IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Improving Transcription of Manuscripts with Multimodality and Interaction

Emilio Granell, Carlos-D. Martínez-Hinarejos, Verónica Romero

Pattern Recognition and Human Language Technology Research Center,
Universitat Politècnica de València, Camí de Vera s/n, 46022, València, Spain
{egranell,cmartine,vromero}@dsic.upv.es

Abstract

State-of-the-art Natural Language Recognition systems allow transcribers to speed-up the transcription of audio, video or image documents. These systems provide transcribers an initial draft transcription that can be corrected with less effort than transcribing the documents from scratch. However, even the drafts offered by the most advanced systems based on Deep Learning contain errors. Therefore, the supervision of those drafts by a human transcriber is still necessary to obtain the correct transcription. This supervision can be eased by using interactive and assistive transcription systems, where the transcriber and the automatic system cooperate in the amending process. Moreover, the interactive system can combine different sources of information in order to improve their performance, such as text line images and the dictation of their textual contents.

In this paper, the performance of a multimodal interactive and assistive transcription system is evaluated on one Spanish historical manuscript. Although the quality of the draft transcriptions provided by a Handwriting Text Recognition system based on Deep Learning is pretty good, the proposed interactive and assistive approach reveals an additional reduction of transcription effort. Besides, this effort reduction is increased when using speech dictations over an Automatic Speech Recognition system, allowing for a faster transcription process.

Index Terms: speech recognition, human-computer interaction, handwriting recognition, assistive transcription, deep learning

1. Introduction

Many documents used every day include handwritten text and, in many cases, such as for detecting fraudulent bank checks [1], it would be interesting to recognise these text images automatically. However, state-of-the-art handwritten text recognition (HTR) systems can not suppress the need of human supervision when high quality transcriptions are needed [2, 3, 4, 5].

A way of taking advantage of the HTR system is to combine it with the knowledge of a human transcriber, constituting the so-called Computer Assisted Transcription of Text Images (CATTI) scenario [6]. In this framework, the HTR system and the transcriber cooperate interactively to obtain the perfect transcription of the text line images. At each interaction step, the system uses the text line image and a previously validated part (prefix) of its transcription to propose an improved output. Then, the user finds and corrects the next system error, thereby providing a longer prefix which the system uses to suggest a new, hopefully better continuation. Moreover, the accuracy of the interactive system can be improved by providing it with additional sources of information, such as the speech dictation of the handwritten text over an automatic speech recognition (ASR) system.

In previous related works, multimodal combination was used to integrate the transcriber feedback into the main stream of information for word correction, by using on-line handwriting and speech [7, 8, 9]. In this work, we explore the idea of using speech dictation for feeding the interactive system with an additional source of information of the full text to transcribe.

The rest of the paper is organised as follows: Section 2 presents our multimodal proposal; Section 3 introduces the experimental framework; Section 4 explains the performed experiments and the obtained results; finally, Section 5 offers the conclusions and future work lines.

2. Multimodal Computer Assisted Transcription of Text Images

This section presents our proposal, which is composed of two parts, multimodal recognition and interaction.

2.1. Multimodal Recognition Framework

The natural language recognition problem aims to recover the text represented in an input signal. In the case of HTR, this input signal is usually a segmented line of a digitalised handwritten document [10]. Then, given a handwritten text line image or a speech signal represented by a feature vector sequence $x = (x_1, x_2, \ldots, x_{|x|})$, the problem for HTR and ASR is finding the most likely word sequence $\hat{w}$ [2], that is:

$$\hat{w} = \arg\max_{w \in W} \mathrm{P}(w \mid x) = \arg\max_{w \in W} \mathrm{P}(x \mid w)\mathrm{P}(w) \qquad (1)$$

where W denotes the set of all permissible sentences, $\mathrm{P}(w)$ is the probability of $w = (w_1, w_2, \ldots, w_{|w|})$ approximated by the language model (usually n-gram) [11], and $\mathrm{P}(x \mid w)$ is the probability of observing x by assuming that w is the underlying word sequence for x, evaluated by the optical or acoustical models for HTR and ASR, respectively. For both modalities, the state-of-the-art morphological models are based on deep neural networks [3, 12].

In this work, the search (or decoding) of $\hat{w}$, for both modalities, was performed by using the EESEN decoding method [13], which is based on Weighted Finite State Transducers (WFST). The main reason for using this decoding method is that it allows obtaining not only a single best hypothesis, but also a huge set of best hypotheses compactly represented into a unimodal word lattice, in systems with morphological models based on neural networks.

To obtain the multimodal final output lattice, the lattices generated by the unimodal systems can be combined by removing the total cost of all paths from the unimodal lattices and by doing a union of the reweighted lattices [14]. The architecture of our multimodal recognition framework is presented in Figure 1, that shows that the output is a word lattice (more details in Section 3.3).

10.21437/IberSPEECH.2018-20

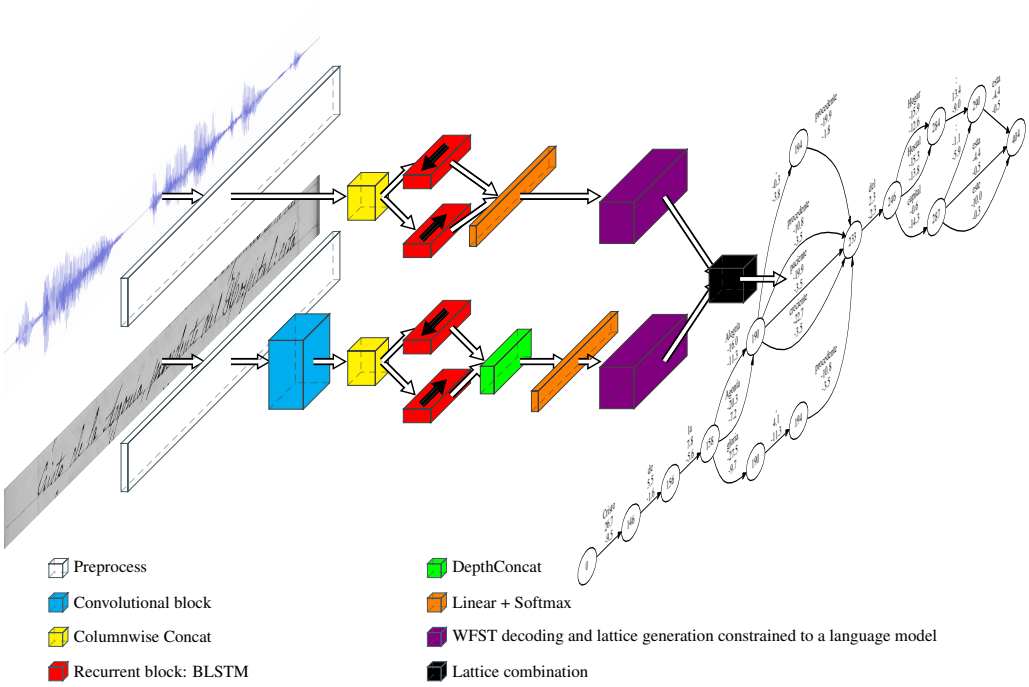

Figure 1: *Multimodal (handwriting and speech) recognition system architecture.*

2.2. Multimodal and Interactive Framework

In the CATTI framework the transcriber is involved in the transcription process, since he/she is responsible for validating and/or correcting the system hypothesis. The system takes into account the handwritten text image and the transcriber feedback in order to improve the proposed hypotheses [6]. An example of a CATTI operation is shown in Figure 2. In this example, in the traditional post-edition approach, a transcriber should have to correct about two errors from the recognised hypothesis (*Agonía* and *Hospital*). However, using our interactive approach only one explicit user-correction is necessary to get the correct transcription.

Formally, in the traditional CATTI framework [6], the system uses a given feature sequence, x_{htr}, representing a handwritten text line image and a user validated prefix p of the transcription. In this work, in addition to x_{htr}, a sequence of feature vectors x_{asr}, which represents the speech dictation of the textual contents of the text line image, is used to improve the system performance. Therefore, the CATTI system should try to complete the validated prefix by searching for a most likely suffix $\hat{s}$ taking into account both sequences of feature vectors. Following the asumptions presented in [15], the CATTI problem can be formulatted as:

$$\hat{s} = \arg\max_{s} \mathrm{P}(x_{htr} \mid p, s) \cdot \mathrm{P}(x_{asr} \mid p, s) \cdot \mathrm{P}(s \mid p) \qquad (2)$$

where the concatenation of p and s is w. As in conventional HTR and ASR, $\mathrm{P}(x_{htr} \mid p, s)$ and $\mathrm{P}(x_{asr} \mid p, s)$ can be approximated by morphological models and $\mathrm{P}(s \mid p)$ by a language model conditioned by p. Therefore, the search must be performed over all possible suffixes of p [6].

This suffix search can be efficiently carried out by using lattices [6] obtained from the combination of the HTR and ASR recognition outputs. In each interaction step, the decoder parses the validated prefix p over the lattice and then continues searching for a suffix which maximises the posterior probability according to Equation (2). This process is repeated until a complete and correct transcription of the input text line image is obtained.

3. Experimental Framework

This section presents the datasets, the preprocess, the models, the system setup, and the evaluation metrics used in the experiments.

3.1. Datasets

The datasets used in this work correspond to a Spanish historical manuscript, a Spanish phonetic corpus, and a set of speech samples provided by five different native Spanish speakers.

3.1.1. Historical Manuscript: The Cristo Salvador Corpus

The Cristo Salvador corpus is a 19th century Spanish manuscript provided by *Biblioteca Valenciana Digital* (BiValDi), and it is publicly available for research purposes on the website of the Pattern Recognition and Human Language Technology (PRHLT) research center [1]. It is a single writer book composed of 53 pages (the page 41 is presented in Figure 3)

[1] https://www.prhlt.upv.es

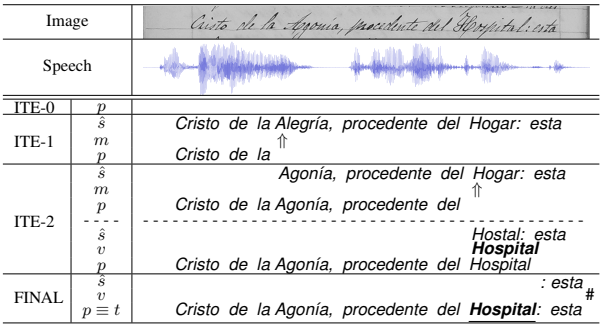

Figure 2: *Example of CATTI operation using mouse-actions (MA). Starting with an initial recognised hypothesis $\hat{s}$ from the combination of both modalities, the user validates its longest well-recognised prefix p, making a MA m (that is, positioning the cursor in the place where the error is), and the system emits a new recognised hypothesis $\hat{s}$. As the new hypothesis corrects the erroneous word, a new cycle starts. After the new validation, the system provides a new suffix $\hat{s}$ that does not correct the mistake; thus, the user types the correct word v, generating a new validated prefix p that is used to suggests a new hypothesis $\hat{s}$. This process is repeated until the final error-free transcription t is obtained. The underlined boldface word in the final transcription is the only one which was corrected by the user.*

that were manually divided into lines (such as the line shown at the top of Figure 4). This corpus presents some problematic image features, such as smear, background variations, differences in bright, and bleed-through (ink that trespasses to the other surface of the sheet).

We followed the directives of the *hard* partition defined in previous works [16, 17]. The first 30 pages (662 text lines) were used for training the optical and language models, while the following 3 pages (78 text lines) were used for validation purposes. The test set was composed of the lines of the page 41 (24 lines, 222 words); this page was selected for being, according to preliminary error recognition results, a representative page of the whole test set (the remaining 20 pages, 473 lines). This corpus contains 1213 lines, with a vocabulary of 3451 different words, and a set of 92 different characters, taking into account lowercase and uppercase letters, numbers, punctuation marks, special symbols, and blank spaces.

3.1.2. Speech Dataset: Albayzin and Cristo Salvador

The Spanish phonetic corpus Albayzin [18] was used for training the ASR acoustical models. This corpus consists of a set of three sub-corpus recorded by 304 speakers using a sampling rate of 16 KHz and a 16 bit quantisation. The training partition used in this work includes a set of 6800 phonetically balanced utterances, specifically, 200 utterances read by four speakers, 25 utterances read by 160 speakers, and 50 sentences read by 40 speakers with a total duration of about 6 hours. A set of 25 acoustical classes, 23 monophones, short silence, and long silence, was estimated from this corpus.

Test data for ASR was the product of the acquisition of the dictation of the contents of the lines of the page 41 by five different native Spanish speakers (i.e., a total set of 120 utterances, with a total duration of about 9 minutes) using a sample rate of 16 KHz and an encoding of 16 bits (to match the conditions of Albayzin data).

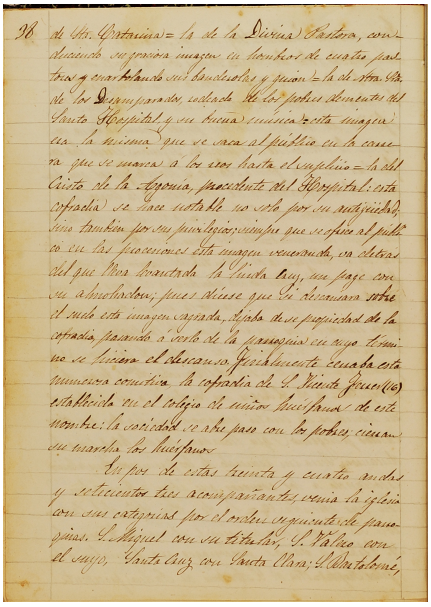

Figure 3: *Page 41 of the Cristo Salvador corpus.*

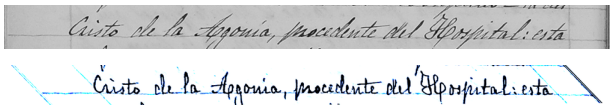

Figure 4: *Examples of an extracted text line image (top) and the result of the preprocess given to the neural network (bottom).*

3.2. Preprocess and Feature Extraction

All the text line images were scaled to 64 pixels in height and a pre-processing was applied for correcting the slant and removing the background noise [19]. A text line image and the resulting image after the image preprocess are presented in Figure 4.

With respect to speech feature extraction, 39 Mel-Frequency Cepstral Coefficients composed of the first 12 cepstrals and log frame energy with first and second order derivatives were extracted from the audio files [20].

3.3. Models

Optical models are Convolutional Recurrent Neural Networks (CRNN), which consist of a convolutional and a recurrent blocks [21]. The convolutional blocks are composed of 3 convolutional layers of 16, 32, and 48 features maps. Each convolutional layer has kernel sizes of 3×3 pixels, horizontal and vertical strides of 1 pixel, LeakyReLU as activation function, and a maximum pooling layer with non-overlapping kernels of 2×2 pixels only at the output of the first two layers. Then, the recurrent blocks are composed of 3 recurrent layers. Each recurrent layer is composed of 256 Bidirectional Long-Short Term Memory (BLSTM) units. Finally, a linear fully-connected output layer is used after the recurrent block. Those models were trained using Laia [22].

Acoustical models were trained using EESEN [13]. This acoustical model is a Recurrent Neural Network (RNN) composed of 351 inputs for 9 neighbouring frames of cepstral features, 6 hidden layers with 250 BLSTM units, and an output layer with a softmax function [12].

The lexicon models for both modalities are in HTK lexicon format, where each word is modelled as a concatenation of characters for HTR or phonemes for ASR.

The particularities of historical manuscripts, such as, writing style, epoch and subject, make it very difficult to find external resources that allow to improve the models. In general, a part of the book is used to train the models that are used to automatically transcribe the rest of the book. Therefore, the language model (LM) was estimated directly from the transcriptions of the pages included on the HTR training set using the SRILM *ngram-count* tool [23]. This language model is a 2-gram with Kneser-Ney back-off smoothing [24] interpolated with the whole lexicon in order to avoid out-of-vocabulary words, and it presents a perplexity of 742.8 for the test data.

3.4. System Setup

As previously stated, the decoding and lattice generation based on WFST for both modalities were implemented using the EESEN recogniser [13], however, the multimodal lattice combination was performed using *lattice-combine* from Kaldi [25].

In order to optimise the presented multimodal and interactive framework, the values of the main variables were set up on a validation set, as well as the limit of mouse actions for correcting each erroneous word on the interactive transcription experiments, that was set to 3 [6].

3.5. Evaluation Metrics

The quality of the transcriptions is assessed using the Word Error Rate (WER), which allows us to obtain a good estimation for the transcriber post-edition effort. The WER is based on the Levenshtein edit distance [26] and it can be defined as the minimum number of words that have to be substituted, deleted and inserted to transform the transcription into the reference text, divided by the number of words in the reference text.

The quality of the lattices can be defined as the quality of the best hypotheses contained in them, and it is known as oracle error rates. Then, the quality of the word lattices is estimated by the oracle WER, which represents the smaller WER that can be obtained from the word sequences contained in them.

The overall interactive performance is given by Word Stroke Ratio (WSR), which can also be computed by using the reference text. After each hypothesis proposed by the system, the longest common prefix between the hypothesis and the reference text is obtained and the first error from the hypothesis is corrected. This process is iterated until a full match is achieved. Therefore, the WSR can be defined as the number of user corrections that are necessary to produce correct transcriptions using the interactive system, divided by the total number of words in the reference text. This definition makes the WER comparable to the WSR. The relative difference between them gives us the effort reduction (EFR), which is an estimation of the reduction of the transcription effort that can be achieved by using the interactive system.

The statistical significance of the experimental results is estimated by means of p-values with a threshold of significance of $\alpha = 0.05$ that were calculated through the Welch t-test [27] using the statistical computing tool R [28].

4. Experimental Results

Table 1 presents the obtained experimental results. As it can be observed, in the post-edition results the quality of the lattices offered by the handwritten text recognition system is pretty good, concretely it presents a WER equal to 8.9% and an oracle WER equal to 1.8%. In this case, speech recognition does not seem to be a good substitute for handwriting recognition. The quality of the lattices obtained by the speech recognition system present a WER equal to 31.4% and an oracle WER equal to 8.5%.

Table 1: *Experimental Results.*

Experiment	Post-edition		CATTI		
	WER	**Oracle WER**	**WSR**	**EFR**	**P-value**
HTR	8.9%	1.8%	4.1%	53.9%	0.0511
ASR	31.4%	8.5%	10.4%	-16.9%	0.3898
Multimodal	10.6%	0.8%	1.8%	79.8%	0.0004

Regarding multimodality, the quality of the lattices obtained from the lattice combination of both modalities presents a WER equal to 10.6%. However, these multimodal lattices presents an oracle WER equal to 0.8%. Even though the combination technique does not improve the unimodal HTR WER, it allows to reduce the oracle WER substantially. Therefore, an outstanding effect on interactive transcription can be expected, since the oracle WER is related to the quality of the alternatives offered by the interactive and assistive system (the lower the oracle WER, the better the alternatives).

Concerning the CATTI results, 4.1% of estimated interactive human effort (WSR) was required for obtaining the perfect transcription from the HTR lattices, which represents 53.9% of relative effort reduction (EFR) over the HTR baseline (WER equal to 8.9%, $p = .051$). On the other side, no effort reduction can be considered when only ASR is used at the input of the interactive system. However as expected, the multimodal combination not only represents 56.1% of relative improvement on the estimate interactive human effort (1.8% over 4.1%, $p = .091$), but these improvements are statistically significant when compared with the HTR baseline (EFR equal to 79.8%, $p < .001$).

5. Conclusions

In this paper, the use of multimodal combination for improving the CATTI system presented in previous works has been studied. Multimodal combination allows us to provide additional sources of information to the assistive transcription system, such as speech dictation of the textual contents of the document to transcribe.

The obtained results show that the combination technique used, even though it does not improve the best hypothesis offered by the unimodal HTR system, it may produce new bigrams that increase the search alternatives. Moreover, the adjustment of the word posterior probabilities can increase the probabilities of the correct words, reaching better hypotheses that allows the assistive transcription system to provide an additional and significant reduction of the human effort.

In future work, we will study the use of other combination techniques, the use of sentences in the handwritten text corpus instead of lines (in order to make multimodality more natural), and the use of the information of context given by the previous lines.

6. Acknowledgments

This work was partially supported by the following research projects: READ - 674943 (European Union's H2020) and CoMUN-HaT - TIN2015-70924-C2-1-R (MINECO / FEDER).

7. References

[1] S. Chhabra, G. Gupta, M. Gupta, and G. Gupta, "Detecting Fraudulent Bank Checks," in *Proc. of the 15th IFIP International Conference on Digital Forensics*, 2017, pp. 245–266.

[2] 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," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 4, pp. 519–539, 2004.

[3] T. Bluche, H. Ney, and C. Kermorvant, "A Comparison of Sequence-Trained Deep Neural Networks and Recurrent Neural Networks Optical Modeling for Handwriting Recognition," in *Proc. of the 2nd SLSP*, 2014, pp. 199–210.

[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] S. España-Boquera, M. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martínez, "Improving offline handwriting text recongition with hybrid HMM/ANN models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 767–779, 2011.

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

[7] E. Granell, V. Romero, and C. D. Marínez-Hinarejos, "An Interactive Approach with *Off-line* and *On-line* Handwritten Text Recognition Combination for Transcribing Historical Documents," in *Proc. of the 12th IAPR-DAS*, 2016, pp. 269–274.

[8] C.-D. Martínez-Hinarejos, E. Granell, and V. Romero, "Comparing different feedback modalities in assisted transcription of manuscripts," in *Proc. of the 13th IAPR-DAS*, 2018, pp. 115–120.

[9] A. Toselli, V. Romero, M. Pastor, and E. Vidal, "Multimodal interactive transcription of text images," *Pattern Recognition*, vol. 43, no. 5, pp. 1824–1825, 2010.

[10] V. Romero, J. A. Sanchez, V. Bosch, K. Depuydt, and J. de Does, "Influence of text line segmentation in handwritten text recognition," in *Proc. of the 13th ICDAR*, 2015, pp. 536–540.

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

[12] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. of ICASSP*, 2013, pp. 6645–6649.

[13] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. of ASRU*, 2015, pp. 167–174.

[14] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[15] E. Granell, V. Romero, and C.-D. Martínez-Hinarejos, "Multimodality, interactivity, and crowdsourcing for document transcription," *Computational Intelligence*, vol. 34, no. 2, pp. 398–419, 2018.

[16] V. Romero, A. H. Toselli, L. Rodríguez, and E. Vidal, "Computer Assisted Transcription for Ancient Text Images," in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, M. Kamel and A. Campilho, Eds. Springer Berlin Heidelberg, 2007, vol. 4633, pp. 1182–1193.

[17] V. Alabau, V. Romero, A. L. Lagarda, and C. D. Martínez-Hinarejos, "A Multimodal Approach to Dictation of Handwritten Historical Documents." in *Proc. 12th Interspeech*, 2011, pp. 2245–2248.

[18] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mariño, and C. Nadeu, "Albayzin speech database: design of the phonetic corpus," in *Proc. of EuroSpeech*, 1993, pp. 175–178.

[19] M. Villegas, V. Romero, and J. A. Sánchez, "On the modification of binarization algorithms to retain grayscale information for handwritten tet recognition," in *Proc. of IbPRIA*, 2015, pp. 208–215.

[20] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[21] J. Puigcerver, "Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition?" in *Proc. of the 14th ICDAR*, vol. 1. IEEE, 2017, pp. 67–72.

[22] J. Puigcerver, D. Martin-Albo, and M. Villegas, "Laia: A deep learning toolkit for HTR," 2016. [Online]. Available: https://github.com/jpuigcerver/Laia/

[23] A. Stolcke, "SRILM-an extensible language modeling toolkit." in *Proc. of the 3rd Interspeech*, 2002, pp. 901–904.

[24] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. of ICASSP*, vol. 1, 1995, pp. 181–184.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. of ASRU*, 2011.

[26] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, February 1966.

[27] B. L. Welch, "The Generalization of 'Student's' Problem when Several Different Population Variances are Involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.

[28] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017, https://www.R-project.org/. Last access: May 2017.