(Article begins on next page)

Learning to Read *L'Infinito*: Handwritten Text Recognition with Synthetic Training Data

Silvia Cascianelli[0000−0001−7885−6050], Marcella Cornia[0000−0001−9640−9385], Lorenzo Baraldi[0000−0001−5125−4957], Maria Ludovica Piazzi, Rosiana Schiuma, and Rita Cucchiara[0000−0002−2239−283X]

University of Modena and Reggio Emilia, Modena, Italy
{name.surname}@unimore.it

Abstract. Deep learning-based approaches to Handwritten Text Recognition (HTR) have shown remarkable results on publicly available large datasets, both modern and historical. However, it is often the case that historical manuscripts are preserved in small collections, most of the time with unique characteristics in terms of paper support, author handwriting style, and language. State-of-the-art HTR approaches struggle to obtain good performance on such small manuscript collections, for which few training samples are available. In this paper, we focus on HTR on small historical datasets and propose a new historical dataset, which we call Leopardi, with the typical characteristics of small manuscript collections, consisting of letters by the poet Giacomo Leopardi, and devise strategies to deal with the training data scarcity scenario. In particular, we explore the use of carefully designed but cost-effective synthetic data for pre-training HTR models to be applied to small single-author manuscripts. Extensive experiments validate the suitability of the proposed approach, and both the Leopardi dataset and synthetic data will be available to favor further research in this direction.

Keywords: Handwritten Text Recognition · Historical Documents · Synthetic Data.

1 Introduction

Transcribing ancient manuscripts is key in their conservation and valorization. Having searchable and easily accessible digital text will both ease scholars in their research activity and allow a broad lay public to read ancient texts (either by very famous authors and unknown people) without being experts in paleography and philology. AI-based techniques have already proven to be high-performing when it comes to recognizing patterns from visual inputs, and images of manuscripts make no exception [3,7]. In this context, state-of-the-art Handwritten Text Recognition (HTR) systems have shown remarkable performance when applied to both modern and historical manuscripts [23,18,6,12,1]. This is

allowed by the modeling capability of the deep convolutional and recurrent neural networks that constitute these systems, which, however, need a significant quantity of data to train properly. In this respect, historical documents pose critical challenges related to the historical period and geographic area in which they were edited. Calligraphical handwriting has significantly changed over the centuries, and many of the preserved more recent manuscripts have been written in the personal handwriting style of their authors. For this reason, HTR systems trained on a specific handwriting style (or even on multiple handwriting styles) can be ineffective when applied to manuscripts in a different style. Digital libraries and archives can contain valuable manuscripts by historically and culturally important authors, which could be valorized and made available to the large public by automatic transcription. However, such manuscripts can be in a small quantity, making the training of HTR systems ineffective or even unnecessary. In this respect, strategies should be developed to enable high-quality automatic transcription from few or none annotated training manuscripts.

To this end, we propose a new historical HTR dataset containing letters from the famous Italian writer Giacomo Leopardi[1]. The dataset is relatively small and poses the typical challenges of HTR for small but relevant document collections preserved in historical libraries and archives. In particular, the paper support has scratches and creases, there are stains and bleed-through ink, and the used language is early-seventeenth century Italian (thus, different from other languages such as modern English, it is not sufficiently represented in standard language corpora). The dataset is annotated at line-level, which is quicker and cheaper to obtain compared to word-level annotation and allows gathering a fair number of training samples compared to paragraph-level or page-level annotation. Moreover, line-level HTR systems are the most common in literature and can be easily integrated into a transcription pipeline in conjunction with layout analysis and line-level segmentation. The dataset is used as a test bench for the task of HTR on documents with limited training data and can be used for further research. Moreover, to deal with the aforementioned task, we consider direct transfer learning from a number of both modern and historical source datasets and pre-training plus fine-tuning on a specifically designed synthetic dataset.

2   Related Work

An important factor for the development of efficient deep learning solutions is the availability of big datasets. Here we review some of the most commonly used western-characters datasets for line-level HTR. The most commonly used line-level modern datasets are the IAM [16] and the RIMES [2] datasets. The former is a collection of handwritten English sentences from the Lancaster-Oslo/Berge (LOB) corpus [14], produced by approximately 40 different writers, for a total of 10 373 lines. The latter contains free-layout letters handwritten by multiple

---

[1] Giacomo Leopardi (Recanati, 1798 – Naples, 1837) was an Italian philologist, writer, and poet, considered to be one of the most relevant authors of the Italian Romanticism literary current. *L'Infinito* (The Infinite) is one of his most known poems.
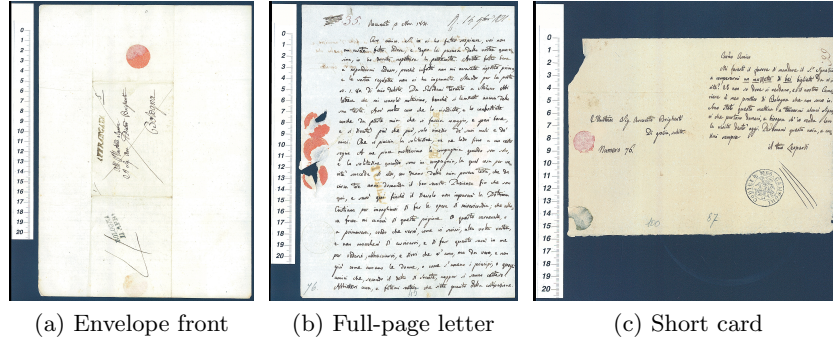
(a) Envelope front   (b) Full-page letter   (c) Short card

Fig. 1. Sample pages of the three categories considered for obtaining our devised Leopardi dataset.

authors in modern French, for a total of 12 111 lines. Among historical datasets, the largest and most commonly used line-level ones are those used for the International Conference on Frontiers of Handwriting Recognition 2014 and 2016 HTR challenges, namely the ICFHR14 [20] and ICFHR16 [21] datasets. The former consist of legal forms and drafts from the Bentham Papers collection [4], handwritten by the English philosopher Jeremy Bentham and his collaborators between mid-18th century and mid-19th century, for a total of 11 473 lines. The latter contains registers of the Bozen's Town Council's discussions from the Ratsprotokolle collection, handwritten by multiple writers from 1470 to 1805 in old German, for a total of 10 550 lines. Most of the available historical manuscript datasets are of small size. Among the most used there are those collected in the IAM-HisDB project [10]. These are the George Washington dataset, consisting of 656 lines from letters written in 18th century English by George Washington and one of his collaborators, the Parzival dataset [9], consisting of 4 477 lines of a Medieval German manuscript by two different writers, and the Saint Gall dataset [11], consisting of 1 410 lines from a single-writer 9th century Latin manuscript. Another interesting historical dataset is the Esposalles [19], a collection of 5 447 lines written in 17th century Catalan by a single writer. The number of lines in our presented Leopardi dataset is comparable to those of the aforementioned small historical datasets, is single-author and entirely in Italian.

A possible strategy to deal with HTR with small training sets is to apply data augmentation [17,26,18,27]. Typically, data augmentation consists of generic random distortions and color modifications. For big historical datasets, in [5], the authors demonstrated the benefits of carefully designed data augmentation. Another strategy is to apply transfer learning [12,13,25,15,1], *i.e.*, pretraining the HTR model on a big HTR dataset and fine-tuning it on the small training set of the dataset of interest. For HTR on small single-writer historical datasets, pretraining plus fine-tuning has been proven to be a more effective strategy than data augmentation [1]. Hence, in this work, we focus on this strategy. The dataset used for pretraining is, in general, one of the available benchmark ones.
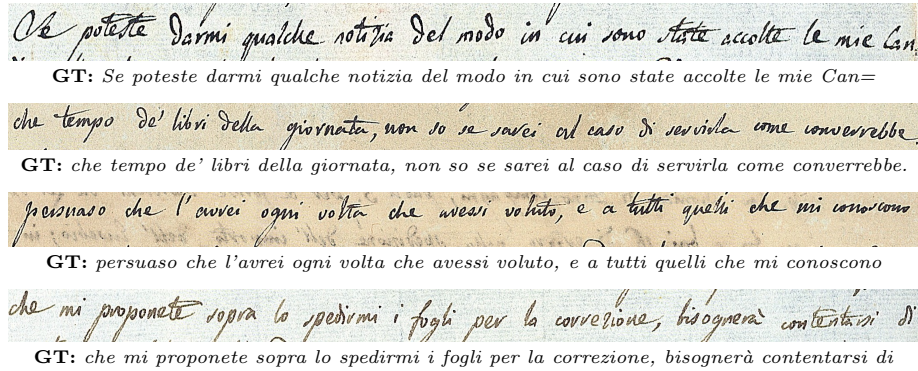
GT: *Se poteste darmi qualche notizia del modo in cui sono state accolte le mie Can=*

GT: *che tempo de' libri della giornata, non so se sarei al caso di servirla come converrebbe.*

GT: *persuaso che l'avrei ogni volta che avessi voluto, e a tutti quelli che mi conoscono*

GT: *che mi proponete sopra lo spedirmi i fogli per la correzione, bisognerà contentarsi di*

Fig. 2. Sample images and corresponding ground-truth (GT) transcription of the text lines contained in the collected Leopardi dataset.

In [22], instead, the authors used a synthetic dataset obtained by concatenating isolated handwritten Chinese characters to obtain synthetic text lines. Moreover, in [15], the authors proposed to pretrain a model on generic modern synthetic data and apply it to modern HTR data. Different from their approach, here, we carefully design synthetic data for pretraining a model for HTR on historical manuscripts.

3   Training HTR Models on Small Datasets

In this section, we present our Leopardi dataset and detail the semi-automatic procedure we propose for building a large quantity of synthetic training data. Moreover, we outline the HTR models considered for validating our approach.

3.1   Leopardi Dataset

To favor the research towards HTR systems able to work on historical documents even in the absence of large training datasets, we devise a new dataset consisting of a small collection of early 19th Century letters written in Italian by Giacomo Leopardi. The letters are preserved at the Estense Library in Modena, and their high-resolution scans are also available at its Digital Library[2]. In particular, there are 168 pages containing text in Giacomo Leopardi's handwriting, both letter bodies and envelope fronts.

   For this study, the pages have been categorized in envelope fronts, full-page letters, and short cards (see Fig. 1). Half of the pages in each category have been used to obtain the training set (for a total of 84 pages), and the remaining pages in each category have been equally divided for obtaining the validation and test sets (both consisting of 42 pages). We manually segmented the pages at line level, with rectangular bounding boxes oriented to follow the line inclination. Each

---

[2] https://edl.beniculturali.it

line has been manually transcribed and double-checked by experts. Note that we obtained diplomatic transcriptions (*i.e.*, the text is transcribed as it appears on paper, with no editorial intervention such as abbreviated forms extensions), where stroke-out text is transcribed as "#". Some examples are reported in Fig. 2. The resulting dataset, which we call Leopardi dataset, contains 1 303 training lines, 587 validation lines, and 569 test lines. The total number of non-*blank* characters in this dataset is 77.

3.2 Collecting Synthetic Data

Ancient manuscripts preserved in digital libraries are often organized in small collections authored by a single, culturally relevant author. This is also the case of the collection from which we obtained the Leopardi dataset. To deal with the training data scarcity of this scenario, we propose to semi-automatically generate synthetic data that closely mimic the real ones and can be used for training HTR models able to adapt easily to the real set.

In particular, we isolate two to four variants of Giacomo Leopardi's most used glyphs (*i.e.*, punctuation marks, typographic symbols, characters, and their ligatures) and combine them in a randomized font built with a professional tool. Then, we collect free e-books of Giacomo Leopardi's main prose works to better capture the author's writing style and language and split their text into lines of random length ranging from 6 to 10 words, as this is the typical length of the lines in the real dataset. Finally, we type the so obtained lines with the built font at different sizes, and we superimpose on the lines a piece of paper with no writes on it, from the same collection as the letters, to simulate the typical background of ancient documents. Some examples of text in the synthetic font in comparison with real Giacomo Leopardi's handwriting are reported in Fig. 3.

Note that isolating and transcribing a text line for the real Leopardi dataset took around 120s. With our automatic procedure, we are able to generate arbitrarily big synthetic datasets in negligible time once the font is built. For this work, we obtain 111 465 text lines, 89 068 of which are used for training, and the remaining 22 397 for validation. The total number of non-*blank* characters in the synthetic dataset is 114. Furthermore, during the training with these text synthetic data, we apply random distortions to simulate variations in the shape of the handwritten characters and ink stains.

3.3 Baseline Models

Many state-of-the-art approaches for HTR work at line-level, *i.e.*, take as input an image representing a handwritten text line. The image is fed to a convolutional neural network to extract a sequence of visual feature vectors from the feature map of the last layer. In particular, the C-dimensional vectors of each of the H rows of the $H \times W \times C$ feature map are stacked to obtain a sequence of W feature vectors with $(H \cdot C)$-elements each. Each feature vector of the sequence corresponds to a region of the original image.

(a) Leopardi Real   (b) Leopardi Synth

Fig. 3. Sample text parts from the real Leopardi dataset (a) and the corresponding text in the font and style used for obtaining the synthetic dataset (b).

The sequence is fed to a recurrent neural network that outputs, for each feature vector, the probability that the corresponding image region contains one of the characters in the charset, plus a special *blank* character meaning "no other characters". Finally, a decoding block outputs the final transcription by taking the most probable character at each timestep, removing duplicate characters not separated by a *blank*, and then removing the blanks. Such architectures are usually trained to maximize the Connectionist Temporal Classifier (CTC) probability of the transcribed sequence.

In this work, we consider two variants of this pipeline, commonly adopted as backbone for HTR systems, namely the method proposed by Shi *et al.* [23] for sequence recognition and that proposed by Puigcerver [18] for HTR. The first variant has a deeper convolutional part, while the second has a deeper recurrent part. Moreover, we consider two further variants of these models that employ Deformable Convolutions (DefConvs) [8] in place of standard convolutions in the convolutional part, as proposed in [6].

Shi *et al.*, 2016 [23]. In this variant, the convolutional part has the same architecture as VGG-11 [24] up to the sixth convolutional block, plus a 7th one with a 2×2 kernel. Moreover, the receptive field of the 3rd and 4th max-pooling layers are rectangular 2×1 instead of 2×2, so that the resulting feature maps are wider and better reflect the height-width ratio of text-lines images. The recurrent part of this variant consists of a stack of two Bidirectional Long Short-Term Memory networks (BLSTMs).

Puigcerver, 2017 [18]. In this variant, the convolutional part consists of five blocks containing a convolutional layer with 3×3 kernels, a Batch Normalization layer, and a LeakyReLU activation function. The convolutional layers at the $k = 1, ..., 5$ blocks have $16k$ filters, respectively. Moreover, a 2×2 max-pooling operation is applied to the output of the first three blocks. The recurrent part of this variant consists of a stack of five BLSTMs.

Applying Deformable Convolutions [6]. The mentioned variants use standard convolutions to represent the input image. In our recent work [6], we demonstrated the suitability of DefConvs to the HTR task due to the capability of their kernel to adapt to handwritten strokes and thus better deal with character variations in shape, scale, and orientation. In particular, the kernel grid of DefConvs is deformed depending on the processed input thanks to a set of learnable trans-

lation vectors, each applied to a separate element of the kernel grid. In this work, we further test the performance of DefConvs in the HTR task when trained on small historical datasets. For this reason, we replace the standard convolutional layers in the Shi *et al.* [23] and Puigcerver [18] variants with DefConvs layers. A DefConv layer is obtained by concatenating a standard convolutional layer in charge of learning the translation vectors and another convolutional layer in charge of learning the kernel weights.

4   Experimental Evaluation

In this section, we evaluate the performance of the considered HTR models when pretrained on the semi-automatically obtained synthetic data and then applied or fine-tuned on the presented Leopardi dataset. We compare our approach to pretraining on different datasets, both synthetic and real, and to training from scratch on the Leopardi dataset. The results are expressed in terms of the commonly used Character Error Rate (CER) and Word Error Rate (WER) scores.

4.1   Implementation Details

In the considered models, the output of the convolutional part is a $2 \times W \times 512$ tensor for the variants based on [23] and a $16 \times W \times 80$ tensor for those based on [18]. These are collapsed in a sequence of W vectors of $1\,024$ and $1\,280$ elements, respectively. The BLSTMs have 512 hidden units each in the Shi *et al.* [23] variants and 256 in the Puigcerver [18] variants. The recurrent layers are separated by a dropout layer with dropout probability equal to 0.5.

When training from scratch or fine-tuning, we normalize the text line images between -1 and 1 and rescale them in height, keeping the original aspect ratio. In particular, the rescaled height is 60 pixels for the variants based on Shi *et al.* [23], 128 for those based on Puigcerver [18]. When pretraining, additionally to these pre-processing steps, we randomly alter the image brightness (with factor chosen between 0.5 and 5), its contrast (with factor chosen between 0.1 and 10), its saturation (with factor chosen between 0 and 5), and its hue (with factor chosen between -0.1 and 0.1), and apply Gaussian blur with size 5 and standard deviation randomly chosen between 0.1 and 2. Moreover, we randomly apply one of the following geometric distortions: random rotation between $-1°$ and $1°$, affine transformation randomly rotating the image between $-1°$ and $1°$ and randomly shearing it between $-50°$ and $30°$, or random homography.

When pretraining the models, we use batch size 16, while for fine-tuning and training from scratch, the batch size is 8. We use a learning rate of 10^{-4} in all experiments except for training from scratch the variants based on [18], where we use a learning rate equal to $3 \cdot 10^{-4}$. The proposed models have been trained using Adam as optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, in combination with a scheduler that reduces the learning rate by a factor 0.1 when the CER on the validation set reaches a plateau. We train the models until the CER on the validation set stops improving for 20 epochs in case of pretraining, 80 epochs in case of fine-tuning and training from scratch.

Table 1. Performance of the considered models when pretrained on different real and synthetic datasets and directly applied to the Leopardi dataset test set.

	[23]		[18]		[23]+DefConv[6]		[18]+DefConv[6]	
	CER	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**	**WER**
ICFHR14	47.2	102.7	59.7	120.1	77.8	112.4	103.9	147.1
ICFHR16	76.0	129.8	79.1	111.0	83.1	109.3	86.9	144.6
IAM	46.5	92.9	68.0	96.4	60.4	97.5	74.4	98.9
RIMES	43.4	88.0	73.9	103.9	72.8	100.0	69.5	97.1
Modern Leopardi Synth	59.2	100.9	65.6	101.8	61.5	100.5	64.1	101.6
Leopardi Synth	**35.9**	**86.4**	**38.5**	**93.9**	**40.2**	**92.1**	**36.3**	**94.0**

4.2   Experimental Results

First, we compare the performance of the four presented HTR models when pretrained on a big HTR dataset (source) and then directly applied to the test set of the devised Leopardi dataset (target). We consider both real benchmark datasets and synthetic datasets. In particular, the real datasets are the modern IAM and RIMES datasets and the historical ICFHR14 and ICFHR16 datasets, introduced in Sec. 2. The synthetic datasets are Leopardi Synth and Modern Leopardi Synth. Both are obtained by following the procedure explained in Sec. 3.2, the former by using the font inspired by Giacomo Leopardi's handwriting, the latter by using three freely available modern handwriting fonts. Note that the text, background, number of lines, and random transformations applied during training are the same for both these synthetic datasets. The results are reported in Table 1 and clearly show that directly applying models trained on the Leopardi Synth data allows reaching the lowest CER and WER on the real Leopardi dataset.

Further, we compare the performance of the considered models pretrained on the Leopardi Synth dataset and then fine-tuned on the 100% and 50% of the lines in the real Leopardi training, with the same models trained from scratch on the same quantity of real Leopardi training lines. The results are reported in Table 2, where we also report the results of the direct application of the models pretrained on Leopardi Synth for comparison. The benefits of pretraining are more evident in the case of fine-tuning on 50% of the training lines. In fact, in this case, the CER decreases by 1.8 and the WER of 4.1 on average, while in the case of fine-tuning on 100% of the training lines, the CER decreases by 0.8 and the WER of 3.1 on average.

5   Conclusion

In this paper, we explored line-level HTR on historical manuscripts when limited training data are available. To this end, we devised a newly collected dataset taken from letters authored by the Italian writer Giacomo Leopardi, which poses the typical challenges of small collections of handwritten historical documents. To deal with this scenario, we propose pretraining on a large quantity of synthetic data that reflect the real target manuscripts, which we built with a semi-automatic procedure, and fine-tuning on a portion of real data. The obtained

Table 2. Experimental results of the considered models when pretrained on the Leopardi Synth dataset and then fine-tuned on different portions of the real Leopardi dataset, compared to their performance when trained from scratch on the same lines of the real Leopardi dataset.

	[23]		[18]		[23]+DefConv[6]		[18]+DefConv[6]	
	CER	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**	**WER**
Real Training Data: 0%								
Leopardi Synth	35.9	86.4	38.5	93.9	40.2	92.1	36.3	94.0
Real Training Data: 50%								
Leopardi Real	7.2	27.2	8.2	27.8	6.7	24.6	8.0	28.1
Leopardi Synth	**5.7**	**22.1**	**5.7**	**21.6**	**6.1**	**23.0**	**5.3**	**20.3**
Real Training Data: 100%								
Leopardi Real	4.3	17.2	5.6	20.5	3.4	13.3	4.1	15.4
Leopardi Synth	**3.8**	**14.3**	**3.8**	**14.3**	**3.2**	**12.7**	**3.3**	**12.8**

experimental results demonstrate the suitability of the proposed approach for several HTR models, both in a direct transfer learning and a pretraining and fine-tuning scenario.

References

1. Aradillas, J.C., Murillo-Fuentes, J.J., Olmos, P.M.: Boosting offline handwritten text recognition in historical documents with few labeled lines. arXiv preprint arXiv:2012.02544 (2020)
2. Augustin, E., Carré, M., Grosicki, E., Brodin, J.M., Geoffrois, E., Prêteux, F.: RIMES evaluation campaign for handwritten mail processing. In: IWFHR (2006)
3. Baraldi, L., Cornia, M., Grana, C., Cucchiara, R.: Aligning text and document illustrations: towards visually explainable digital humanities. In: ICPR (2018)
4. Causer, T., Wallace, V.: Building a volunteer community: results and findings from transcribe bentham. Digital Humanities Quarterly **6** (2012)
5. Chammas, E., Mokbel, C., Likforman-Sulem, L.: Handwriting recognition of historical documents with few labeled data. In: DAS (2018)
6. Cojocaru, I., Cascianelli, S., Baraldi, L., Corsini, M., Cucchiara, R.: Watch Your Strokes: Improving Handwritten Text Recognition with Deformable Convolutions. In: ICPR (2020)
7. Cornia, M., Stefanini, M., Baraldi, L., Corsini, M., Cucchiara, R.: Explaining digital humanities by aligning images and textual descriptions. Pattern Recognition Letters **129**, 166–172 (2020)
8. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: CVPR (2017)

9. Fischer, A., Frinken, V., Fornés, A., Bunke, H.: Transcription alignment of Latin manuscripts using hidden Markov models. In: Proc. of the Workshop on Hist. Doc. Imag. and Proc. (2011)
10. Fischer, A., Keller, A., Frinken, V., Bunke, H.: Lexicon-free handwritten word spotting using character HMMs. Pattern Recognition Letters **33**(7), 934–942 (2012)
11. Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., Stolz, M.: Automatic transcription of handwritten medieval documents. In: VSMM (2009)
12. Granet, A., Morin, E., Mouchère, H., Quiniou, S., Viard-Gaudin, C.: Transfer learning for handwriting recognition on historical documents. In: ICPRAM (2018)
13. Jaramillo, J.C.A., Murillo-Fuentes, J.J., Olmos, P.M.: Boosting handwriting text recognition in small databases with transfer learning. In: ICFHR (2018)
14. Johansson, S., Leech, G.N., Goodluck, H.: Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computer. Department of English, University of Oslo (1978)
15. Kang, L., Riba, P., Rusiñol, M., Fornés, A., Villegas, M.: Pay Attention to What You Read: Non-recurrent Handwritten Text-Line Recognition. arXiv preprint arXiv:2005.13044 (2020)
16. Marti, U.V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. IJDAR **5**(1), 39–46 (2002)
17. Poznanski, A., Wolf, L.: CNN-N-Gram for Handwriting Word Recognition. In: CVPR (2016)
18. Puigcerver, J.: Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition? In: ICDAR (2017)
19. Romero, V., Fornés, A., Serrano, N., Sánchez, J.A., Toselli, A.H., Frinken, V., Vidal, E., Lladós, J.: The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognition **46**(6), 1658–1669 (2013)
20. Sánchez, J.A., Romero, V., Toselli, A.H., Vidal, E.: ICFHR2014 competition on handwritten text recognition on transcriptorium datasets (HTRtS). In: ICFHR (2014)
21. Sanchez, J.A., Romero, V., Toselli, A.H., Vidal, E.: ICFHR2016 competition on handwritten text recognition on the READ dataset. In: ICFHR (2016)
22. Shen, X., Messina, R.: A method of synthesizing handwritten chinese images for data augmentation. In: ICFHR (2016)
23. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. PAMI **39**(11), 2298–2304 (2016)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
25. Soullard, Y., Swaileh, W., Tranouez, P., Paquet, T., Chatelain, C.: Improving text recognition using optical and language model writer adaptation. In: ICDAR (2019)
26. Voigtlaender, P., Doetsch, P., Ney, H.: Handwriting recognition with large multidimensional long short-term memory recurrent neural networks. In: ICFHR (2016)
27. Wigington, C., Stewart, S., Davis, B., Barrett, B., Price, B., Cohen, S.: Data augmentation for recognition of handwritten words and lines using a CNN-LSTM network. In: ICDAR (2017)