A Scalable Handwritten Text Recognition System

R. Reeve Ingle, Yasuhisa Fujii, Thomas Deselaers, Jonathan Baccash, Ashok C. Popat

Google Research

Mountain View, CA 94043, USA

Email: {reeveingle,yasuhisaf,deselaers,jbaccash,popat}@google.com

Abstract—Many studies on (Offline) Handwritten Text Recognition (HTR) systems have focused on building state-of-the-art models for line recognition on small corpora. However, adding HTR capability to a large scale multilingual OCR system poses new challenges. This paper addresses three problems in building such systems: data, efficiency, and integration. Firstly, one of the biggest challenges is obtaining sufficient amounts of high quality training data. We address the problem by using online handwriting data collected for a large scale production online handwriting recognition system. We describe our image data generation pipeline and study how online data can be used to build HTR models. We show that the data improve the models significantly under the condition where only a small number of real images is available, which is usually the case for HTR models. It enables us to support a new script at substantially lower cost. Secondly, we propose a line recognition model based on neural networks without recurrent connections. The model achieves a comparable accuracy with LSTM-based models while allowing for better parallelism in training and inference. Finally, we present a simple way to integrate HTR models into an OCR system. These constitute a solution to bring HTR capability into a large scale OCR system.

I. Introduction

Handwritten text can be found in many types of images: handwritten notes, memos, whiteboards, medical records, historical documents, text input by stylus, etc. Therefore, a complete OCR solution has to include support for recognizing handwritten text in images. This emphasizes the need for research into the area of building large scale handwriting recognition systems for many languages and scripts.

Many of the recent offline handwritten text recognition (HTR) systems have adopted line-level recognition strategies that use a combination of convolutional neural networks (CNN) and Long Short-Term Memory (LSTM) [1] recurrent neural networks for feature extraction, trained with Connectionst Temporal Classification (CTC [2]) loss [3]–[9]. Similarly to other fields, these data-driven deep-learning approaches learn features directly from the training data, in contrast to traditional methods that employ hand-engineered features. While these methods have resulted in substantial improvements in handwriting recognition accuracy on public benchmarks [5], [6], [8], scaling these approaches to support new domains or new languages can be challenging due to the cost and difficulty associated with collecting and labeling a large corpus of handwritten training data of line images.

We believe that the biggest challenge in developing an HTR system is not modelling but obtaining sufficient amounts of high quality training data. In this paper we address this problem by leveraging a large amount of online handwriting data and only few images of handwriting: We use the *ink* data from our online handwriting recognition system [10], [11] to train the line recognizer in our existing OCR system [12]. An *ink* is the sequence of strokes of (x, y) coordinates with timestamps of a user writing with their finger (or a stylus) on a screen. Online handwriting recognition is important in the context of entering text into computing devices in a world where large parts of humanity gain access to computing devices through mobile devices. Many of these users are using languages written in scripts that are not as easy to type as English [13].

To the best of our knowledge, this is the first work that leverages the relationship between online and offline handwriting recognition by re-using an existing dataset of online handwriting samples to build a full HTR system. Our HTR system is trained on images that are rendered from the trajectory data of the online handwriting recognition system. To achieve sufficient accuracy on real images of handwriting, we have adopted an image degradation approach that is commonly used to generate realistic synthetic training images for OCR systems [14]. After rendering clean handwriting images from trajectory data, the images are processed with an image degradation pipeline that applies realistic image transformations to the data. In addition to the collected online handwriting data we also use a handwriting synthesis pipeline to enrich the variability in our training dataset to obtain better results. We are demonstrating the feasibility of this approach on Latin script, but we believe that this approach will work for other scripts as well.

For the line recognizer we experiment with two different model architectures: We describe an LSTM-based architecture similar to many of the state-of-the-art methods. However, a problem with recurrent architectures is that they do not train and run as easily on specialized hardware (GPUs, TPUs) as feed forward networks. Thus we also propose a fully feed-forward network architecture that achieves comparable accuracy as our LSTM-based model.

While the line recognition engine is often the primary focus of HTR research, it is just one component of a full text recognition system. We describe the additional steps that were taken to integrate HTR support into an end-to-end text recgognition system consisting of text detection, direction identification, script identification, and text line recognition models [15].

The remainder of this paper is structured as follows:

17

Section II describes our line recognition model. Section III describes how we use online handwriting data to generate training images suitable for training an offline recognizer including the rendering process, the image degradation pipeline, and our handwriting synthesis system. In section IV we describe how the line recognizer is integrated into our full OCR system. Sections V and VI describe and discuss our experimental evaluation and section VII concludes the paper.

II. HANDWRITTEN TEXT LINE RECOGNITION

A. Model

The task of handwritten text line recognition is to produce a sequence of Unicode code points for the handwritten text in a single line image. Let x be a line image and y be a sequence of Unicode points. We consider a probabilistic model $P(y|x)$ to represent the relationship between x and y.

We follow the approach described in [15] to model $P(y|x)$: We scale the height of the input image to a fixed size of 40 pixels and collapse the color space down to luminance (gray-scale). The processed image is then fed into a neural network to produce a 1-D sequence of logits where each logit corresponds to a single character or a special blank symbol. The network is trained using a CTC loss [2]. The model naturally handles a variable width (length) of image. At inference time, the logits cost is combined with the cost of a character-based n-gram language model in a log-linear way and beam search is used to find the best recognition result.

We experiment with two different model architectures:

1) LSTM-based model: LSTMs and other recurrent neural networks are dominantly used for modern handwritten text line recognition models [3], [5]–[8], [16]–[18]. Our model is inspired by the CLDNN (Convolutions, LSTMs, Deep Neural Network) architecture proposed by [19]. For the convolutional layers, we use the inception [20] style architecture described in [15]. For the LSTM layers, we use between one and four stacked bidirectional LSTMs (BLSTMs). More details are given in Section V.

2) Recurrence-free model: One disadvantage of recurrent models is that they are inherently sequential and cannot take full advantage of the parallelism offered by accelerators such as GPUs and TPUs. Therefore, network architectures having only feed-forward connections are preferred over those with recurrent connections for a large scale production system when comparable accuracy is achieved. In order to simulate the recurrency and gating mechanism of LSTMs with a feed-forward structure, we follow an idea from [21] and use 1-D gated recurrent convolutional layers (GRCLs)[1] as a fully-feed-forward alternative to LSTMs (see Figure 1). Briefly speaking, GRCL is a recurrent convolutional network [22] with gating mechanism. The recurrency is along depth, and information propagation is controlled by the gating mechanism, which is computed using a sigmoid activation that is multiplied by the original ReLU activation. Although the recurrency is

along depth, information can be propagated along time with a large receptive field. [21] uses GRCLs as an alternative to CNNs, and the GRCLs are still followed by BLSTMs. We use GRCLs as an alternative to LSTMs, and the CNN layers are not changed. In this way, the model has only feed-forward connections. We show that the combination yields comparable accuracy with LSTM-based systems.[2]

B. Training Data

The model needs line images (x) along with their transcriptions (y) for training. One way to get such data is to manually annotate handwritten text lines in images collected from various sources. This provides the most valuable data. However, it is time-consuming and costly to do the annotation. Moreover, it is not trivial to find a sufficiently large amount of images containing handwritten text. Thus, the availability of manually labeled data tends to be limited especially when the goal is to support many languages.

In the following section we describe how we leverage a large amount of stroke data that was collected to build an online-handwriting recognition system [10], [11]. This method enables us to support a new script at substantially lower cost.

III. DATA SYNTHESIS, RENDERING, AND DEGRADATION PIPELINE

We use a mix of different sources of data to obtain the best possible HTR system: We have access to a *large amount of ink data* collected for building an online handwriting recognition system [10], [11] in many languages. For the purpose of this paper we are only looking at Latin script but are planning to expand our system to other scripts and languages. We further are using an *online handwriting synthesis pipeline* to enrich this data with handwriting styles that are not well represented in the original corpus. This data is rendered into images using a rendering pipeline described below and then degraded using the same degrading pipeline that we are using to train an OCR system on synthetic typeset data. To improve recognition accuracy on typeset text, the training data also includes degraded synthetic typeset data. Additionally, we include historic image data from several public datasets to improve recognition accuracy on historical handwriting [23]–[26]. Finally, we have a small amount of image data with modern handwriting labeled at the line level.

Table I gives an overview of the amount of data of each type that is used to train our final system and Figure 2 shows samples of our training data. Figure 3 shows some samples of our labeled real image data.

A. Rendering Images from Stroke Data

The ink data that was used to train this system has been obtained through a variety of sources including prompted data collection of individual words on mobile phones and logged

---

[1]The term "recurrent" in GRCL is used to mean recurrency along depth. The recurrency we want to avoid here is recurrency along time.

[2]We can also consider CNN-only models to design feed-forward models as in [15]. However, we observed that such models performed significantly worse than LSTM-based models on handwritten text and did not explore those models further.
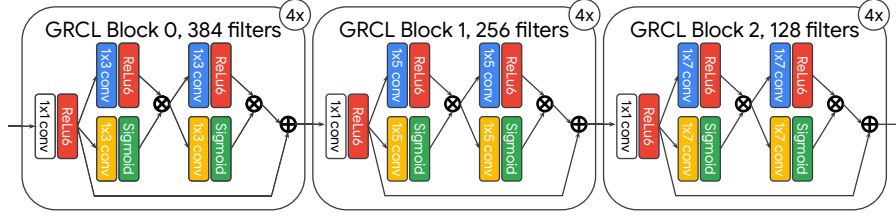
18

Fig. 1. Gated recurrent convolutional layer. Each GRCL block contains two convolutional layers with ReLU6 activations, with shared weights. The output of these layers are multiplied element-wise by the output of convolutional layers with sigmoid activations, which serve as a learned gating mechanism. Convolutional weights for the gates are also shared within each GRCL block.

label	rendered	degraded
important		
because		
down upon		
Zagrebu And fro		
and Two Viola Ground you		
language of the law remains		
Academy of Design		
PARADOXY A KRITIKA		
o distribuie la microprocesor,		

Fig. 2. Examples of label, rendered ink, and degraded image. The first 3 rows are samples that are rendered from the training data of the online handwriting recognition system (sec. III-A). The fourth and fifth rows are examples of a long line created by concatenating multiple short inks (sec. III-A). The sixth and seventh rows are synthetic inks (sec. III-B). The final two rows are synthetic typeset text lines. The last column shows some typical degradations (sec. III-C).

TABLE I. Number of training samples from different sources in our final system.

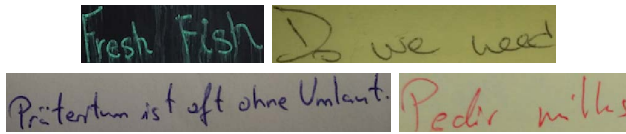| Type of data	# training samples	# test samples
Online data	1,115,822	34,895
Long lines data	500,000	0
Synthetic online data	1,033,680	0
Synthetic typeset data	500,000	0
Labeled image data	62,477	8,362
Historic image data	159,326	0

Fig. 3. Example line images from our collected real handwriting image set.

traffic of users of a handwriting recognition web service [10], [11]. Most of the data was obtained with the goal of building a handwritten text input method for mobile phones and thus most of the written items are individual words, partial words, or even individual letters. In most cases the writing is in a single line, but in some cases the data was written over multiple lines, e.g. when users had no space to finish a word at the edge of the screen.

To obtain data feasible to train an OCR line recognizer, we first processed the entire dataset using forced alignment to get character segmentation using our segment-and-decode online handwriting recognition system [10], [11]. Then, we apply a simple heuristic to split multi-line inks into single lines: for each character, we look at the bounding box – if the highest point of a new character is at most 20% overlapping with the existing line, it is considered to be the start of a new line. After this step all inks are single-line inks.

Since most of the inks are short, we supplement these with longer lines that are created by randomly combining individual inks into longer lines. For this we concatenate the label (with a space) and arrange the corresponding inks with a randomly chosen amount of space in between [27]. The target line lengths are distributed roughly uniformly between 500 and 2500, at a height of 100.

Finally we render the inks into images using a randomly chosen stroke width, a randomly chosen slant, and a randomly chosen boundary size with black pixels on white background at an image height of 100 pixels. Figure 2 shows some examples of this data without and with concatenated long lines.

B. Synthesis of Handwriting Stroke Data

From inspecting our data, we observed that the training data from the online handwriting recognition system has two

19

deficiencies. 1) Most of the written items are very short. We are addressing this by combining individual items into longer lines, however the resulting data, while long enough, does not create a realistic text sample, nor does it have a consistent writing style within the line. 2) Most of the written items have well-separated characters (rather than a cursive style). This is due to most of the data coming from users entering text on mobile phones - where users write pragmatically. For handwriting recognition in images, we need to recognize a wider variety of styles, including print, cursive, and non-natural writing styles, e.g. a restaurant menu written with neat, stylized handwriting.

To alleviate these two problems we employ a *handwriting synthesis pipeline*. This pipeline was trained on the same training data, but it allows for controlling the "*neatness*" and the writing style.

Our handwriting stroke synthesis pipeline is based on the system described in [28]. In addition to the original system, we have added a style embedding module to the network. The style embedding module consists of two LSTM layers that take as input the output from the original synthesis model. We compute the sum of the output of the style-embedding LSTM as a style embedding and pass that as an additional input to the mixture density output layer of the original synthesis model. This style embedding captures the writing style because input/output are independent of the label. The style embedding is trained as part of the normal training procedure without any extra loss function.

In order to synthesize handwriting with the style of a given sample, we compute the style embedding for that sample (by passing it through the style embedding directly) and then feed the fixed style vector into the mixture density layer when synthesizing. We use this method to synthesize additional cursive data by using the style embeddings from a small number of handpicked cursive writing samples.

In order to obtain meaningful text samples, we synthesize the same text corpus as we use in the synthetic data pipeline of our OCR system.

Finally, in order to create neat handwriting samples, we reduce the sampling temperature in the mixture density network layer - effectively sampling very close to the mean handwriting for a given style. This typically leads to visually appealing samples with a high regularity - as you expect when someone is trying to write neatly.

See Figure 2 for some examples of synthetic handwriting samples.

C. Image Degradation Pipeline

In order to make the images of rendered ink appear realistic, we apply our image degradation pipeline that we also use to degrade synthetic OCR training data. The degradation pipeline is a multi-step system where different degradation steps are performed in order. Most degradation steps use randomness to vary how they transform the input image, e.g. a "background" degrader has access to a large set of images, picks one of these at random and then superimposes the ink image

onto this image, and a "transform" degrader would pick a transformation (e.g. scale, rotate) at random and then pick the parameters for the transformation at random. The steps of our degradation pipeline are shown in Figure 4. Figure 2 shows some images before and after the degradation process.

IV. INTEGRATION INTO THE OCR SYSTEM

Line recognition is a component of a full text recognition system. We briefly describe our line-based OCR system and propose a simple way to integrate the handwritten text line recognition into the full system.

A. Line-based Multilingual OCR System

Figure 5 shows the end-to-end pipeline of our line-based multilingual OCR system. The first step extracts text lines from the image. Next, direction identification determines the writing direction of each line and rotates the line images to make the characters in the images upright. Once the directions are fixed, script identification classifies the lines into scripts. Text recognition selects a model based on the script identification result and runs the model to obtain a sequence of Unicode points for each line. Finally, layout postprocessing infers the structure of the text (e.g. paragraph). The recognition results may be enhanced during the process, for example, by filtering results with low confidence values and by removing overlapping lines. Up to the layout postprocessing step, lines are the fundamental processing unit. More information on the system can be found in [12].

B. Dual-Head Script and Style Identification

Script identification is the third step in our pipeline and is responsible for selecting the appropriate recognition model for each line. In order to integrate the handwriting line recognizer, we extended our model to allow for identifying a line as handwritten or printed which then selects the appropriate model.

We adopt the *sequence-to-label* paradigm proposed in [15] to implement the script identification model. The model performs a single classification for the identification. We modify the model by adding an extra classification head to perform the style classification between printed and handwritten text. This approach is more effective than the naive approach where each script has has a 'printed' and a 'handwritten' sub-class, which doubles the number of classes in the model.

Figure 6 shows the proposed dual-head script and style identification model. The script identification head selects a script and the style identification head selects a style. The handwritten text model is selected if the probability of handwritten text is greater than or equal to a predefined threshold θ^{hwr}.

V. EXPERIMENTAL SETUP

A. Data

We have been using data from three different datasets for our experiments: The IAM offline handwriting database, the
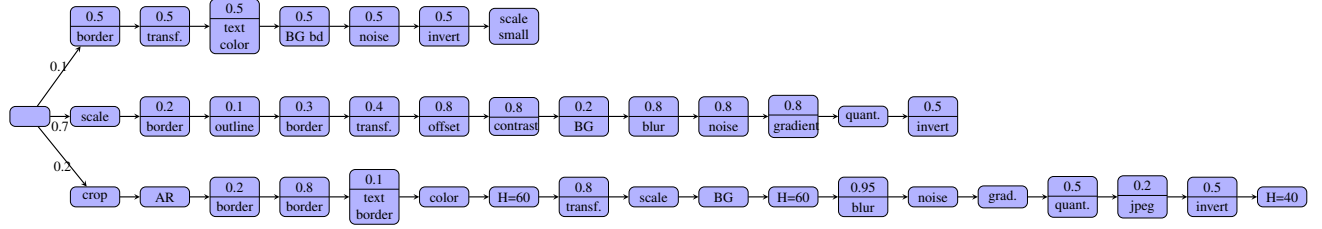
20

Fig. 4. A flow diagram describing the degradation system. Each box is a degradation step, numbers on arrows indicate probabilities that a path is chosen, numbers in the top part of a degradation step indicate probability that the node step is executed. Degradation steps are: border: adds a border to an image; outline: adds a line around an image; transf.: performs a spline, rotation, or projective transformation; BG: puts the text onto a background image; scale: scales the image between 50% and 150%; offset: applies an offset to the intensity values of all pixels in the image; contrast: changes the contrast of the image; blur: applies a blur filter to the image; noise: adds additive noise to the image; grad.: puts a gradient background behind the text; quant.: quantizes the intensities in the image; invert: inverts the image; crop: crop up to 20% of the image; AR: randomly scales - changing aspect ratio; text color: changes the intensity of the text with respect to the background; jpeg: adds jpeg compression artifacts to the image; H=X: scales the image to be of height X pixels; text border: adds text fragments to the border of the image - mimicking bad segmentation.
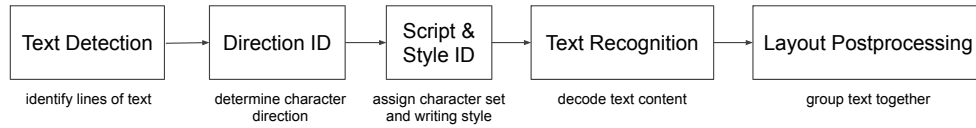

Fig. 5. Multilingual HTR and OCR system with line-level script identification and writing style classification.
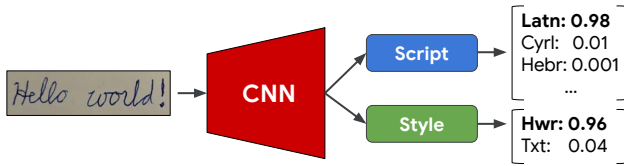

Fig. 6. Dual-head script and style classifier.

IAM online handwriting database, and an internal dataset of both online and offline handwriting samples.

The IAM offline handwriting database [29] (IAM-Off) contains scanned pages of handwritten text passages, which were written by 500 different writers using prompts from the Lancaster-Oslo/Bergen (LOB) text corpus. The database contains line images that have been partitioned into writer-disjunct training, validation, and test sets containing 6161, 976, and 2915 lines, respectively.

The IAM online handwriting database [30] (IAM-On) contains trajectory data of handwritten text written by 221 different writers, acquired on a digital whiteboard. The database is partitioned into training, validation, and test sets containing 5363, 1438, and 3859 lines, respectively. Text prompts for the IAM online handwriting database were also drawn from the LOB corpus. In contrast to the IAM offline database, the partitioning into train, validation, and test sets did not ensure that the train/valid/test sets use a disjunct set of prompts. For our experiments, we have rendered the IAM online database using the method described in section III-A. We also created an augmented version of this dataset (IAM-On-Long) con-

taining 500'000 longer lines for training and 20'000 lines for validation by randomly concatenating samples as described in section III-A

Our internal dataset consists of a large dataset of online handwriting samples [10], [11], long lines obtained by concatenating samples from this dataset (see sec. III-A), and synthetic samples generated from a model trained on this dataset (see sec. III-B). In addition we have a smaller set of scanned and photographed handwriting images for training and evaluation. The sizes of these individual datases are given in Table I and Figure 2 shows some samples for each of these datasets.

For the full system experiments, two datasets are used: printed text images and handwritten text images. The printed text images come from the Web and consist of 508 images with 4868 lines. The handwritten text images are a subset of the internal dataset, explained above, having full image annotations. The internal dataset contains some data that consists solely of line images, and these are excluded from the full system experiments. The data consists of 433 images with 2555 lines.

B. Models

All neural network models were implemented with TensorFlow [31]. A standard asynchronous stochastic gradient descent with multiple workers was used to train all models. The learning rate was decayed exponentially through the training steps. The hyper parameters of the training were tuned on the development set if available; otherwise on a holdout set of the training data.

We trained two types of handwritten text recognition models. One uses BLSTM layers and the other uses GRCL blocks described in Section II-A2. For the former, we trained models with 1, 2, and 4 BLSTM layers. For the latter, we trained a model with the topology shown in Figure 1. The full model architecture is given in Appendix A. The models were combined with a character-based n-gram language model trained on text collected from the Web and Wikipedia. We used $n = 9$.

We implemented the dual-head model for script and style classifications by adding an extra summarizer to the shared inception-style encoder. The full model architecture is given in Appendix B. We incorporated the handwritten recognition model into the multilingual OCR system described in [15] by using the dual-head script and style identifier. We treated the data used for [15] as the printed text data to train the style identifier. The OCR (printed text) model was trained with the printed text data used in [15] with a small amount of additional handwritten text data. Thus, the OCR model should be able to recognize handwritten text to some extent. However, as we observe in the experimental results, the handwritten-specific model is superior to the printed text model on handwritten text. We expect that the dual-head identifier provides better results than either of these models in isolation.

C. Evaluation

For line recognition experiments, we use character error rate (CER) and word error rate (WER) as error metrics. CER is defined as a ratio of the edit distance between the target and the reference character sequences to the number of characters in the reference. WER is defined analogously on words.

For the full system experiments, we use word recall and precision as error metrics. In order to consider the positional information in the metrics, we compute IoU between word pairs and ignore ones whose IoU are less than 0.01. We use the small number for the threshold to make the score less sensitive to the errors of bounding boxes in groundtruth. If a word in the recognition result does not overlap with any words in the reference, the word is considered as a false positive. If a word in the reference does not match with any words in the recognition result, the word is considered as a false negative. AUC (Area Under a Curve) is computed with recall and precision by changing the threshold to remove words whose confidence value is lower than the threshold.

VI. RESULTS AND DISCUSSIONS

A. Line Recognition

Table II shows line-level WER and CER of our GRCL-based and our LSTM-based system on the IAM offline test data set while varying the training data sets. As training data we used the IAM Offline training dataset, the IAM online training dataset with and without long-lines, rendered as described in section III-A and also with no degradations (ND), as well as a combination of these. No additional sources of training data were used when training these models.

TABLE II. Line-level results of the GRCL and two-layer BLSTM system on the IAM-Offline test data using different training sets.

		GRCL		BLSTM	
Training Data	# tr. obs.	WER	CER	WER	CER
IAM-Off	6'161	35.2	14.1	30.7	12.8
IAM-On	5'363	27.8	12.0	21.7	9.1
IAM-On-Long	505'363	24.8	10.5	18.3	7.4
IAM-Off + IAM-On-Long-ND	511'524	22.3	8.8	23.0	9.2
IAM-Off + IAM-On-Long	511'524	17.0	6.7	15.1	5.7

TABLE III. Line-level results of our system trained on internal data and evaluated on internal data and the IAM offline testdata.

		Off-Test		Internal	
Model	#params	WER	CER	WER	CER
1 BLSTM	3'913'065	16.4	6.1	32.5	13.4
2 BLSTM	4'307'305	14.9	5.5	**30.1**	**12.4**
4 BLSTM	5'095'785	15.7	5.8	32.1	13.1
GRCL	10'614'889	14.1	5.4	30.8	12.6
GRCL + tuning		**10.8**	**4.0**	38.6	15.4
Bluche et al. [7]		10.5	3.2		
Voigtlander et al. [6]		9.3	3.5		
Poznanski et al. [32]		6.4	3.4		
Castro et al. [9]		10.5	3.6		

The most surprising finding was that using the online dataset for training we obtain better results than when using the offline training data. The reason for this is that the *effective* size of the online training set is larger than the 5'363 samples because we add random degradations during training (see section III-C), which we do not do on the natural images in the IAM offline dataset. These random degradations facilitate training and help to mitigate overfitting. Further, we found that it helps substantially to add the concatenated long lines into the training set, and that a combination of both datasets, with random degradations applied to the IAM online data, leads to the best results. This trend is identical for both the GRCL and the LSTM model.

Table III shows results on the IAM Offline test data and on our internal dataset for a few different variants of our full line recognition system trained on internal data. The first four rows were trained on the entire training data as described in Table I and the row "GRCL + tuning" is the same system as the GRCL system but tuned to perform well on the IAM dataset by 1) adding the IAM offline training data 50 times to the training data; 2) using a language model trained on the LOB corpus; 3) tuning the decoder weights on the IAM validation set.

For the internal dataset, we observe that the LSTM model with 2 layers performs best, larger models become worse, probably because of overfitting. On the IAM test set, the GRCL model performs a bit better than the best LSTM model, probably also because of less overfitting. This is surprising given that the number of parameters of GRCL model is as twice as the biggest BLSTM model (10'614'889 vs. 5'095'785). GRCL is less prone to overfitting than LSTM. Tuning the GRCL model for the IAM dataset leads to a substantial improvement of 25% relative from 5.4% to 4.0% CER - which is close to the state of the art.

Table IV shows the impact of using the different parts of our

22

TABLE IV. Experiments evaluating the impact of different training datasets on the recognition accuracy using a GRCL-based system (compare line GRCL in table III). The columns indicate which of the training datasets were used: (1) labeled image data, (2) rendered online data including long lines data, (3) synthetic typeset data, (4) synthetic online data, (5) historic data.

	Training dataset used					IAM-Off-Test		Internal	
	(1)	(2)	(3)	(4)	(5)	WER	CER	WER	CER
A	+					45.7	19.5	58.8	27.2
B		+				19.2	7.0	42.1	18.6
C			+			51.8	25.4	54.6	27.0
D		+		+		20.9	7.7	42.5	18.3
E	+	+		+		18.5	6.6	36.5	15.1
F		+	+	+		18.6	6.8	35.8	14.7
G	+	+	+	+		17.5	6.3	33.7	13.7
H	+	+	+	+	+	14.1	5.4	30.8	12.6

TABLE V. Word precision and recall of full text recognition system.

	Printed			Handwritten		
Model	Prec.	Recall	AUC	Prec.	Recall	AUC
OCR	**0.705**	**0.702**	0.594	0.562	0.596	0.449
HTR	0.678	0.645	0.562	0.584	0.602	**0.476**
Dual-head	**0.705**	0.701	**0.595**	**0.586**	**0.612**	0.474

training datasets on the error rates on the IAM Offline test data and on our internal dataset. The first interesting result is the difference between lines A and B where we found that using the large dataset of rendered online handwriting data (line B) leads to much better results than training on a (much smaller) dataset of real handwriting images (line A). When training the recognizer only on (synthetic) printed OCR training data, results are very poor (line C).

Another interesting insight is that adding the dataset of rendered synthetic handwriting data to the handwriting data alone does not lead to an improvement but rather to a slightly worse result (line D). Adding the real handwriting data (line E) or the synthetic printed OCR data (line F) to this experiment gives a clear improvement in both cases and very similar results. Finally, line G, shows the yet again improved result obtained from combining all of the previous data sources and line (H) corresponds to additionally using historic data as training data - which leads to a surprisingly large improvement on both the IAM Offline test data and on our internal dataset although neither of these datasets contain any historic samples.

B. Integration to the OCR System

Table V shows results for the full text recognition system with and without the dual-head script and style identifier. We used $\theta^{hwr} = 0.95$ in the experiments. As expected, we observe that the OCR model outperforms the HTR model on printed text while the HTR model outperforms the OCR model on handwritten text. The dual-head system yields comparable accuracy with the OCR model and the HTR model on printed text and handwritten text, respectively. 98.9% of the lines from the printed text images were classified as printed text by the style identifier, and 59.6% of the lines from the handwritten text images were classified as handwritten. The low handwriting classification accuracy is due to the high classification threshold ($\theta^{hwr} = 0.95$) as well as the fact that handwritten images frequently contain some lines of printed text.

VII. CONCLUSION

We described three techniques to build a scalable handwritten text recognition system.

We explored the use of online handwriting data to train handwritten text recognition models. We described the data generation pipeline and presented a series of techniques to generate better data. The experimental results showed that models can be improved by using the generated data, and we obtained the best model by combining the generated data and small amount of real images.

When a large mount of data is available, the model with GRCL blocks was comparable with the model with LSTM layers. GRCL blocks are a good alternative to LSTM layers in a situation where feed-forward networks are preferred over recurrent networks for performance reasons.

We integrated the handwritten recognition model into the full text recognition system by augmenting the script identification model with an additional classification between printed text and handwritten text. The dual-head model yielded the best accuracy on both printed and handwritten text compared to printed and handwritten-specific models.

We applied these techniques only to Latin script. It is future work to apply them to other scripts and confirm the effectiveness of the approach on more scripts.

REFERENCES

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[2] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.
[3] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *NIPS*, 2009.
[4] A. Graves, M. Liwicki, S. Fernndez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *PAMI*, vol. 31, no. 5, pp. 855–868, 2009.
[5] V. Pham, T. Bluche, C. Kermorvant, and J. Louradour, "Dropout improves recurrent neural networks for handwriting recognition," in *ICFHR*, 2014.
[6] P. Voigtlaender, P. Doetsch, and H. Ney, "Handwriting recognition with large multidimensional long short-term memory recurrent neural networks," in *ICFHR*, 2016.
[7] T. Bluche and R. Messina, "Gated convolutional recurrent neural networks for multilingual handwriting recognition," in *ICDAR*, vol. 01, 2017.
[8] J. Puigcerver, "Are multidimensional recurrent layers really necessary for handwritten text recognition?" in *ICDAR*, 2017.
[9] D. Castro, B. L. D. Bezerra, and M. Valena, "Boosting the deep multidimensional long-short term memory network for handwritten recognition systems," in *ICFHR*, 2018.

23

[10] D. Keysers, T. Deselaers, H. A. Rowley, L. Wang, and V. Carbune, "Multi-language online handwriting recognition," *PAMI*, vol. 39, no. 6, pp. 1180–1194, 2017.

[11] V. Carbune, P. Gonnet, T. Deselaers, H. A. Rowley, A. Daryin, M. Calvo, L.-L. Wang, D. Keysers, S. Feuz, and P. Gervais, "Fast multi-language lstm-based online handwriting recognition," *ArXiV*, 2019.

[12] J. Walker, Y. Fujii, and A. C. Popat, "A web-based ocr service for documents," in *DAS*, Apr 2018, pp. 21–22.

[13] S. Ghosh and A. Joshi, "Text entry in indian languages on mobile: User perspectives," in *India HCI*, 2014.

[14] T. M. Breuel, "Tutorial on ocr and layout analysis," in *DAS*, 2018.

[15] Y. Fujii, K. Driesen, J. Baccash, A. Hurst, and A. C. Popat, "Sequence-to-label script identification for multilingual ocr," in *ICDAR*, 2017.

[16] M. Kozielski, P. Doetsch, and H. Ney, "Improvements in rwth's system for off-line handwriting recognition," in *ICDAR*, 2013.

[17] P. Doetsch, M. Kozielski, and H. Ney, "Fast and robust training of recurrent neural networks for offline handwriting recognition," in *ICFHR*, 2014.

[18] P. Voigtlaender, P. Doetsch, S. Wiesler, R. Schlter, and H. Ney, "Sequence-discriminative training of recurrent neural networks," in *ICASSP*, 2015.

[19] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *ICASSP*, 2015.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[21] J. Wang and X. Hu, "Gated recurrent convolution neural network for ocr," in *NIPS*, 2017.

[22] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *CVPR*, 2015, pp. 3367–3375.

[23] J. Sánchez, A. Toselli, V. Romero, and E. Vidal, "ICDAR 2015 Competition HTRtS: Handwritten Text Recognition on the tranScriptorium Dataset," in *ICDAR*, 2015.

[24] A. Toselli, V. Romero, M. Villegas, E. Vidal, and J. Sánchez, "ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset," in *ICFHR*, 2016.

[25] J. Sánchez, V. Romero, A. Toselli, M. Villegas, and E. Vidal, "ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset," in *ICDAR*, 2017.

[26] Folger Shakespeare Library, "Early Modern Manuscripts Online (EMMO)." [Online]. Available: https://emmo.folger.edu

[27] K. Chen, L. Tian, H. Ding, M. Cai, L. Sun, S. Liang, and Q. Huo, "A compact cnn-dblstm based character model for online handwritten chinese text recognition," in *ICDAR*, 2017.

[28] A. Graves, "Generating sequences with recurrent neural networks," *ArXiV*, 2013.

[29] U. Marti and H. Bunke, "The IAM-database: An english sentence database for off-line handwriting recognition," *IJDAR*, vol. 5, pp. 39–46, 2002.

[30] M. Liwicki and H. Bunke, "Iam-ondb - an on-line english sentence database acquired from handwritten text on a whiteboard," in *ICDAR*, 2005.

[31] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *OSDI*, 2016.

[32] A. Poznanski and L. Wolf, "Cnn-n-gram for handwritingword recognition," in *CVPR*, 2016.

APPENDIX

A. Line Recognition Model

Figure 8 shows our inception-style CNN used for all line recognition models. LSTM-based models use stacked BLSTMs on top of this network. GRCL models use the GRCL layers shown in Figure 1 on top of the network. Figures 11, 9, and 10 give more details on the building blocks.

B. Dual-head Script and Style Identifier Model

Figure 8 shows the encoder network used for the dual-head script and style identifier. The output of the encoder network is fed to two summarizer networks for script and style prediction. We use the same Gate summarizer described in [15].
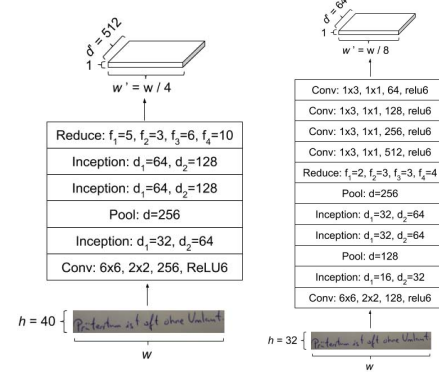

Fig. 7. Left: Inception-style CNNs used for all line recognizers: "Conv: $h_f \times w_f, h_s \times w_s, d, f$" represents a convolutional neural network layer of filter size $h_f \times w_f$, stride $h_s \times w_s$, channels d, activation f, and SAME padding. The Inception module, the Pool module, and the Reduce module are shown in Figure 11, Figure 9, and Figure 10, respectively.

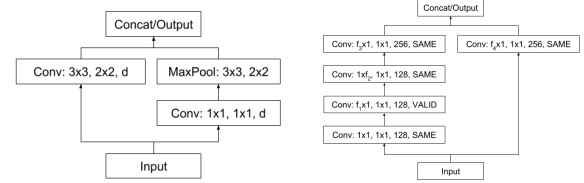Fig. 8. Right: Encoder used for the dual-head script and style identifier.

Fig. 9. Left: Pool module: "Conv: $h_f \times w_f, h_s \times w_s, d$" represents a convolutional neural network layer of filter size $h_f \times w_f$, stride $h_s \times w_s$, channels d, ReLU6 activation, and SAME padding. "MaxPool: $h_f \times w_f$" represents a max pooling layer of filter size $h_f \times w_f$, stride 1×1, and SAME padding.

Fig. 10. Right: Reduce module: "Conv: $h_f \times w_f, h_s \times w_s, d, p$" represents a convolutional neural network layer of filter size $h_f \times w_f$, stride $h_s \times w_s$, channels d, ReLU6 activation, and padding p.
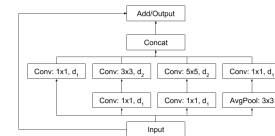

Fig. 11. Inception module: "Conv: $h_f \times w_f, d$" represents a convolutional neural network layer of filter size $h_f \times w_f$, stride 1×1, channels d, ReLU6 activation, and SAME padding. "AvgPool: $h_f \times w_f$" represents an average pooling layer of filter size $h_f \times w_f$, stride 1×1, and SAME padding.