 (5.3%) to 313 (4.0%).
- **CITlab-Re:** CITlab system pre-processed the line images by first carrying out a contrast normalisation based on foreground/background pixel intensity levels, size normalisation and slant correction. For feature extraction, they used the same technique described in [15]. Optical modelling was carried out with NN. The lexicon was composed by all words in the training and validation sets and complemented by additional verbs with "-ing" and "-ed" endings. Hyphenated words were taken as one entire word instead of including its parts. CITlab used Backpropagation-Through-Time (BPTT) [16] using the CTC algorithm [15] for network training[13].
- **LIMSI-Re:** This system was in fact a systems combination, and therefore, this description summarises the main characteristics of these systems. In the pre-processing stage the images were converted to grey-level. Then, the lines were deslanted and the contrast was enhanced. All line images were normalised in height to 72px. Two type of features were researched and each system used just one of them: i) handcrafted features obtained from a sliding window [17], and ii) pixel features with a sliding window where the pixel values were normalised to lie in the interval $[0, 1]$. Hybrid NN/HMM models were used for modelling optical models. Two type of NN were considered and each system used just one of them: Deep NN (DNN) and Bidirectional Long Short-Term Memory (BLSTM) NN. For training DNN, first a GMM-HMM system was used. The GMM-HMM system was trained with the standard EM algorithm. The forced alignment computed with the GMM-HMM was used to create a training set for DNNs. The networks were pre-trained with a unsupervised layerwise training method [18], and fine-tuned with cross-entropy training and stochastic gradient descent. Different number of hidden layers were tested and the best networks were further trained with a sequence-discriminative criterion [19]. Both type of features previously mentioned were also used to train BLSTM-RNNs systems [9] with the CTC objective function [15] and with the dropout technique [20]. For language modelling, some normalisation was carried out (to isolate currency symbols and split sequences of digits and capital letters). Pyphen[14] was used for dealing with hyphened words. It was used to generate hyphenated words from the most frequent word in the vocabulary and the obtained parts of the words were used to complete full words that were added to the language model as unigrams. Recognition was carried out at line level with a 4-gram. According to the authors, the OOV in the validation set without adding hyphenated words was 7.1%[15] and after adding the full words the OOV was 5.6% with an increment of the lexicon from 7,318 to 32,692. Lattices were

[11]http://www.numen.fr/en/innovation-rd/project-improve-text-capture

[12]http://oll.libertyfund.org/titles/bentham-works-of-jeremy-bentham-11-vols

[13]The software modules behind that as well as the basic utility technologies are essentially powered by PLANET's ARGUS framework for intelligent text recognition and image processing.

[14]http://pyphen.org/

[15]Note that this value is computed from Table I as 857 running words divided by the difference between 12962 running words minus 857 running OOV, that is 857*100/(12962-857)

generated with all systems, and a lattice-based system combination was carried out.

- **LIMSI-Un:** This system was also a system combination. The difference between this system and the previous system was that additional training data was used in this system, but only systems with handcrafted features were trained. The additional training data for OM was the same as in the A2IA-Un system.
  The Open American National Corpus (OANC)[16] was used as external resource for training the LM. All words with single counts were removed before adding partial words obtained with the hyphenation technique previously described. Completed words were added only for words with count higher than 100. The recognition was carried out at line level with a bigram, and then a reescoring with the lattices was carried out with a trigram. The combination included 7 systems: the 2 DNNs features and pixels (trained on restricted data), and 5 RNNs (features, pixels, and three trained with unrestricted data), and used the unrestricted LM.

V. Results and Discussion

The best results obtained by each participant can be seen en Table III. A clear improvement is achieved when using more training data, as column *Unrestricted track* shows. In the restricted track both entrants obtained similar results. In the unrestricted track, A2IA clearly obtained the best result, even though the techniques used by both participants were quite similar. This such a significant improvement could be due to two important aspects: i) the optical models were trained and progressively improved with several datasets, including a large dataset of historical documents; ii) the language model included transcripts from additional Bentham manuscripts (different from those used in the test set).

Table III
Best Word Error Rate and Character Error Rate (WER/CER) obtained by the participants on each track.

	Restricted track	Unrestricted track
A2IA	-	8.6 / 2.9
CITlab	14.6 / 5.0	-
LIMSI	15.0 / 5.5	11.0 / 3.9

Figure 4 shows the histogram of the positions of the errors with to regard to the correct line transcripts. We used the lines with more than three words in the reference transcripts for computing this histogram. Note that most of the errors were concentrated in the initial and last parts of the lines for all results. Note that hyphenated words may have large influence in the errors. Note that the LIMSI technique for dealing with hyphenated words seemed to have a positive effect when dealing with these hyphenated words as top and bottom plots show. However a more deep research on this issue would be interesting.

[16]http://www.americannationalcorpus.org/OANC/index.html

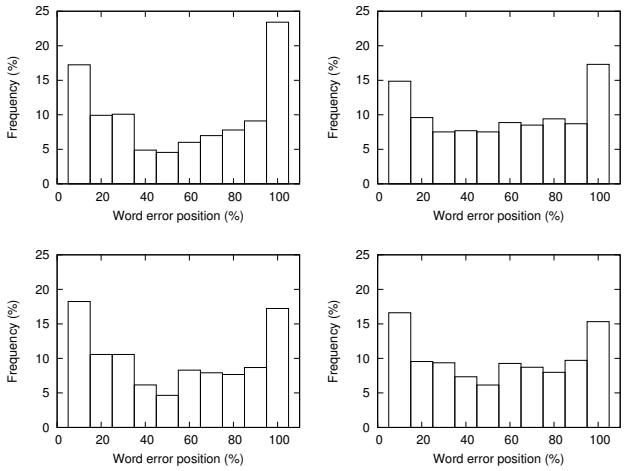

Figure 4. Histogram of the positions of the errors in the reference transcripts. The plots from top to down and from left to right correspond to: A2IA-Un, CITlab-Re, LIMSI-Re, LIMSI-Un

Figure 5 shows in boxes the histogram of the length of words in the test set, and with dashed lines the histogram of the length of words that were incorrectly transcribed. For the errors, only substitutions and deletions of the reference transcripts were considered. Most of the errors were concentrated in short words, specially in words with only one character. This is specially noticeable for A2IA-Un, LIMSI-Re, and LIMSI-Un systems.

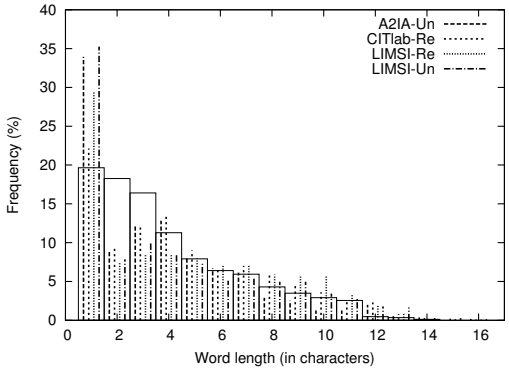

Figure 5. Normalised histograms of the length of words that took part in the error events.

To conclude this section, we can say that the winners of the contest were: CITlab-Re for the restricted track and A2IA-Un for the unrestricted track.

VI. Conclusion

This paper described the HTRtS contest that was organised in the context of the ICFHR 2014 conference with part of the English Bentham dataset that was prepared in

the TRANSCRIPTORIUM project. The three entrants obtained very good results with this dataset in the two tracks that were defined. For future work, we plan to carry out this contest with more data both for training and test that will include more difficulties like writing styles, crossed-out texts, fainted texts and larger vocabularies.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 600707 - tranScriptorium. The authors would like to thank all the TRANSCRIPTORIUM members for their collaboration and the entrants for their participation in this contest.

REFERENCES

[1] J. A. Sánchez, G. Mühlberger, B. Gatos, P. Schofield, K. Depuydt, R. Davis, E. Vidal, and J. de Does, "tranScriptorium: an European project on handwritten text recognition," in *DocEng*, 2013, pp. 227–228.

[2] U.-V. Marti and H. Bunke, "Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System," *IJPRAI*, vol. 15, no. 1, pp. 65–90, 2001.

[3] A. H. Toselli, A. Juan, D. Keysers, J. González, I. Salvador, H. Ney, E. Vidal, and F. Casacuberta, "Integrated Handwriting Recognition and Interpretation using Finite-State Models," *IJPRAI*, vol. 18, no. 4, pp. 519–539, June 2004.

[4] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A Novel Connectionist System for Unconstrained Handwriting Recognition," *IEEE Transaction on PAMI*, vol. 31, no. 5, pp. 855–868, 2009.

[5] T. Causer and V. Wallace, "Building a volunteer community: results and findings from Transcribe Bentham," *Digital Humanities Quarterly*, vol. 6, no. 2, 2012.

[6] B. Gatos, G. Louloudis, T. Causer, K. Grint, V. Romero, J. Sánchez, A. Toselli, and E. Vidal, "Ground-truth production in the tranScriptorium project," in *DAS*, Tours, France, April 2014, pp. 237–241.

[7] C. Clausner, S. Pletschacher, and A. Antonacopulos, "Aletheia - an advanced document laout and text ground-truthing system for production environments," in *Proc. of the International Conference on Document Analysis and Recognition*, Beijing, China, Sept. 2011, pp. 48–52.

[8] S. Pletschacher and A. Antonacopoulos, "The PAGE (page analysis and ground-truth elements) format framework," in *Proc. ICPR*, 2010, pp. 257–260.

[9] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. on PAMI*, vol. 31, no. 5, pp. 855–868, 2009.

[10] B. Moysset, T. Bluche, M. Knibbe, M. F. Benzeghiba, R. Messina, J. Louradour, and C. Kermorvant, "The A2iA multi-lingual text recognition system at the second Maurdor evaluation," in *Submitted to ICFHR*, 2014.

[11] E. Grosicki and H. E. Abed, "ICDAR 2009 handwriting recognition competition," in *ICDAR*, 2009, pp. 1398–1402.

[12] U. V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," *IJDAR*, vol. 1, no. 5, pp. 39–46, 2002.

[13] S. Brunessaux, P. Giroux, B. Grilheres, M. Manta, M. Bodin, K. Choukri, O. Galibert, and J. Kahn, "The Maurdor project - improving automatic processing of digital documents," in *Proc. DAS*, 2014, pp. 349–354.

[14] R. Messina and C. Kermorvant, "Over-generative finite state transducer n-gram for out-of-vocabulary word recognition," in *International Workshop on Document Analysis Systems (DAS)*, 2014, pp. 212–217.

[15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[16] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks." in *NIPS*, 2008, pp. 545–552.

[17] A. L. Bianne, F. Menasri, R. Al-Hajj, C. Mokbel, C. Kermorvant, and L. Likforman-Sulem, "Dynamic and contextual information in hmm modeling for handwriting recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 2066–2080, 2011.

[18] G. E. Hinton, S. Osindero, , and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–554, 2006.

[19] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Interspeech*, vol. 1, 2013, pp. 3–7.

[20] V. Pham, C. Kermorvant, and J. Louradour, "Dropout improves recurrent neural networks for handwriting recognition," 2013, [Online]. Available: http://arxiv.org/abs/1312.4569.