A Set of Benchmarks for Handwritten Text Recognition on Historical Documents

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

[a]*Pattern Recognition and Human Language Technologies Center*
Universitat Politècnica de València
Camino de Vera s/n, 46022 , València, Spain

Abstract

Handwritten Text Recognition is a important requirement in order to make visible the contents of the myriads of historical documents residing in public and private archives and libraries world wide. Automatic Handwritten Text Recognition (HTR) is a challenging problem that requires a careful combination of several advanced Pattern Recognition techniques, including but not limited to Image Processing, Document Image Analysis, Feature Extraction, Neural Network approaches and Language Modeling. The progress of this kind of systems is strongly bound by the availability of adequate benchmarking datasets, software tools and reproducible results achieved using the corresponding tools and datasets. Based on English and German historical documents proposed in recent open competitions at ICDAR and ICFHR conferences between 2014 and 2017, this paper introduces four HTR benchmarks in order of increasing complexity from several points of view. For each benchmark, a specific system is proposed which overcomes results published so far under comparable conditions. Therefore, this paper establishes new state of the art baseline systems and results which aim at becoming new challenges that would hopefully drive further improvement of HTR technologies. Both the datasets and the software tools used to implement the baseline systems are made freely accessible for research purposes.

Keywords: Historical Handwritten Text Recognition, Hidden Markov Models, Convolutional Neural Networks, Recurrent Neural Networks, Language Modeling.

*Tel: +34 96 387 7358, Fax: +34 96 387 7239. e-mail: jandreu@prhlt.upv.es

1. Introduction

Off-line Handwritten Text Recognition (HTR) is a fundamental requirement to unveil the substance of billions of historical manuscripts residing in archives and libraries. Many of these documents are digitized, but the access to their contents is very limited since they are just raw images. HTR systems aim at transcribing these documents in order to make their textual contents accessible and searchable.

HTR has progressed enormously in the last two decades due mainly to two reasons: first, the use of holistic training and recognition concepts and techniques which were previously developed in the field of Automatic Speech Recognition (ASR); and second, the existence of an increasing number of publicly available datasets for training and testing the HTR systems.

The need for holistic techniques in HTR has been known for many years given that the processes of handwriting and speech share many similar properties and challenges [1, 2, 3, 4, 5]: i) in both cases the production process is sequential through time; ii) the resulting images or signals are often largely distorted and severely contaminated with different kinds of noise; iii) due to the sequential production process, it is not possible in general to accurately recognize isolated words or characters/phonemes because none of these units can be reliably and consistently segmented or isolated; and iv) handwriting images and speech signals typically exhibit similar forms of lexical and syntactical regularity and ambiguity. Because of these similarities it is not surprising that the same basic Pattern Recognition techniques which had proved successful in ASR also become successful in HTR. To name a few: hidden Markov models (HMM) and recurrent neural networks (RNN) for optical character/phoneme modeling and statistical N-gram models for language modeling. These models are trained both in ASR and HTR with identical machine learning techniques based on the use of annotated data. The availability of sufficiently large amounts of annotated data is currently one of the bottlenecks to move forward in HTR since the annotation is generally performed by human experts and is, therefore, expensive and time-consuming.

Currently, several freely available datasets exist which are commonly used in HTR experimentation. We now mention just a few. One of the earliest and best known is IAM [6][1]. It is often considered "semi-artificial" in that it consists of short fragments of modern English printed (electronic) text,

[1] `www.fki.inf.unibe.ch/databases/iam-handwriting-database`

copied by hand (i.e., handwritten) by volunteers on clean, white paper. The best Word Error Rate (WER) for the standard evaluation (test) images of this dataset is 9.3% [7]. Another well-know dataset is RIMES[2], composed of handwritten letters written by more than 1 300 people. The best word error rate WER with this dataset is 11.2% [5]. One of the first historical handwritten datasets used for HTR was the so called George Washington (GW)[3], although most results reported with this dataset are related to Key Word Spotting. In a different language, the Esposalles dataset[4] [8] is a relatively small set of page images from a marriage register book written in old Catalan by a single hand, which belongs to a large XVII century collection. The best WER for the evaluation partition in this dataset is 10.1% [9]. On the other hand, the Rodrigo dataset[5] is the result of digitizing and annotating a full manuscript dated 1545. It is completely written in old Castilian (Spanish) by a single author. The best WER for the evaluation partition in this dataset is 14% [10]. A historical Arabic dataset is referred to in [11], but it is quite recent and no WER results are provided in the reference. There exists other handwritten text corpora for different languages like Chinese[6], although some of them are not historical documents.

In recent years several projects have been supporting research in HTR, both at national and at European level. Among several important projects which have contributed to the advance of HTR, it is worth to mention the series of HisDoc projects[7], funded by the Swiss National Science Foundation since 2009. In the present work, we focus on two specific European projects, tranScriptorium[8] and READ[9]. The involvement of archives and libraries in these projects has been crucial to allow learning first-hand what are the most important HTR challenges associated with the historical documents residing in these organizations. Many medium- and large-size manuscript collections have been processed in the framework of these projects and parts of some of them have been used in HTR competitions organized in the context

[2] www.a2ialab.com/doku.php?id=rimes_database:start

[3] www.fki.inf.unibe.ch/databases/iam-historical-document-database/washington-database

[4] http://dag.cvc.uab.es/the-esposalles-database

[5] www.prhlt.upv.es/wp/resources/the-rodrigo-corpus

[6] www.nlpr.ia.ac.cn/databases/handwriting/Home.html

[7] http://diuf.unifr.ch/main/hisdoc (visit also the links: HisDoc 2.0, HisDoc III)

[8] http://transcriptorium.eu/

[9] https://read.transkribus.eu/

of the ICDAR and ICFHR conferences in the last few years. The goals of these competitions were to bring together researchers to share and compare new techniques and ideas on HTR. These competitions have actually been successful in promoting a great, fast progress in HTR.

The main contributions of this paper are: i) to describe the datasets used in these HTR competitions; ii) to summarize the main challenges raised by them and the results obtained for the corresponding datasets; iii) to present the details of new up-to-date baseline systems and results for these datasets, based on the prevalent technologies that allow obtaining competitive results; and iv) to provide freely available tools, specific code and corresponding scripts which allow full reproductivity of the baseline results reported here.

The article is organized as follows: Section 2 summarizes the prevalent technology used nowadays for HTR. The baseline systems have been developed with this technology. Section 4 explains the relation among projects and the datasets used in the competitions. Then, Sections 5, 6, 7 and 8 describe each dataset and the corresponding baseline system and results. These systems are freely accessible through GitHub.[10]

2. HTR technologies

The most traditional approaches to HTR are based on N-gram language models (LM) and optical modeling of characters by means of HMMs with Gaussian mixture emission distributions (HMM-GMM) [3, 5]. However, significant improvements in optical modeling were demonstrated by approaching emission probabilities with multilayer perceptrons (HMM-MLP) [5] and also by training the HMM-GMMs with discriminative training techniques [12]. More recently, notable improvements in HTR accuracy have been achieved by using RNNs for optical modeling.

Optical models are trained with pairs of line images and their corresponding transcripts. If transcripts exactly follow the text written in the images, they are usually called "diplomatic". In the cases considered in this paper, diplomatic transcripts are assumed.

Language Models, on the other hand, are usually trained using only training text, typically just the transcripts of the training images.

[10] `https://github.com/PRHLT/htr-contests-exps`

2.1. Convolutional and Recurrent Neural Network Optical Modeling

Current state-of-the-art optical modeling HTR technologies are based on deeply layered neural network models which consist of a stack of several *convolutional* layers followed by one or more layers of RNNs composed of special "neurons" called *Bidirectional Long Short Term Memory* (BLSTM) units [4, 5]. Finally, a softmax output layer computes an estimate of the probabilities of each character in the training alphabet plus a special "non-character" symbol. The overall architecture is often referred to as *Convolutional-Recurrent Neural Networks* (CRNN) [13].

In [14], a more complex, "multidimensional" version of BLSTM architecture was introduced, leading to the so called *Multidimensional Recurrent Neural Networks* (MDLSTM), which became fairly popular for some years because of their superior performance. However, a more recent paper [13] has shown that, by adequately configuring the stack of convolutional and recurrent layers, similar HTR accuracy can be obtained using only plain BLSTM NNs, leading to simpler and much more efficient CRNN architectures.

The baseline results reported in this paper were obtained using this simpler kind of CRNNs for character optical modeling. A CRNN is trained by stochastic gradient descend with the RMSProp method [13] on mini-batches to minimize the so called *Connectionist Temporal Classification* (CTC) cost function [4]. Dropout techniques are used to reduce training over-fitting, which has been proved to effectively improve recognition accuracy [5]. In order to decide when to stop the training iterations, a *development set* (which may be an excerpt of the training set) is used. Therefore, to take further advantage of the labeled data contained in this set, the standard training process is generally finalized by running a few more training iterations with all the training data available (including the development set). It is worth noting that all the techniques, associated and software tools required for this approach are now implemented and readily available in the HTR *Laia Toolkit* [13], based on the Torch machine learning platform.

The exact architecture of a typical CRNN is defined through a large number of hyper-parameters. These include, at least, the size of the input line image, the number of convolutional layers, the number of filters in each layer, the kernel sizes and resolution-reduction factors ("max-pooling") of these filters, the number of recurrent layers, the type of recurrent units (BLSTM or other), the number of these units in each recurrent layer, the types of activation functions used in each layer, and the number of different characters to be predicted in the output layer. In addition, other hyper-parameters are

needed to specify the CRNN training details including, at least, the size of the mini-batches, the the rate of dropout, the base learning rate, the convergence criterion (possibly using a development set) and the number of possible extra iterations using the development data.

Clearly, optimizing such a large number of hyper-parameters for each HTR task is an important bottleneck of CRNN character optical modeling. Aiming to overcome this bottleneck, the *Laia Toolkit* offers default architecture definitions and corresponding training hyper-parameters which generally allow in practice to obtain accurate CRNN optical models for typical handwritten documents in most languages, historical periods and writing styles. Of course, it is up to the practitioner to adapt the basic, default settings to the singularities of each document collection considered. Small tweaks of the basic settings do not generally lead to large performance differences. But in the current state of the affairs, experience and intuition often allow to achieve some improvements by tuning some settings taking into account dataset features such as image resolution, writing density and average text size, overall image quality and amount of available training data, among others.

To obtain the baseline results reported in this paper, only small variations to the default *Laia* settings were made. In general, unless otherwise stated, CRNN optical models include a stack of four convolutional layers and three recurrent layers, each with 256 BLSTM units. Full architecture and training details are included in the scripts available for each dataset in GitHub.[10]

As previously discussed, for a given text line image, a trained CRNN estimates a sequence of character posterior probability vectors (often referred to as "ConfMat"). While raw images can be directly accepted as input, results can often be improved if images are previously deskewed, deslanted, cleaned, contrast-enhanced, and/or size-normalized [15, 16, 17, 18].

2.2. Language modeling

CRNNs have proved able to capture by themselves lexical and linguistic context to some extent. However, classical LM methods, which explicitly aim at modeling contextual regularities and constraints, can often help CRNNs to further improve HTR results. Moreover, these models can be very easily and efficiently trained using only plain text (i.e., only transcripts, without images). In this paper, statistical character N-grams, estimated using the

SRILM Toolkit[11], are used, with a default N-gram order of $N = 8$ and Kneser-Ney back-off smoothing [19].

Trained N-gram contextual constraints can be applied to the CRNN output character probabilities in several ways [5]. Here N-grams are represented as a stochastic finite-state transducer. The edge probabilities of this transducer are then obtained by adequately combining the estimated N-gram probabilities with CRNN output character posteriors, suitably scaled with character priors [5]. The resulting stochastic transducer, along with the classical Viterbi decoding algorithm (also known as "token- or message-passing"), are used to obtain an optimal transcription hypothesis of the original input line image. For these combination and decoding processes, the KALDI toolkit [20] is used in this paper.

3. Evaluation measures

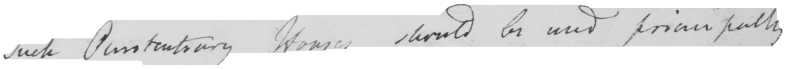
The most usual evaluation metrics for measuring the performance of an HTR system are the Word Error Rate (WER) and the Character Error Rate (CER). WER is defined as the minimum number of words that need to be substituted, deleted, or inserted to match the recognition output with the corresponding reference ground truth, divided by the total number of words in the reference transcripts. CER is defined in the same way but at character level. See examples in Figure 1.

Generally speaking, WER is fairly well correlated with CER, but this correlation is not always strong or systematic. Therefore, *both* measures are important and complementary to assess the quality of an automatic transcript. A low CER but a relatively high WER reveals that the character errors are spread among many words. Conversely, a transcript with the same CER as before but lower WER indicates that errors are concentrated in few words. A good language model typically helps to achieve greater improvements in WER than in CER. It is worth mentioning that punctuation symbols, like commas, are separated from the preceding words. This becomes relevant if punctuation is frequent and/or when WER is small.

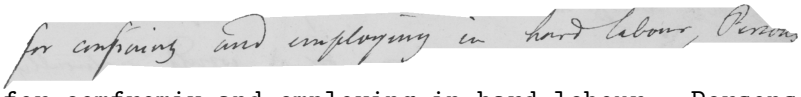
WER tends to be better than CER at indicating how difficult is to understand a transcript by human beings. Similarly, even if CER is low, a high WER may dramatically harm the performance of information extraction or

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

searching systems which rely on automatic transcripts. Figure 1 illustrates these facts for two samples from the ICFHR-2014 benchmark which exhibit similar CER but different WER.



such Penstentrary Hoases should be anid priapalty — WER = 4/7 = 57%, CER = 8/50 = 16%



for eomfromiy and employing in hard lebour , Persons — WER = 2/9 = 22%, CER = 8/52 = 15%

Figure 1: Two examples of test line images and *automatic* transcripts, along with the corresponding WER and CER. While CER is similar in both transcripts, that with higher WER may be harder to understand. *Reference* transcripts for the top and bottom line images are, respectively: "`such Penitentiary Houses should be and principally`" and "`for confining and employing in hard labour , Persons`".

4. Relations among projects, datasets and competitions

Four HTR competitions have been organized and supported by the TRANSCRIPTORIUM and READ European projects. Organizing these competitions and making the datasets freely available for research purposes was one of the goals of these projects. The datasets and the challenges in each competition were defined according to the needs and the evolution of the projects. The authors of this paper did not participate in these competitions.

The first two of these competitions were based on parts of the so called Bentham Papers, handwritten in English by several writers. The whole digitized collection encompasses 100 000 page images of text authored by the renowned English philosopher and reformer Jeremy Bentham (1748-1832)[12] [21]. It mainly contains legal forms and drafts in English, but also some pages are in French and Latin. Documents handwritten by Bentham himself over a period of sixty years, as well as fair copies handwritten by Bentham's secretarial staff, are included. Many images entail important pre-processing and layout analysis difficulties, like marginal notes, faint ink, stamps, skewed images, lines with large slope variation within the same page, slanted script, inter-line text, etc.

[12] `http://www.ucl.ac.uk/Bentham-Project/`

The Bentham Papers were being transcribed following a crowdsourcing initiative [21]. One of the main objectives of the tranScriptorium project was to develop interactive-predictive transcription technology [22] and to adapt it to help the crowdsourcers transcribe more easily and efficiently.

The first HTR competition was organized in 2014 within the "International Conference on Frontiers of Handwriting Recognition" (ICFHR-2014) [23]. The corresponding dataset was a small part of the Bentham Papers containing only relatively "easy" images, manually selected to avoid the most severe difficulties entailed by this collection.

Then a second competition was organized in 2015 within the "International Conference on Document Analysis and Recognition" (ICDAR-2015) [24]. The dataset was again a part of the Bentham Papers, but it was significantly larger than the previous one and contained much more "difficult" images, including plenty of crossed-out text, marginalia, added interline text, and more difficult writing styles. In addition to these and other new challenges (see details in [24]), the following motivation was perhaps most interesting: A usual HTR scenario is that some (or many) transcripts are available for some parts of the collection, but they are not aligned with their corresponding line images. These transcripts might be profitably used to better train both the optical and the language models but, for optical model training, it is necessary to pair line images with their corresponding transcripts. In this dataset, the transcripts of a subset of training data were provided at page level, and the participants had to use their own techniques for detecting the lines and for aligning the detected lines with the corresponding text of the page transcripts. This challenge was in part motivated by the need of non-expensive ways to prepare training ground truth (GT).

The third competition was organized within ICFHR-2016 in the framework of the READ project [25]. This time the Ratsprotokolle collection, composed of handwritten minutes of council meetings held from 1470 to 1805, was considered. It was considered in READ as a good example of archive manuscripts written in old German. Thus, a main challenge stated in this competition was to deal with a language different from English. The German language is similar in some aspects to English, specially from the optical modeling point of view. But compound words make word-level language modeling more challenging. One important characteristic of this collection is that the lines are short, each one containing very few (long) words, which makes it difficult to take much advantage of using a word LM.

The fourth HTR competition was organized within ICDAR-2017 [26],

again in the framework of the READ project. In this case, the dataset was a part of the Alfred Escher Letter Collection (AEC)[13] which is composed of letters handwritten mainly in German but it also has pages in French and Italian. This collection includes many images whose transcripts are only provided at the page level; i.e., without any alignment of transcribed text with image lines. This competition went thus further in the direction initiated in ICFHR-2014. That is, the problem was to automatically detect the lines and align them with corresponding text from the training page transcripts and then use the transcribed line images for training an HTR system. In this case, it was feasible to proceed in the traditional way for a few pages; i.e., accurately detected and (manually supervised) lines and the corresponding exact transcripts were provided to be used for training an initial HTR system. This seed system could then be used to automatically align page transcripts with detected lines, thereby increasing the amount of training material.

The details of the datasets used in these contests and in the benchmarks of this paper are described in the following sections. The actual datasets are available for research purpose at ZENODO: Table 1 provides the specific web addresses from which these datasets can be downloaded. The GT is provided in PAGE format [27].

Table 1: The datasets described in this paper are publicly available for research purposes at the following web URLs. All of them are in PAGE format [27].

Dataset	Internet address
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`

For each dataset, a script based on the previously mentioned Laia toolkit is provided. These scripts, available at GITHUB[14], allow to reproduce the HTR baseline experiments reported in the following sections. Each script downloads the corresponding dataset, executes the training and recognition processes and provides the WER and CER results reported in this paper.

[13] `https://www.briefedition.alfred-escher.ch/`
[14] `https://github.com/PRHLT/htr-contests-exps`

5. The ICFHR-2014 benchmark

5.1. Dataset description

The dataset for this benchmark was taken from the Bentham Papers. It is a small sample of so called *"easy pages"* of this collection. Figure 2 shows some examples of the images included in this dataset.

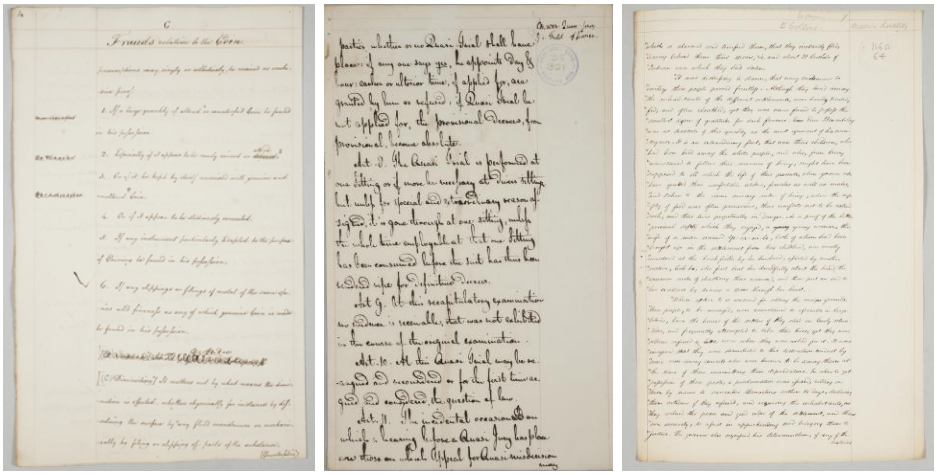


Figure 2: Document samples of the ICFHR-2014 dataset.

Ground truth (GT) line detection and transcription were produced semi-automatically and registered in PAGE format [27]. Table 2 summarizes the basic statistics of this dataset. Most of its 433 page images encompass a single text block each. In total, they contain 11 473 lines with nearly 107 000 running words and a vocabulary of more than 9 700 different words. The dataset was divided into three subsets for training, development and test, respectively encompassing 350, 50 and 33 images. The rows "Running words" and "Lexicon" show total number of words and number of *different* words, respectively. Development Out-Of-Vocabulary (OOV) refers to words that do not appear in the training set, while Test OOV are words not appearing in the training or the development sets. The row "OOV Lexicon" shows numbers of *different* OOV words. The images and the transcripts of the training and development data are provided in the dataset both at page level and at line level. The images of the test data are provided at line level.

In the ICFHR 2014 competition, the so called *restricted track* was established in order to allow fair comparison of HTR techniques under identical training data conditions. In this track, participants were allowed to use just the data provided by the organizers for training and tuning their systems. In

Table 2: Main statistics of the ICFHR-2014 dataset. The images were scanned at 300 dots per inch (dpi). The exact number of writers is unknown, but it is believed to be about 20 writers.

Number of:	Training	Development	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
Running OOV (%)	-	6.6	5.3	-
Lexicon	8 658	2 709	1 946	9 716
Lexicon OOV	-	681	377	-
Character set size	86	86	86	86
Running characters	442 336	67 400	40 938	550 674

addition an *unrestricted track* was also considered, where participants were free to train their systems with any (amount of) data of their choice.

5.2. Summary of results obtained with the ICFHR-2014 dataset

This dataset has been used in several HTR research works. Table 3 summarizes the most relevant results obtained in the *Restricted track*.

Table 3: Results obtained with the test set of the ICFHR-2014 dataset in the *Restricted track*. CER & WER are, respectively, character and word error rate percentages. The row marked with "†" corresponds to the winner of the competition.

References	Approach	CER(%)	WER(%)
[23]†	CRNN + regex/lexicon LM [28]	**5.0**	14.6
[5]	CRNN + word 2-gram	**5.0**	14.1
[12]	Discriminative HMMs + word 2-gram	6.7	17.2
This paper	CRNN (Laia)	6.2	12.7
	+ character 7-gram	**5.0**	**9.7**

The first row corresponds to the winner of the ICFHR 2014 competition [23]. In that work, usual pre-processing techniques, as described in Section 2, were used and the line images were scaled to a height of 96 pixels. State-of-the-art (as of 2014) CRNN technology, was used for optical modeling. The same architecture defined in [14] was adopted, but rather than a MDLSTM architecture MDLeaky cells were used because, according to this participant, these cells were more stable. In addition, a lexicon (roughly equivalent to a 1-gram LM) derived from the training transcripts, was used to finally obtain excellent results.

A similar CRNN approach, with some variations in the architecture, was used in [5]. In this case, however, rather than a plain lexicon, a standard word 2-gram LM was used. This PhD work contains several interesting comparisons, such as using the raw line image pixels as input, versus extracting handcrafted line image features; or relative improvements achieved different variants of MDLSTM architectures, etc. Moreover, combinations of some of these alternatives, by means of word lattices obtained as byproducts of Viterbi decoding, were also studied. The WER reported in the second row of Table 3 is the best system-combination result achieved in [5].

Finally, the third row of Table 3 shows results achieved in a work that did *not* use CRNN optical modeling; instead, traditional HMM modeling was enhanced by means of (also traditional) HMM discriminative training techniques [12]. The reported WER was obtained through conventional Viterbi decoding using a word 2-gram LM.

The baseline system proposed in the present work for the ICFHR-2014 dataset is based on the CRNN technology described in Section 2. Preprocessing applied to this dataset includes line image scaling, contrast enhancement and noise removal, as in [29]. Both the training data and the test data are pre-processed in this way.

The CRNN architecture and training process for optical modeling were as discussed in Sec. 2.1. Details can be seen in the scripts provided for this specific dataset in GitHub.[10] Given the relatively scarce text data available in the training transcripts, a character 7-gram was used in this case.

The most relevant differences between our proposed baseline and the systems described in [23] and [5] is that we adopt plain a BLSTM (rather than MDLSTM) architecture, with the improved architecture discussed in [13], and use a relatively high-order *character* N-gram, rather than a simple lexicon, or the more traditional *word* 2-gram LM.

5.3. Analysis of results and summary of pending challenges

A main conclusion in this first benchmark is that using CRNN for optical modeling and character N-grams for LM is currently the best approach for this dataset. In a more detailed analysis of the results, we observed that although the CER was fairly low for all the systems based on RNN optical modeling, there are important differences in WER. For the benchmark system here proposed, it is important to remark that including a character N-gram for LM resulted in a decrease of CER from 6.2% to 5.0% (*17% relative*) and WER from 12.7% to 9.7% (*24% relative*). A comparison of these results with

the others shown in Table 3 suggests that our WER improvement is due to the use of a character N-gram for LM.

Looking at the results in further detail reveals that, in contrast with the most traditional 2-gram word LMs, a long-span character LM permits a simpler and more effective treatment of OOV words in the decoding phase. Using a character LM leads to output text where characters tend to be concatenated according to the regularities observed in the training transcripts. This also applies to the blank-space and the punctuation signs, which act as "word" separators. However, in contrast with using a *word LM* (which ensures that output words do belong to a fixed lexicon), the "words" obtained using a *character LM* are by no means ensured to be real English words.

To better understand this issue, we analyzed the amount of output "words" that were not really English words. First we applied an English speller to the test reference transcripts and observed that 1.5% of the running words were not English. Most of them were actually parts of hyphenated words and a few were GT mistakes. Some examples are: *ar, gued, effec, tive, compleat, enquiries, tioned*, etc. Then, we did the same for the automatic transcripts provided by our baseline systems. Without any LM, the corresponding value was 5.9%, but using the character 7-gram for LM, this value decreased to 3.0%. So, we conclude that the character LM not only decreases the WER, but also definitely helps making the transcripts more readable.

In order to shed further light into possible causes of errors, we studied how the WER depends on two features of the words involved in the errors. The left plot in Figure 3 shows the WER for different relative positions of the words along the recognized text lines. Most of the errors stockpile at both ends of the lines. A likely cause of this effect is the lesser linguistic context available in these line positions. This happens both with the raw CRNN output and also when a LM is used, but the gain of using a LM is generally greater towards the middle of the lines. Clearly, the capability of the LM to help avoiding errors is greater in the middle of the lines, where more context is available.

The right plot in Figure 3 shows how the WER depends on the length of the considered words. For each word length, the box shows the percentage of words of this length and the vertical lines represent percentages of errors in words of the corresponding lengths. It can be observed that about 16% of words consist of single-character words (11.6% are punctuation marks and 4.4% are normal words. About 23% of the 12.7% erroneous words (see Table 3) are produced in these single-character words. Results are similar

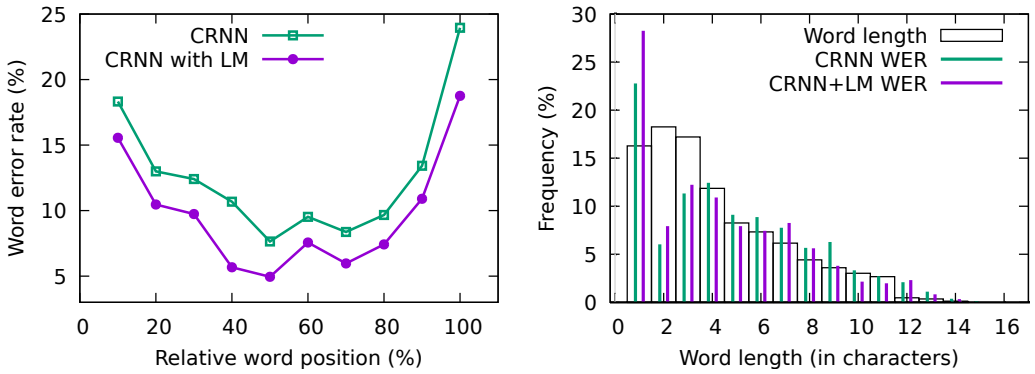


Figure 3: ICFHR-2014 benchmark. Left: WER for different relative positions of the words within the line. Right: normalized word length histogram, along with the the percentage of word errors for each specific length. Results with and without using a LM are shown.

using a LM, with error frequency picking at 28% for single-character words. The WER on punctuation symbols is 2.9% without LM and 2.6% with LM. Much in the same way as the context is more or less helpful depending on the word position, it is clear here that the error-avoiding capability of the (character) LM is lesser for single-character words or symbols, and becomes greater for longer words.

These results suggest possible ways to reduce the WER; namely: i) accuracy at line boundaries might be improved by concatenating the line images (per page or paragraph) and training the LM at page or paragraph level. Note that this will require a special treatment of hyphenated words; and ii) develop specific techniques to improve the recognition of punctuation marks.

A general lesson that was learned from the organization of the ICFHR-2014 competition was that preparing the GT for training a good HTR system is very time-consuming and expensive. This suggested that future challenges should involve a semi-automatic or fully automatic use of existing transcripts, with minimal supervision. Another important challenge suggested for upcoming competitions was to include more difficult page images, including crossed-out words, marginalia, faint text and bleed-trough, as well as promoting the use additional text to help training better LMs.

6. The ICDAR-2015 benchmark

6.1. Dataset description

The ICDAR-2015 dataset contains more difficult pages from a layout analysis point of view, drawn again from the Bentham collection. The im-

ages have marginal notes, fainted writing, stamps, skewed images, lines with different slope in the same page, variable slanted writing, inter-line text, etc. Figure 4 shows some examples of these images.

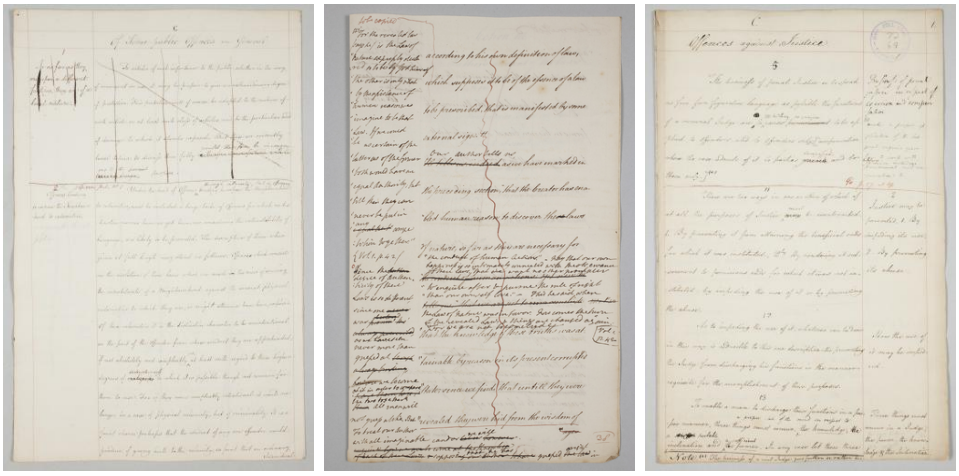


Figure 4: Document samples of the ICDAR-2015 dataset.

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. The images were scanned at 300 dpi. The exact number of writers is unknown, but it is believed to be about 20 writers.

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
Running OOV (%)	-	-	-	12.4	-
Lexicon	9 716	11 152	16 881	2 493	17 948
Lexicon OOV	-	-	-	1 067	-
Character set size	86	87	87	84	87
Running characters	550 674	357 672	908 346	47 286	955 632

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

6.2. Summary of results obtained with the ICDAR-2015 dataset

Table 5 shows a some of the most relevant results obtained in the *Restricted track*. The third row corresponds to the winner team of the ICDAR 2015 competition [24], which was also the winner of the previous HTR competition (ICFHR 2014) and used the same system described in the previous section. They did not provide any information about the way of detecting the lines in Train-B2. The first and the second rows of Table 5 are taken from [30], which provides further results and details which allow better comparison with our benchmark results, reported in the last two rows.

Table 5: Test set CER & WER for the ICDAR-2015 dataset in the *restricted track*. The row marked with "†" corresponds to the winner of the competition.

References	Approach	CER(%)	WER(%)
[30]	CRNN B1&B2	20.0	51.2
[30]	+ regexp/lexicon	15.1	33.8
[30, 24]†	+ ROVER combination	15.5	30.2
This paper	CRNN B1&B2 (Laia)	15.2	39.3
	+ char 8-gram	**12.8**	**30.0**

The image pre-processing and baseline system used for this dataset are similar to those of Section 5.2. The CRNN general architecture and training process for optical modeling were as discussed in Sec. 2.1, with details available in the scripts provided for this specific dataset in GitHub.[10] Given the larger amount of text data available in this dataset, the default N-gram order (8) was used for the character LM in this case.

Additional training data for optical modeling were obtained from Train-B2 as follows: First text lines of Train-B2 images were automatically detected using the well known open source system tesseract[15]. Also, a first CRNN model was trained on Train-B1 only, and used to recognize all the detected lines. Then, each of these automatic line transcripts was aligned with its best-matching GT line transcript from the same page. Finally, 4328 pairs of image-lines and GT transcripts with sufficiently high matching score were selected as additional training data.

[15] `https://hub.docker.com/r/mauvilsa/tesseract-recognize/`

6.3. Analysis of results and summary of pending challenges

All the results on this dataset evince that it is notably more difficult than that used in the ICFHR-2014 benchmark. This is what was expected, given the much more complex and noisy (and realistic!) images considered. For the proposed benchmark systems, the present CER is more than 2.5 times higher (from 5.0% to 12.8%) than that achieved with the ICFHR-2014 dataset, and the WER is more than three times worse (from 9.0% to 30.0%). The comparison is similar if we consider the results of the winner systems of the ICFHR-2014 and ICDAR-2015 HTR competitions.

The best results for this dataset are also achieved by the proposed baseline system, in this case both in terms of CER and WER. The impact of using a LM is also significant here, with a relative improvement of 24% of WER. However, in this case, the difference of using a long-span character N-gram, with respect to the simple lexicon used by the ICDAR-2015 winner system, is only minor. This is in sharp contrast with the important difference observed for the ICFHR-2014 dataset and it might be explained by the ROVER combination of results from different systems used in [30] (and [24]) which was not used in [28] (and [23]).

As in the previous experiments, Fig. 5 shows how the errors are distributed with respect to word positions and lengths. While the tendencies observed in Fig. 5 are similar to those observed for the ICFHR-2014 dataset in Sec. 5.3, here the WER variations with word position are somewhat less pronounced. It can be observed that about 18.0% of words consist of just one character, where 11.0% are punctuation symbols and 7.0% are normal words. The WER on punctuation symbols is 5.1%, and goes slightly down to 5.0% with LM.

To finish this section we suggest that in order to improve the word (and also the character) recognition accuracy in this benchmark, the same approaches hinted at the end of Sec. 5.3 apply.

7. The ICFHR-2016 benchmark

7.1. Dataset description

The ICFHR-2016 benchmark was based on a small part of the German Ratsprotokolle collection. The dataset for this benchmark is 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 pages

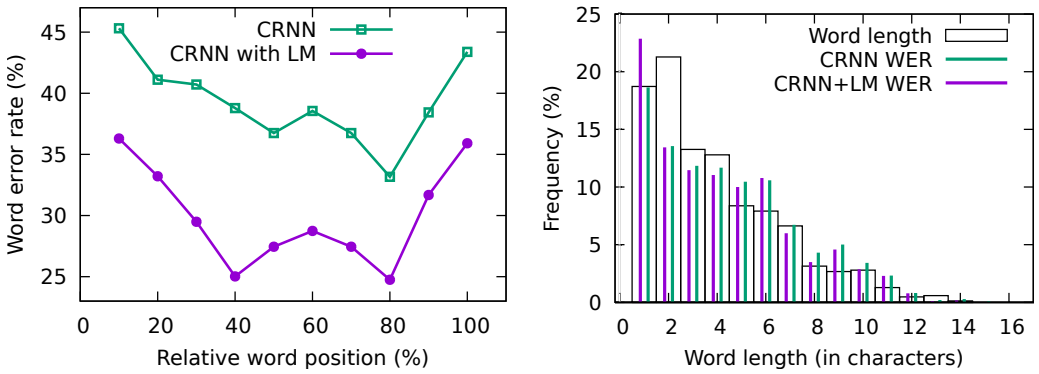


Figure 5: ICDAR-2015 benchmark. Left: WER for different relative positions of the words within the line. Right: normalized word length histogram, along with the percentage of word errors for each specific length. Results with and without using a LM are shown.

entail several line detection and transcription difficulties, including significant amounts of bleed-through. The corresponding GT was produced semi-automatically, with final manual revision. Figure 6 shows some images of this dataset.

These 450 pages contain 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 overall statistics of these images.

The dataset was divided into three subsets for training, development and testing, respectively encompassing 350, 50 and 50 page images. The GT in both training and development sets is in PAGE format and it is fully annotated at line level. On the other hand, the PAGE files of the test set only contain the line regions, but not the transcripts. Table 6 shows the details of these partitions.

The same tracks defined for the ICFHR 2014 and ICDAR 2015 competitions were defined with this dataset: a *Restricted track* and an *Unrestricted track*, the former allowing to use just the material provided by the organizers.

7.2. Summary of results obtained with the ICFHR-2016 dataset

Table 7 shows some of the most relevant results obtained with this dataset in the *Restricted track*. The first row corresponds to the winner of the HTR ICFHR 2016 competition. This participant applied usual pre-processing techniques to line images, which were fed to a neural network of five layers of CNNs followed by MDLSTMs. A character 10-gram was used for LM in the final decoding phase [25].

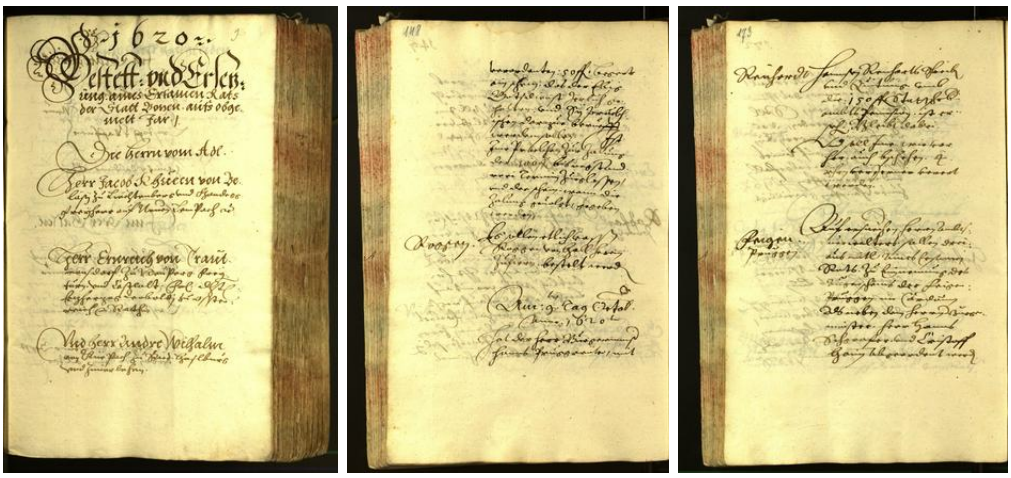


Figure 6: Document samples of the ICFHR-2016 dataset.

Table 6: Main statistics of the ICFHR-2016 dataset. Image resolution is 300 dpi. The pages were written by several writers but the exact number of writers is unknown.

Number of:	Train	Development	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
Running OOV (%)	-	16.8	14.7	-
Lexicon	6 985	1 526	1 656	8 120
Lexicon OOV	-	574	563	-
Character set size	92	80	83	92
Running characters	208 595	26 654	25 179	260 428

The baseline system presented here is also based on the general setup described in Section 2, including the character 8-gram LM. The CRNN architecture and training details for optical modeling are available in the scripts provided for this specific dataset in GitHub.[10]

7.3. Analysis of results and summary of pending challenges

The proposed baseline system achieves the best CER and WER results also in this case, with a WER relative improvement of 16.3% with respect to the winner of the ICFHR-2016 competition.

As in the previously considered datasets, here we computed how the errors are distributed with respect to word positions and lengths. The results are shown in Figure 7. The right plot reveals some of the specifies of the Ger-

Table 7: CER & WER obtained with the ICFHR-2016 dataset in the *Restricted track*. The row marked with "†" corresponds to the winner of the competition.

References	Approach	CER(%)	WER(%)
[25]†	CRNN + char 10-gram	4.8	20.9
This paper	CRNN (Laia)	4.8	19.0
	+ char 8-gram	**4.5**	**17.5**

man language mentioned above: words are generally longer and very short words are now very scarce. In this case, punctuation symbols in the *Test* represent 15.4% of the single-character symbols. However it is important to remark that the number single-character words is really small as it is shown in Figure 7. However, error rates for single-character words are similar or even higher than in the previous benchmarks. In this case the WER on punctuation symbols is 1.8% without LM and decreases to 1.6% with LM. Issues related with the language might also explain the behavior observed in the left plot. Here, in contrast with previous benchmarks, the WER without a LM is almost invariable with the position of the words in the line. When a LM is used, the overall WER improves 8% relative, but the WER for words towards the right end of the line does increase very significantly, as in the ICFHR-2014 benchmark. Rather surprisingly, though, now this increase does not happen for words at the left end of the line. Further studies are thus needed to try to better understand this unexpected behavior.

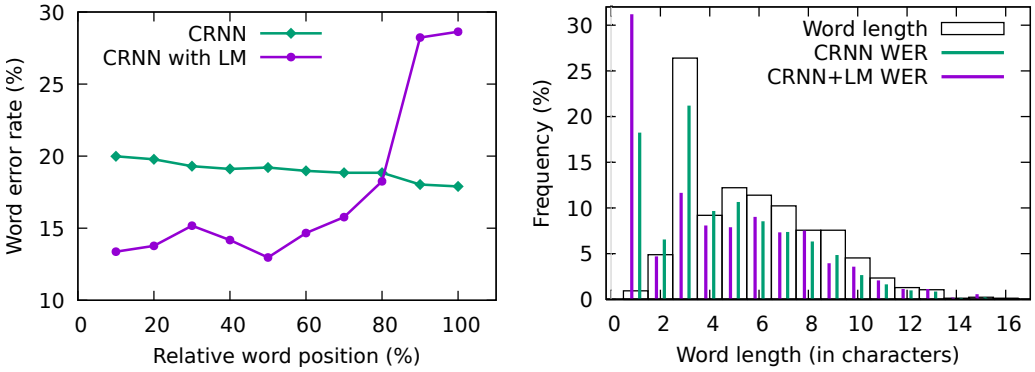


Figure 7: ICFHR-2016 benchmark. Left: WER for different relative positions of the words within the line. Right: normalized word length histogram, along with the percentage of word errors for each specific length. Results with and without using a LM are shown.

According to Table 6, the running OOV rate of this dataset is significantly higher than for the other datasets considered in this work. Longer German

words may contribute to this fact. While using a character N-gram for LM quite nicely deals with OOV words, it may also allow producing output character strings that are not really German words. To shed light into this issue, we used a German speller to evaluate the number of words that were really German words. First, the percentage of non-German words in the GT was 17.5%. We think that this is mainly due to the abbreviations and hyphenated words. The percentage of non-German words when the line images were recognized without a LM was 20.3%, while when a LM was used the percentage of non-German words was 18.3%, which is close to the GT rate.

As in the previous benchmarks, we think that recognition accuracy may be improved by concatenating line images and training a LM at page level without line breaks. In the ICFHR-2016 benchmark, this strategy might be even more rewarding, given the much shorter length of the lines involved.

8. The ICDAR-2017 benchmark

8.1. Dataset description

Most of the images used in the ICDAR-2017 benchmark were taken from the AEC 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. The text considered in this benchmark has been written by several hands, but the precise number of writers is unknown. Overall, the writing styles are quite heterogeneous in this dataset. Fig. 8 shows examples of these images. 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). Table 8 provides basic statistics of this dataset and partitions.

Train-A consists of 50 page images, each including one or more text blocks, making a total of about 1 000 lines. These pages entail several line detection and transcription difficulties and the corresponding, fully detailed, 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 improve line detection

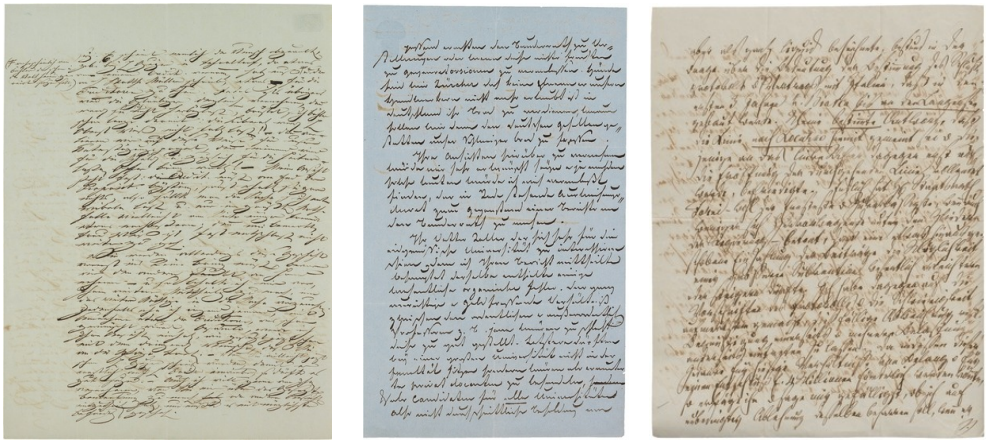


Figure 8: Document samples. The image on the left belongs to the Alfred Escher Letter Collection and the two images on the right belong to different collections.

Table 8: Main statistics of the ICDAR-2017 dataset. The resolution of the images is variable and ranges from 75 dpi to 300 dpi. The precise number of writers is unknown.

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
Running OOV (%)	-	-	-	5.5	6.0
Lexicon	4 637	98 993	99 530	4 635	4 648
OOV Lexicon	-	-	-	739	771
Character set size	102	168	168	104	104
Running characters	70 268	8 290 607	8 360 875	81 626	80 568

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 challenges are defined for this benchmark:

- *Traditional challenge*: The images from Test-A, annotated with baselines, are provided for usual transcription and evaluation.
- *Advanced challenge*: The Test-B2 images are provided without GT baselines. Text lines must be detected and then transcribed and submitted to evaluation. Clearly, this setting is more realistic and difficult.

The evaluation metrics used in the *Traditional challenge* are the WER and the CER, as usual. For the *Advanced challenge*, the *Bilingual Evaluation Understudy* (BLEU) metric was additionally proposed. However, a clear correlation of BLEU with WER and CER has been observed in this dataset [26] and, for homogeneity with previous sections, we will report only WER and CER also for the *Advanced challenge* in this paper.

In the *Advanced challenge*, the lines are not provided and therefore an automatic detection method has to be used in advance. The regions where the lines are located are provided. If an automatic line detection/extraction method is used, then some lines can be lost, and also their transcripts. Therefore, all lines detected and transcribed are concatenated for each region and the WER is computed with this concatenated string.

8.2. Summary of results obtained with the ICDAR-2017 dataset

Table 9 shows a summary of the most relevant results obtained with this dataset in ICDAR 2017 in the *Traditional challenge*. The first row corresponds to the winner of the ICDAR 2017 competition [26]. This participant used optical modeling consisting of 7 CNN layers and 2 layers of BLSTM, along with a character 10-gram for decoding. An initial optical model was trained using the provided segmented line images of Train-A. Then, lines of Train-B were automatically detected and recognized using the models trained with Train-A. Finally line images and GT transcripts were aligned using edit-distance methods. The optical model was re-trained with Train-A and these automatically aligned lines. This strategy was iterated a few times in order to improve the alignments and increase the total amount of training lines obtained. See all the available details in [26]. The second row corresponds to a system that used 13 CNN layers and 3 layers of BLSTM, along with a word 2-gram for decoding [31].

Table 9: CER and WER obtained for Test-A in the *Traditional challenge* of ICDAR-2017. The row marked with "†" corresponds to the winner of the competition.

References	Approach	CER(%)	WER(%)
[26]†	CRNN + char 10-gram	7.0	19.1
[31]	CRNN + word 2-gram	7.7	21.6
This paper	CRNN	6.7	21.6
	+ char 8-gram	**5.8**	**17.6**

The results of the here proposed baseline system are also shown in Table 9. To take full advantage of both training sets available, a procedure similar to that adopted by the winner of this competition was followed. To this end, a basic HTR system was trained with Train-A using the provided segmented line images and the default CRNN and language modeling settings described in Section 2.

Then, lines of Train-B were automatically detected using the approach presented in [32] and then recognized using the models trained with Train-A. Given that both line detection and recognition were error-prone, a dynamic programming alignment between the GT line transcripts and the hypothesized transcripts was carried out [33]. Pairs of line-images and transcripts aligned with high confidence were used to train a new optical model on both Train-A and Train-B. About 114 300 additional lines out of 200 000 lines in the 10 000 pages were used to train this new optical model. Unlike the winner of the ICDAR 2017 competition, in our case, this process was not iterated.

The writing style and other variabilities exhibited by the images in this collection are much larger than in the previous benchmarks considered in this paper. However, in this case a much larger amount of data is also available to afford the training of larger CRNNs which hopefully can to cope with these variabilities. Therefore, in this case a significant departure form the default CRNN settings discussed in Section 2.1 was adopted. Specifically, the number of convolutional and recurrent layers was increased from 4 to to 5, and from 3 to 4, respectively. In addition, the number of BLSTM units in each recurrent layer was increased from 256 to 512. As in the previous benchmarks, full architecture and training details for optical modeling are available in the scripts provided for this specific dataset in GitHub.[10] Finally, a standard 8-gram character LM was trained using the transcripts of both Train-A and Train-B.

Regarding the *Advanced challenge*, we performed experiments using the same models trained with Train-A and Train-B. The text lines of Test-B2 were detected using the advanced line detection tool of the Transkribus text image processing platform,[16] which proved more robust than the one we used in our training process. The quality of the detected lines has been assessed using the algorithm proposed in [35], obtaining a F-mesure of 97.0%. Table 10 shows the results obtained, along with those achieved by the win-

[16] Openly available at `https://transkribus.eu/Transkribus` – see details in [34].

ner of this challenge[17] in ICDAR 2017 (also the winner of the *Traditional challenge*).

Table 10: Comparative results obtained in the *Advanced challenge* in the ICDAR-2017 benchmark with Test-B2. All values were computed at region level by concatenating the line transcripts. The row marked with "†" corresponds to the winner of the competition.

References	Approach	CER(%)	WER(%)
[26]†	CRNN + char 10-gram	**6.4**	**16.8**
This paper	CRNN, char 8-gram, automatic line detection	7.0	20.0
	Same, but error-free line detection	6.3	18.5

8.3. Analysis of results and summary of pending challenges

As we commented in the previous section, German has some singularities that complicate handwritten text recognition. While both are in German language, there are two main factors which make the ICDAR-2017 dataset clearly more difficult than that used in ICFHR-2016: first, the quality of the ICDAR-2017 test-set images is very variable and generally worse; second, the number of different characters involved is significantly larger (168 vs. 92). These differences likely account for the 40% and 14% worse CER and WER respectively achieved in ICFHR-2016 with respect to ICDAR-2017, using our benchmark systems without a LM (from 4.8% to 6.7%, see tables 7 and 9).

The difficulties of the ICDAR-2017 dataset are partially overcome when a LM is used: using our benchmark systems the CER is now only 29% worse and almost the same WER is finally achieved in both datasets. These results clearly indicate that the LM has a larger impact in ICDAR-2017, which rather obviously due to the much larger amount of characters (8 360 875) available for character N-gram training in this dataset, with respect to ICFHR-2016 (only 235 249).

Regarding the *Traditional challenge* (see Table 9), all the systems achieved reasonably good results. Without using a LM, our baseline system achieves a slightly lower CER, which can be explained by differences in the CNNN architecture. However our WER results are comparatively worse. This can be due by two likely causes: First, our baseline system was less accurate at detecting punctuation marks. And second the character errors committed

[17]No information was given by the winner about how the lines were detected.

by our system are more uniformly spread among words than those of the winner system. As mentioned in Sec. 3, this generally leads to a higher WER with respect to output transcripts with character errors more concentrated within individual words. By using a LM, both our CER and WER results improved significantly and the proposed baseline system achieves the best HTR accuracy published so far for the ICDAR-2017 *Traditional challenge*.

Concerning the *Advanced challenge* (see Table 10), the accuracy of the winner system was excellent. This result confirmed that the currently technology for detecting and extracting line images in a completely automatic way is mature enough for achieving excellent recognition results. As opposed to all the other benchmarks discussed in this paper, in this challenge our results were not the best ones. A visual analysis revealed that many of our errors were caused by incorrectly detected lines. In order to provide an objective confirmation of this observation, in Table 10 we also report results after manually fixing all the line detection errors. It is worth noting that in this case we achieve practically the same CER as the ICDAR 2017 winner, but somewhat worse WER.

As in the previous experiments, for the *Traditional challenge*, Fig. 9 shows how the word errors are distributed with respect to word relative positions and lengths. The tendencies observed are similar to those of the ICFHR-2014 and ICDAR-2015 datasets, but it is in this case where the greater positive effect of the LM for words towards the middle of the lines is more clearly and consistently observed. As for single-character symbols, 1.5% are punctuation symbols and 3.6% are normal symbols. The WER on punctuation symbols is 1.1% without LM and is 1.0% with LM.

Again, we consider that for improving the WER in this benchmark, the recognition should be performed at paragraph or page level in order to avoid the loose of linguistic context caused by breaking the text into lines.

9. Conclusions and outlook

A set of four benchmarks aimed at HTR research for historical and legacy documents have been introduced. These benchmarks are based on the datasets and rules previously adopted in well known open HTR competitions. For each dataset, a comprehensive description is given, along with full details of techniques and corresponding open-source tools required to implement a state-of-the-art baseline system. The required information to readily access all the required materials (datasets and tools) is provided. Except

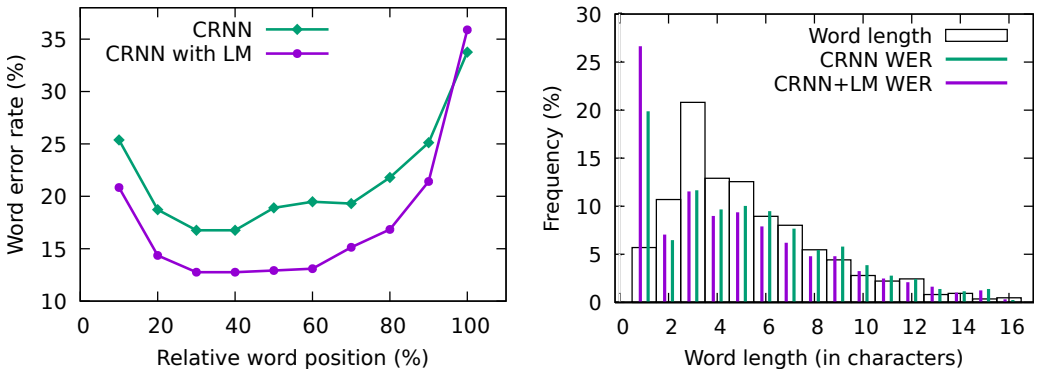


Figure 9: ICDAR-2017 benchmark. Left: WER for different relative positions of the words within the line. Right: normalized word length histogram, along with the percentage of word errors for each specific length. Results with and without using a LM are shown.

in one specific experiment, the proposed baselines overcome all previously published results, as summarized in Table 11.

Table 11: Summary of the best results achieved in all the benchmarks.

Benchmark	Best so far		This paper		Table
	CER(%)	WER(%)	CER(%)	WER(%)	
ICFHR-2014 Restricted	**5.0**	14.6	**5.0**	**9.7**	3
ICDAR-2015 Restricted	15.5	30.2	**12.8**	**30.0**	5
ICFHR-2016 Restricted	4.8	20.9	**4.5**	**17.5**	7
ICDAR-2017 Traditional	7.0	19.1	**5.8**	**17.6**	9
ICDAR-2017 Advanced	6.4	**16.8**	**6.3**	18.5	10

The increasingly demanding challenges introduced in this series of benchmarks faithfully aimed to approach the difficulties generally entailed by real transcription tasks for many kinds of relatively simple historical and archival document collections. Therefore, even though the results reported in this paper only qualify as "laboratory results", they are actually not too far from what can be achieved in the real world for these kinds of collections. In fact, projects like READ[9] and its publicly available TRANSKRIBUS platform[16], have proved very successful at making the technologies which make these results possible available for practical use in real applications of interest to archives, libraries and general public. Of course, even for manageable handwritten documents like those considered in this work, each new collection needs some preparatory work before it can undergo automatic processing. Main issues to consider include: analysis of the required alphabet, decid-

ing how to deal with (generally frequent) abbreviations and which kind of transcription is aimed at (diplomatic or modernized), etc. In addition, using transcribed samples of the documents considered to (re-)train character optical models and language models, generally leads to significant improvements over just using existing models that had previously trained on "similar" documents. Clearly, if a large collection of, say, one million page images is to be transcribed, the cost of transcribing a few hundreds training images is negligible, as compared to the overall cost of the project (including document scanning) and the benefits of obtaining an accurate textual rendering of the *contents* of the collection.

Despite these successful steps, further R&D work is much required to address many important challenges which remain largely unsolved nowadays.

So far, HTR research, including the work presented in this paper, has mainly dealt with strictly "sequential" recognition of text. That is, by capitalizing on line detection, text is assumed to be somehow presented as the outcome of a kind of uni-dimensional process where characters, words and sentences are produced in a pure sequential, left-to-right manner. This view is reminiscent of considering HTR as the same, or a very similar problem as Automatic Speech Recognition (ASR), from which HTR borrows many fruitful concepts and techniques. However, this view side-skips perhaps the most challenging and distinguishing (with respect to ASR) aspect of handwritten documents; namely the intrinsically bi-dimensional nature of the process actually underlying the production of these documents.

Following this discussion, along with the benchmark results presented in this paper, the following topics emerge for future research.

Currently, the layout analysis problems entailed by simple documents like those presented in this paper can be considered practically solved, since the main (or only) objective is to detect the text lines. Therefore, we think that more complex historical documents need to be considered to challenge future research both in layout analysis itself and in more advanced forms of HTR which are explicitly aware of the possible layout regions and elements of the documents to be transcribed. Historical and legacy handwritten documents of this kind can be counted by the billions: loosely formated tabular data, birth, marriage and death records, logbooks, minutes, drawings, mixed printed and handwritten text, etc. Clearly, for many of these documents the problem can become ill-posed if layout analysis is just based on geometric reasoning: In most cases, actually reading some text is the only way to reliably tell which text elements (e.g, lines) belong to different layout regions.

Therefore holistic approaches which integrate HTR and layout analysis, such as that proposed in [36], are interesting ideas to follow.

In all these situations, and many others, a fundamental problem is how to establish a correct *reading order* of text lines, even if they may have been correctly detected and transcribed. Therefore, research in this specific problem will probably be also important in the coming future.

On a more detailed grain, but also related with the above discussion, it is important to acknowledge that most errors committed by current, state-of-the-art systems tend to concentrate on the ends of the text lines. This has been almost systematically the case in all the benchmarks studied in this paper. Therefore, we think that an immediate challenge to be considered in coming research is to find solutions to this problem which, in turn, is related to problems raised by punctuation marks and hyphenated words. Throughout this paper we have hinted a default solution to this problem which consists in concatenating all the lines extracted from a page image or text region into an elongated image to be processed as a whole, single-line document. Clearly, this may be appropriate for simple–layout documents such as those considered in this paper, but it does not seem the right way to follow to approach the problems discussed above.

In addition to these problems, other foreseen challenges include heavily deteriorated documents, different acquisition and scanning conditions, different character sizes, difficult and/or mixed languages, abbreviations, lack of (or prohibitive cost of producing) training data, etc. Although many HTR advances are expected in the future, effective solutions to the many foreseen problems will likely come slowly and, in the near future, there will be billions of important documents for which HTR results will just be not good enough. In particular, for large or very large document collections, perhaps the most we can expect in the coming years is to make them searchable – as opposed to get them transcribed. Consequently, we think that scalable techniques for handwritten text indexing and search such as those presented in [37, 38] will be an important research topic in the near future.

Acknowledgments

This work has been partially supported through the European Union's H2020 grant READ (Recognition and Enrichment of Archival Documents) (Ref: 674943), as well as by the BBVA Foundation through the 2017-2018 and 2018-2019 Digital Humanities research grants "Carabela" and "HisClima

– Dos Siglos de Datos Cilmáticos", and by EU JPICH project "HOME - History Of Medieval Europe" (Spanish PEICTI Ref. PCI2018-093122).

[1] H. Bunke, M. Roth, E. Schukat-Talamazzini, Off-line cursive handwriting recognition using hidden markov models, Pattern Recognition 28 (9) (1995) 1399 – 1413.

[2] I. Bazzi, R. Schwartz, J. Makhoul, An omnifont open-vocabulary OCR system for English and Arabic, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (6) (1999) 495–504.

[3] A. H. Toselli, A. Juan, D. Keysers, J. González, I. Salvador, H. Ney, E. Vidal, F. Casacuberta, Integrated handwriting recognition and interpretation using finite-state models, International Journal of Pattern Recognition and Artificial Intelligence 18 (4) (2004) 519–539.

[4] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (5) (2009) 855–868.

[5] T. Bluche, Deep neural networks for large vocabulary handwritten text recognition, Ph.D. thesis, Ecole Doctorale Informatique de Paris-Sud - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, discipline : Informatique (May 2015).

[6] U. Marti, H. Bunke, The IAM-database: an English sentence database for offline handwriting recognition, International Journal on Document Analysis and Recognition 1 (5) (2002) 39–46.

[7] P. Voigtlaender, P. Doetsch, H. Ney, Handwriting recognition with large multidimensional long short-term memory recurrent neural networks, in: International Conference on Frontiers in Handwriting Recognition, 2016, pp. 228–233.

[8] V. Romero, A. Fornés, N. Serrano, J. A. Sánchez, A. H. Toselli, V. Frinken, E. Vidal, J. Lladós, The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition, Pattern Recognition 46 (6) (2013) 1658–1669.

[9] V. Romero, A. Fornés, E. Vidal, J. A. Sánchez, Information extraction in handwritten marriage licenses books using the MGGI methodology, in: IbPRIA, 2017, pp. 287–294.

[10] E. Granell, E. Chammas, L. Likforman-Sulem, C. D. Martínez-Hinarejos, C. Mokbel, B.-I. Cirstea, Transcription of Spanish historical handwritten documents with deep neural networks, Journal of Imaging 4 (1).

[11] M. Kassis, A. Abdalhaleem, J. El-Sana, VML-HD: The historical Arabic documents dataset for recognition systems, in: ASAR, 2017, pp. 11–14.

[12] A. H. Toselli, E. Vidal, Handwritten text recognition results on the Bentham collection with improved classical n-gram-HMM methods, in: International Workshop on Historical Document Imaging and Processing, 2015, pp. 15–22.

[13] J. Puigcerver, Are multidimensional recurrent layers really necessary for handwritten text recognition?, in: International Conference on Document Analysis and Recognition, Vol. 01, 2017, pp. 67–72.

[14] A. Graves, J. Schmidhuber, Offline handwriting recognition with multidimensional recurrent neural networks., in: NIPS, 2008, pp. 545–552.

[15] S. Basu, C. Chaudhuri, M. Kundu, M. Nasipuri, D. Basu, Text line extraction from multi-skewed handwritten documents, Pattern Recognition 40 (6) (2007) 1825 – 1839.

[16] E. Kavallieratou, N. Fakotakis, G. Kokkinakis, Slant estimation algorithm for ocr systems, Pattern Recognition 34 (12) (2001) 2515 – 2522.

[17] S. He, L. Schomaker, Deepotsu: Document enhancement and binarization using iterative deep learning, Pattern Recognition 91 (2019) 379 – 390.

[18] J. Calvo-Zaragoza, A.-J. Gallego, A selectional auto-encoder approach for document image binarization, Pattern Recognition 86 (2019) 37 – 47.

[19] R. Kneser, H. Ney, Improved backing-off for m-gram language modeling, in: International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, USA, 1995, pp. 181–184.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, The Kaldi speech recognition toolkit, in: Workshop on Automatic Speech Recognition and Understanding, 2011.

[21] T. Causer, V. Wallace, Building a volunteer community: results and findings from Transcribe Bentham, Digital Humanities Quarterly 6 (2).

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

[23] J. A. Sánchez, V. Romero, A. H. Toselli, E. Vidal, ICFHR2014 competition on handwritten text recognition on tranScriptorium datasets (HTRtS), in: International Conference on Frontiers in Handwriting Recognition, 2014, pp. 181–186.

[24] J. A. Sánchez, A. H. Toselli, V. Romero, E. Vidal, ICDAR 2015 competition HTRtS: Handwritten text recognition on the tranScriptorium dataset, in: International Conference on Document Analysis and Recognition, 2015, pp. 1166–1170.

[25] J. A. Sánchez, V. Romero, A. H. Toselli, E. Vidal, ICDAR2016 competition on handwritten text recognition on the READ dataset, in: International Conference on Frontiers in Handwriting Recognition, 2016, pp. 630–635.

[26] J. A. Sánchez, V. Romero, A. H. Toselli, M. Villegas, E. Vidal, ICDAR2017 competition on handwritten text recognition on the READ dataset, in: International Conference on Document Analysis and Recognition, 2017, pp. 1383–1388.

[27] S. Pletschacher, A. Antonacopoulos, The PAGE (Page Analysis and Ground-truth Elements) format framework, in: International Conference on Pattern Recognition, 2010, pp. 257–260.

[28] T. Strauß, T. Grüning, G. Leifert, R. Labahn, Citlab ARGUS for historical handwritten documents, 2014. `arXiv:1412.3949`.

[29] M. Villegas, V. Romero, J. A. Sánchez, On the modification of binarization algorithms to retain grayscale information for handwritten text

recognition, in: R. Paredes, J. Cardoso, X. Pardo (Eds.), Pattern Recognition and Image Analysis: 7th Iberian Conference, IbPRIA 2015, Santiago de Compostela, Spain, June 17-19, 2015, Proceedings, 2015, pp. 208–215.

[30] T. Strauß, Decoding the output of neural networks. a discriminative approach, Ph.D. thesis, Universität Rostock (June 2017).

[31] E. Chammas, C. Mokbel, L. Likforman-Sulem, Handwriting recognition of historical documents with few labeled data, in: Document Analysis Systems, 2018, pp. 43–48.

[32] A. Fawzi, M. Pastor, C. D. Martínez-Hinarejos, Baseline detection on Arabic handwritten documents, in: DocEng, 2017, pp. 193–196.

[33] V. Romero, A. H. Toselli, V. Bosch, J. A. Sánchez, E. Vidal, Automatic alignment of handwritten images and transcripts for training handwritten text recognition systems, in: Document Analysis Systems, 2018, pp. 328–333.

[34] T. Grüning, G. Leifert, T. Strauß, R. Labahn, A Two-Stage Method for Text Line Detection in Historical Documents, 2018. `arXiv:1802.03345`.

[35] M. Diem, F. Kleber, S. Fiel, T. Grüning, B. Gatos, cbad: Icdar2017 competition on baseline detection, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Vol. 01, 2017, pp. 1355–1360.

[36] T. Bluche, Joint line segmentation and transcription for end-to-end handwritten paragraph recognition, in: Neural Information Processing Systems, 2016, pp. 838–846.

[37] A. H. Toselli, E. Vidal, V. Romero, V. Frinken, HMM word graph based keyword spotting in handwritten document images, Information Sciences 370-371 (2016) 497–518.

[38] T. Bluche, S. Hamel, C. Kermorvant, J. Puigcerver, D. Stutzmann, A. H. Toselli, E. Vidal, Preparatory KWS experiments for large-scale indexing of a vast medieval manuscript collection in the HIMANIS project, in: International Conference on Document Analysis and Recognition, 2017, pp. 311–316.