ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset

Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Enrique Vidal
PRHLT Research Center
Universitat Politècnica de València
Valencia, Spain
{*jandreu,vromero,ahector,evidal*}*@prhlt.upv.es*

***Abstract*—This paper describes the Handwritten Text Recognition (HTR) competition on the READ dataset that has been held in the context of the International Conference on Frontiers in Handwriting Recognition 2016. This competition aims to bring together researchers working on off-line HTR and provide them a suitable benchmark to compare their techniques on the task of transcribing typical historical handwritten documents. Two tracks with different conditions on the use of training data were proposed. Ten research groups registered in the competition but finally five submitted results. The handwritten images for this competition were drawn from the German document Ratsprotokolle collection composed of minutes of the council meetings held from 1470 to 1805, used in the READ project. The selected dataset is written by several hands and entails significant variabilities and difficulties. The five participants achieved good results with transcriptions word error rates ranging from 21% to 47% and character error rates rating from 5% to 19%.**

***Keywords*-Handwritten Text Recognition, Historical documents.**

I. Introduction

This paper describes the third edition of the Handwritten Text Recognition (HTR) competition organized for the International Conference on Frontiers in Handwriting Recognition (ICFHR) 2016 in the framework of the EU tranScriptorium project [1] the first two editions [2], [3], and now in the framework of the EU READ project[1]. As previous editions, the goal of this competition was to bring together researchers for sharing new techniques and ideas on HTR for historical documents. A dataset used in the READ project was prepared for the participants and some challenges were defined for this dataset.

The "Recognition and Enrichment of Archival Documents (READ)" project is an European project that started in January 2016 and it is scheduled for 42 months. READ's mission is to revolutionize access to archival documents with the support of cutting-edge technology such as HTR and Keyword Spotting (KWS). READ has three main legs: research, service and networking. In the research part it is scheduled to promote HTR research through competitions along all the project. Many archives are involved in READ from different European countries, and therefore HTR research on many languages is expected in the project, including English, German, Spanish, Finish, Italian, French, Dutch, Greek, Latin, Arabic, etc.

In this edition, German was chosen for the competition. The proposed dataset consisted of a subset of documents from the Ratsprotokolle collection[2] composed of minutes of the council meetings held from 1470 to 1805 (about 30,000 pages), which is used in the READ project. This dataset is written in Early Modern German. The number of writers is unknown. Handwriting in this collection is complex enough to challenge the HTR software. Fig. 1 shows some sample images from the Ratsprotokolle collection.

Page images of the Ratsprotokolle collection generally entail important layout analysis difficulties (see Fig. 1), like marginal notes, fainted writing, bleed-through, skewed images, slanted lines, etc. There are also difficulties from the HTR point of view. They are written by several hands, they have crossed-out, hyphenated words, punctuation symbols, footnote symbols, etc. Even with these difficulties, most of these page images are readable for human beings. HTR results on this collection have not been reported in the past.

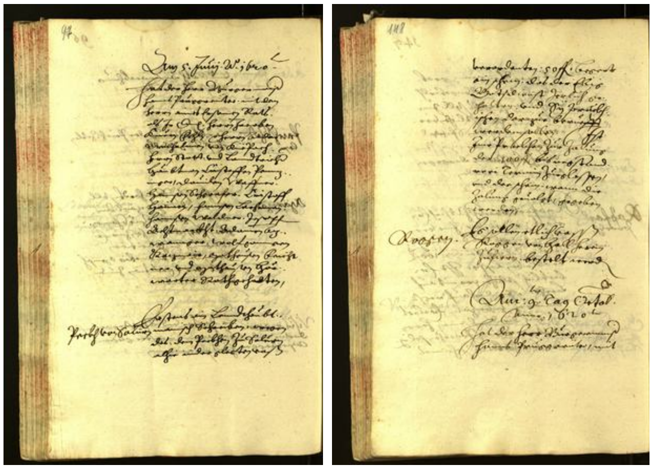

Figure 1. Document samples of the Ratsprotokolle dataset to be processed in READ.

This competition was organized by members of the Pattern Recognition and Human Language Technology research

[1] http://read.transkribus.eu/

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

center[3] that participate in READ, with the help of other members of the READ consortium. In this third edition of the competition, 10 research groups were registered[4] and finally 5 participants actually tested their systems and submitted official results. A HMM-based baseline system was prepared by the organizers. This baseline system was simple enough to guarantee that better results could be obtained easily.

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

II. Challenges of the Competition

This competition aims to bring together researchers working on off-line HTR for historical documents and provide them a suitable benchmark to compare their techniques on the task of transcribing typical historical handwritten documents. It also aims to investigate the performance of the HTR technology for historical documents.

The challenges stated in this third edition taking into account the experience of the competition in the previous editions [2], [3] and from the experience and requirements in the READ project were the following.

Challenge 1. Several approaches exist for HTR [4], [5], [6] and the machinery in each of them can be enormous both for training and decoding. Therefore, comparing different techniques can be sometimes difficult. For making easier the comparison of techniques, in this edition a track with restricted training material was mandatory for all participants.

Challenge 2. HTR techniques for historical documents have been researched in the past for many languages. But the publicly available reference datasets are usually in English [7], [8], [2] or in other very similar languages (from the language modelling point of view), like Spanish [9], [10] or Catalan [11], just to mention a few. In this competition we introduced a new dataset, this time in German. German is also similar in some aspects to English, specially from the optical modelling point of view. But language modelling is more challenging than English due to compound words.

III. Dataset Description

The dataset for this competition was composed of 450 page images, each encompassing of a single text block in most cases, but also with many marginal notes and added interlines. These pages entailed several line detection and transcription difficulties and the corresponding ground truth (GT) was produced semi-automatically and manually reviewed [12] (see examples of extracted lines in Fig. 2). The writing style in these images is characterized by having long and irregular ascenders and descenders and a tight main body text. The GT information was registered in PAGE format [13]. TEI[5] marks were removed and ignored for the competition[6].

These 450 pages contained 10,550 lines with nearly 43,500 running words and a vocabulary of more than 8,000 different words. The last column in Table I summarizes the basic statistics of these pages.

Table I
The Ratsprotokolle dataset used in the HTR competition.

Number of:	Train	Validation	Test	Total
Pages	350	50	50	450
Lines	8,367	1,043	1,140	10,550
Running words	35,169	3,994	4,297	43,460
Lexicon	6,985	1,526	1,656	8,120
Running OOV	-	669	633	-
OOV Lexicon	-	574	563	-
Character set size	92	80	83	92
Running Characters	208,595	26,654	25,179	260,428

The dataset was divided into three subsets for training, validation and testing, respectively encompassing 350, 50 and 50 page images. Since it was not possible to accurately identify the writers in all cases, this characteristic was not taken into account for distributing them over these subsets. This means that some writers could appear in the three sets.

The GT in both training and validation sets is in PAGE format and it was provided annotated at line level in the PAGE files. The transcriptions at line level were also included in the PAGE files. On the other hand, the PAGE files of the test set contained the line regions, but the transcripts were removed. It was delivered just a few days in advance to the deadline.

Table I contains basic statistics of these partitions. The rows "Running words" and "Running OOV" show the total number of words and Out-Of-Vocabulary (OOV) words, respectively. The OOV words in the Validation column are words that do not appear in the training set. The OOV words in the Test column are words that do not appear neither in the training set nor in the validation set. The row "OOV Lexicon" shows the number of *different* running OOV words.

IV. Competition Description

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

[3] http://www.prhlt.upv.es

[4] http://transcriptorium.eu/~htrcontest/

[5] http://www.tei-c.org/index.xml

[6] The dataset is publicly available for research purposes at the READ web page, the TC11 Reading Systems web page, and at the contest web page.

- The original page images of the training and validation sets.
- The PAGE file corresponding to each page image. For each text line in this image, the PAGE file contains a baseline and an automatically obtained bounding polygon [14], and the corresponding diplomatic transcript. All baselines were checked and corrected manually.

The test images, with the transcript fields empty in the PAGE file, were eventually provided in the same format as the train and validation sets for evaluation purposes.

A baseline system based on hidden Markov models trained with the Hidden Markov Model Toolkit[7] (HTK) and 2-gram models trained with the SRILM[8] toolkit was provided. A set of scripts to perform a basic training with the training set and a test with the validation set were included. The participants could use this baseline system as an initial approach. They were allowed to improve this baseline by changing one or several of the following processes: page-level pre-processing and line extraction, line pre-processing and normalization, feature extraction, recognition system approach, types of character and/or language models, etc.

The participants had to send the output transcripts obtained for the test images. Several results per participant were allowed corresponding to different runs of their own systems and all the results were considered for the final decision. Output transcripts were expected with correct capitalization and punctuation and they had to be provided tokenized in the same way as the transcripts in training data. The evaluation metric was a linear combination of the Word Error Rate (WER) and the Character Error Rate (CER) (50% each) between the reference transcript and the transcript provided by the system from each line. Note, that with this evaluation metric, the systems with character errors concentrated in few words would be considered better than systems with character errors scattered in all the words. The winner would be the system which obtained the least value of the linear combination of the metric on the test set. The entrants were informed in advance about this evaluation metric.

Two tracks were planned in this competition:

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

The entrants had to participate necessarily in the Restricted track. The purpose of defining the Restricted track was to have the possibility of comparing techniques with respect to the amount of training data used (*Challenge 1*).

The competition was planned in such a way that the participants had more than four months for preparing their systems before the test set was provided. Then, they had twelve days for sending their transcription results on the test set. Along those twelve days, the participants did not receive any feedback about their results on the test data. When the competition closed, the competitors were informed only about their own results and they were asked for submitting a description of the system for which they obtained their best results. These descriptions are summarized in Section V.

V. Description of Systems

Ten research groups registered at the competition and finally five of them submitted results. Four participants submitted results of several systems to just one track and one participant submitted results of several systems to the two tracks described in Section IV. The five research groups, listed in the same order they registered were:

- Human Language Technology and Pattern Recognition Group, Germany (RWTH)[9].
- Telecom ParisTech, France, and University of Balamand, Lebanon (ParisTech)[10].
- Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes, France (LITIS)[11].
- BYU Computer Science Department, USA (BYU)[12].
- Artificial Intelligence and Image Analysis (A2IA)[13].

The entrants submitted several results that were obtained by several systems. The main characteristics of the best system for each entrant are the following:

- **RWTH.** Unscaled images after applying the image enhancing pipeline provided by the setup were used. The features used were two-dimensional grid of image pixels. For optical modelling, they used ROVER combination [15] of 16 Multi-directional Long-Short Term Memory [16] (MLSTM) networks with about 5 MLSTM and convolutional layers and 3 times max-pooling per net. All networks were trained using the connectionist temporal classification (CTC) [17] objective function. For decoding, single-state HMMs to realize the CTC topology were used. For language modelling, a 10-gram character-based language model with Kneser-Ney smoothing estimated from the training data was used.
- **ParisTech.** Their system was a Bidirectional Long Short-Term Memory (BLSTM) recurrent neural network recognizer that consisted of the coupling of 2 recurrent neural networks. The value of an output unit at time step t is the linear combination of the outputs of the forward and backward hidden layers at this time

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

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

[9] https://www-i6.informatik.rwth-aachen.de/

[10] http://www.tsi.telecom-paristech.fr/mm/themes/equipe-ecrit-et-document/

[11] http://www.litislab.eu/

[12] https://cs.byu.edu/

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

step t. The two hidden layers are both made of 100 LSTM blocks, with one cell per block. The output layer was made of 93 neurons, corresponding to the different characters, numbers and punctuation marks.

The BLSTM recognizer was trained with a gradient-based method on all the provided images during the competition. After each training epoch, the recognition error rate was evaluated on a validation set. If error rates do not improve for 20 epochs, network training was stopped. This strategy avoids data over-fitting.

The BLSTM computed for each frame its corresponding network outputs, each of them being associated to a character class. These outputs were normalized, providing for each character class, the posterior probability. Then the backward-forward token passing algorithm (CTC) took the posteriors as input and provides a sequence of words given the dictionary and the bi-gram language model that was created from the transcription data provided during the competition.

Feature vectors of 20 (geometric and statistical) coefficients were extracted, via a left-to-right sliding window of 9 pixels in width and 2 pixels shift. Geometric coefficients were related to the counts of pixels, the concavity values in the different cells in the window and the position of the baseline of writing and the average position of the pixels with respect to this baseline. The statistical coefficients reflected density of the pixels in different cells and directions

- **LITIS.** They used the training and validation data as it was provided, with the exact same split. Small transformations (rotation, shredding) to artificially extend the number of training images were applied.

  Images were normalized to a 100 pixels height. Histograms of gradient with a 8-pixel wide window and a 1 pixel pace were obtained.

  A three layers $(100, 70, 120)$ BLSTM Recurrent Neural Network that was trained with RNNLIB[14] was used.

  No language model was used and only the words given in the training and validation sets were used as a lexicon.

  The decoding was based on the combination of multiple BLSTM (over 20), and the string output was verified word by word with the lexicon.
- **BYU.** Their system used a CNN and CTC, based on the network described in [18]. Training data was augmented by evenly placing control points across the image and randomly displacing the control points. Ten synthetic instances were produced for every original line image.

  The images were pre-processed using three techniques: the baseline system's pre-processing provided by the organizers, the binarization described in [19], and the gray-scale of the image. The three images for each method were joined as separate channels and then presented to the network.

  Post processing was applied to the page number results. Page images were given in page order. Page numbers with high confidence were used to correct the less-confident neighbouring page numbers.
- **A2iA. Restricted track.** The train and validation sets were used for training the models. For each line two segmentations were used: the original polygon, given in the PAGE xml and an extension of the original polygon towards the boundaries of the neighbour polygons, in order to get lost context back. The lines images were converted to gray-scale.

  For optical modelling, it used a MLSTM-RNN trained with CTC, alternating LSTM layers (in four directions), convolution layers with 2×4 subsampling, feed-forward merging directions and a non-linear function. Drop-out was carried out after each LSTM. The softmax output layer models 87 characters and a blank symbol.

  A hybrid word/character language model was trained using the method presented in [20]. Words were modelled by 2-grams and characters by 7-grams, both with Witten-Bell smoothing. They were estimated on the concatenation of two versions of the transcriptions: the original line transcriptions, which may include parts of hyphenated words, as well as the hyphenation symbol and the entire sentences, delimited by full-stops, with hyphenated words sticked back by removing extra hyphenation symbol and newline characters.

  The decoding was carried out using weighted finite-states transducers with beam search using Kaldi [21].
- **A2iA. Unrestricted track.** The main difference between the system used in this track and the previous one is the quantity of samples used to train the optical models. In addition to the training and validation sets, for this track, lines of HWGL, an in-house dataset of around 500 modern handwritten German letters, have been used. In addition, a first convolution layer with 2×2 subsampling and dropout before the first LSTM layer has been included in the MDLST-RNN.

As a final comment on the description of the systems, it is important to remark that CNN/CTC techniques were used by all the entrants for training the optical models.

VI. Results

The best results obtained by each participant can be seen in Table II. In the restricted track four entrants obtained similar results and the last entrant obtained slightly worse results. In the unrestricted track, only A2IA participated and the obtained results slightly overcome its result in the restricted track.

As previously commented, in order to encourage participation, the initial baseline system provided to the partici-

[14] https://sourceforge.net/projects/rnnl/

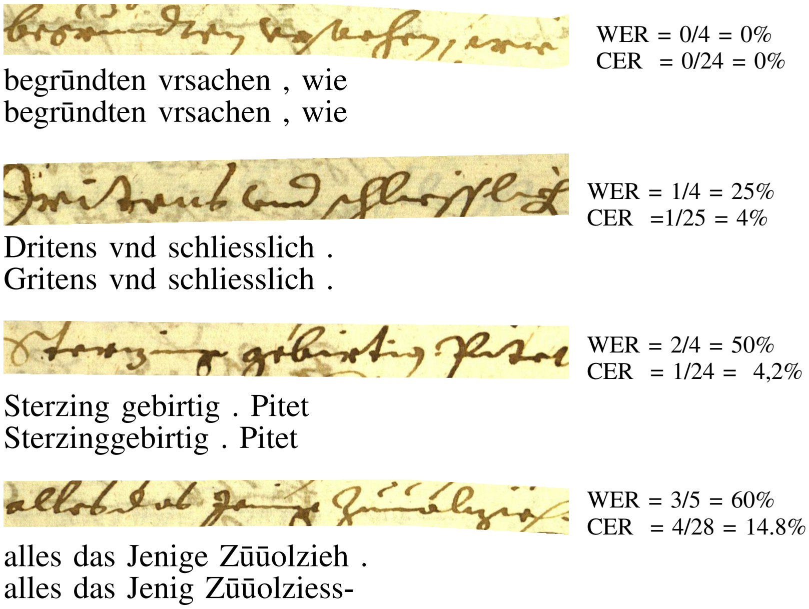
WER = 0/4 = 0%
CER = 0/24 = 0%
begrūndten vrsachen , wie
begrūndten vrsachen , wie
WER = 1/4 = 25%
CER =1/25 = 4%
Dritens vnd schliesslich .
Gritens vnd schliesslich .
WER = 2/4 = 50%
CER = 1/24 = 4,2%
Sterzing gebirtig . Pitet
Sterzinggebirtig . Pitet
WER = 3/5 = 60%
CER = 4/28 = 14.8%
alles das Jenige Zūūolzieh .
alles das Jenig Zūūolziess-

Figure 2. Examples of some test line images sorted according to the WER from top to down obtained by the best system. The reference transcript and the RWTH system hypothesis are displayed (in this order) below of each image. The corresponding WER and CER figures are also shown on the right of each image.

Table II
BEST WORD ERROR RATE AND CHARACTER ERROR RATE (WER / CER) OBTAINED BY THE PARTICIPANTS ON EACH TRACK.

	Restricted track	Unrestricted track
RWTH	20.9±1.2 / 4.8±0.3	-
BYU	21.1±1.2 / 5.1±0.3	-
A2IA	22.1±1.2 / 5.4±0.3	21.0±1.2 / 5.1±0.3
LITIS	26.1±1.3 / 7.3±0.4	-
ParisTech	46.6±1.5 / 18.5±0.5	-

pants was extremely simple. Its WER was as high as 56.1%. All the participants did overcome this result loosely. Note that this result was based on a word-based LM and according to Table I, the Running OOV ratio was about 15%, which means about 30% expected WER[15] because OOV words.

It is interesting to remark that the best result, obtained by RWTH, has been obtained in the restricted track. The result obtained in the unrestricted track, in spite of using additional images for training, do not improve the RWTH result. It is also noticeable that all the systems used CNN/CTC for training and decoding and that results of the four first participants was quite similar. Thus, the organizers obtained 22.7% WER and 5.8% CER using the same technology and additional noisy training data, but without using system combination techniques.

Note also that WER was high for all participant. The reason for this was that those which used LM used just character-based LM (avoiding OOV word problems) or word-based LM with the problem of OOV words.

Figure 2 shows the transcripts for several lines provided by the RWTH system sorted according to their WER. Note that even for the line with the largest WER, the automatic transcript can be useful both for reading and for searching. Finally, the big differences found between CER and WER in Table II can be explained by the wrong segmentation of the words made during the recognition. For example, in the third example, the CER is very low, but a incorrect segmentation of the word increase the WER significantly.

VII. Conclusions

This paper described the HTR competition that was organized in the context of the ICFHR 2016 conference. The competition has been carried out wit the Ratsprotokolle collection that was prepared in the READ project. The five entrants obtained very good results with this dataset at character level. The results at word level were not as good as the character level.

For future work, we plan to carry out this competition with different data that will include more challenges like writing styles, crossed-out text, fainted texts, larger vocabularies and different languages. Another challenge for the future is to deal with less GT data for training.

Acknowledgment

This work has been partially supported through the European Union's H2020 grant READ (Recognition and Enrichment of Archival Documents) (Ref: 674943), the

[15] Each OOV word is responsible of two errors on average: the OOV word itself and the following word because of the n-gram dependency.

MINECO/FEDER-UE project TIN2015-70924-C2-1-R, and the HIMANIS EU project, JPICH programme, (Spanish grant Ref. PCIN-2015-068).

References

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

[2] J. Sánchez, V. Romero, A. Toselli, and E. Vidal, "ICFHR2014 competition on handwritten text recognition on transcriptorium datasets (HTRtS)," in *ICFHR*, 2014, pp. 181–186.

[3] J. Sánchez, A. Toselli, V. Romero, and E. Vidal, "ICDAR 2015 competition HTRtS: Handwritten text recognition on the tranScriptorium dataset," in *13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1166–1170.

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

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

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

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

[8] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character HMMs," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 934–942, 2012.

[9] A. Toselli, V. Romero, L. Rodríguez-Ruiz, and E. Vidal, "Computer assisted transcription of handwritten text," in *9th International Conference on Document Analysis and Recognition (ICDAR 2007)*. IEEE Computer Society, 2007, pp. 944–948.

[10] N. Serrano and A. Juan, "The RODRIGO database," in *LREC*, 2010, pp. 19–21.

[11] V. Romero, A. Fornés, N. Serrano, J. Sánchez, A. Toselli, V. Frinken, E. Vidal, and J. Lladós, "The esposalles database: An ancient marriage license corpus for off-line handwriting recognition," *Pattern Recognition*, vol. 46, no. 6, pp. 1658–1669, 2013.

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

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

[14] V. Romero, J. Sánchez, V. Bosch, K. Depuydt, and J. de Does, "Influence of text line segmentation in handwritten text recognition," in *13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015.

[15] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. of ASRU 1997*, 1997, pp. 347–354.

[16] A. Graves, S. Fernández, and J. Schmidhuber, *Multidimensional Recurrent Neural Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 549–558.

[17] A. Graves, *Connectionist Temporal Classification*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 61–93. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-24797-2_7

[18] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition," 2015, http://arxiv.org/abs/1507.05717.

[19] N. Howe, "A Laplacian energy for document binarization," in *International Conference on Document Analysis and Recognition*, 2011, pp. 6–10.

[20] R. Messina and C. Kermorvant, "Surgenerative finite state transducer n-gram for out-of-vocabulary word recognition," in *International Workshop on Document Analysis Systems (DAS)*, 2014.

[21] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafiat, S. Kombrink, P. Motlcek, Y. Qian, K. Riedhammer, K. Vesely, and N. T. Vu, "Generating exact lattices in the wfst framework." IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), March 2012.