On the Benefits of Convolutional Neural Network Combinations in Offline Handwriting Recognition

Dewi Suryani*†, Patrick Doetsch* and Hermann Ney*
*Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, 52056 Aachen, Germany
†King Mongkut's University of Technology North Bangkok,
The Sirindhorn International Thai-German Graduate School of Engineering, Bangkok, Thailand
dewi.s-sse2014@tggs-bangkok.org, {doetsch, ney}@cs.rwth-aachen.de

***Abstract*—In this paper, we elaborate the advantages of combining two neural network methodologies, convolutional neural networks (CNN) and long short-term memory (LSTM) recurrent neural networks, with the framework of hybrid hidden Markov models (HMM) for recognizing offline handwriting text. CNNs employ shift-invariant filters to generate discriminative features within neural networks. We show that CNNs are powerful tools to extract general purpose features that even work well for unknown classes. We evaluate our system on a Chinese handwritten text database and provide a GPU-based implementation that can be used to reproduce the experiments. All experiments were conducted with RWTH OCR, an open-source system developed at our institute.**

***Keywords*-convolutional neural network; long short-term memory; hybrid HMM; framewise training; offline handwriting; continuous Chinese handwritten text;**

I. Introduction

Handwriting recognition (HWR) is the task of transcribing text data from images of handwritten text. Recognizing handwritten text that is scanned from paper is called offline HWR, and online HWR recognizes text that originates from a special device like digitizer or personal digital assistant (PDA). The difference between them is only the sources of their feature extraction [1]. Nowadays, offline handwriting recognition has increased in terms of popularity by researchers due to the difficulty of its process. The difficulties consist of cursive nature of handwriting, the different size and shape of each character, and large vocabularies. Various methods [2], [3], [4] had been proposed to tackle these problems. Specifically, methods using convolutional neural networks (CNNs) or long short-term memory recurrent neural networks (LSTMs) in combination with training frameworks such as hidden Markov models (HMMs) achieve state of the art performance [2], [3]. However, here the framewise features for either hybrid or tandem HMMs are encoded either by CNNs or LSTMs. Therefore, we study the processing of framewise image features using CNNs followed by LSTMs for offline Chinese handwriting recognition.

This work consists of two different experiments: a preliminary and a main experiment. The preliminary experiment was conducted on the isolated character only, which contains a single character per input image and therefore does not involve sequence modeling. The main experiment then processes full text line images, where LSTMs are used to perform the sequence modeling part while CNNs act as preprocessing layers. We evaluate our results on the openly available database called CASIA [5]. The software used in the experiments is freely available for academic research purpose and setups to reproduce the experiments of this paper will be provided.

The rest of this paper is structured as follows. In Section II, we describe the relation of other approaches to ours. Then, the details of our method are presented in Section III. Section IV describes our experimental settings and its results. Lastly, Section V concludes our works in this paper and discusses the future work.

II. Related Work

CNNs are popular for image and object recognition tasks [6], [7], [8]. One of the most successful implementations is GoogLeNet [8], which is built of many CNN layers with a so-called *Inception* architecture. Moreover, Bluche et al. [2] and Messina et al. [4] successfully implemented CNNs for handwriting recognition. In [2], CNNs are combined with hybrid and tandem HMMs for handwritten English (IAM [9]) and French (RIMES [10]) word recognition. The results are comparable to recurrent neural network (RNN) systems [11] and improved previously used Gaussian HMM systems [12]. Moreover, Messina et al. [4] use the CNNs within multi-dimensional LSTMs with Connectionist Temporal Classification (CTC) for recognizing offline handwritten Chinese (CASIA [13]) text without explicit segmentation. Their result is comparable with the multiple contexts system described in [13]. The combination of a HMM and a neural network is commonly used today [2], [3]. However, in [3], the neural network is an LSTM and the authors combine some successful techniques of object, image, and speech recognition with Bidirectional LSTM (BLSTM) in HMM training systems. Their approach improves the results for text line task on the IAM and RIMES corpora. Compared to the existing approaches, our method combines CNNs

2167-6445/16 $31.00 © 2016 IEEE
DOI 10.1109/ICFHR.2016.43

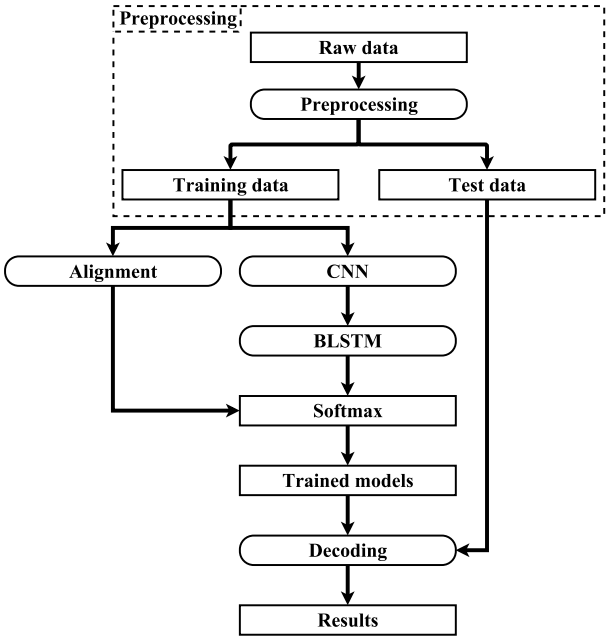

Figure 1. An overview of our system architecture. Input frames are extracted from the raw data and moment-normalized. The training data is aligned to the transcription using a previously trained HMM. The labeled training data is then processed with a cascade of CNN layers, followed by several BLSTM layers. The final model is used to decode unseen test data in a hybrid HMM fashion.

followed by LSTMs with a hybrid HMM. It is motivated by the reported improvements obtained by either combining CNNs and HMMs or LSTMs and HMMs. Furthermore, Donahue et al. [14] also take the advantages of combining CNNs and LSTMs. Their results show good performance of those neural networks for visual recognition and description. However, our method combines not only CNN and LSTM, but also a hybrid HMM and focuses on handwriting task instead of the visual recognition and description. Unlike [2] and [3], we evaluate our approaches on handwritten Chinese (CASIA) text. Our approach is similar to [4], but we will use BLSTMs, study larger networks and evaluate the system in a hybrid HMM instead of using CTC. We furthermore elaborate the reusability of filters learned by a CNN for offline handwriting recognition.

III. Proposed Method

Our proposed method is composed of three major components: a HMM, a CNN and an LSTM. We will describe each component in this section. The full processing pipeline is illustrated in Figure 1.

A. Convolutional neural networks

CNNs are a variant of feed-forward neural networks with a special architecture. The architecture of CNNs usually contains a convolution followed by a pooling operation. Every neuron in a convolutional layer is connected to some region in the input, which is called a local receptive field. Unlike other types of feed-forward neural networks, all weights are shared based on the position within a receptive field. The shared weights are also called filters. The convolutions operation can be formalized as follows:

$$(f * g)(z) = \sum_x \sum_y f(x,y) \cdot g(z-x, z-y) \qquad (1)$$

where $f(x,y)$ is the input image at position (x,y) and $g(z-x, z-y)$ is a trainable filter. Further, a pooling layer is usually used to generate translation invariant features by computing statistics of the convolution activations from different positions along specific windows. One pooling layer that is commonly used in the CNN implementations is the max pooling layer, which takes the maximum value over a processed region. For the two dimensional case, the max pooling operation with a pooling size of $j \times k$ becomes

$$MP(f(x,y))_{jk} = \max_{\substack{m=j-x,\ldots,j+x \\ n=k-y,\ldots,k+y}} \{f(m,n)\} \qquad (2)$$

In this paper, CNNs are applied on two different tasks. In a preliminary experiment, the isolated character data are directly used as the input of the CNN and passed through to a fully connected layer, which is a feed-forward neural network with 1024 neurons. Here, the CNNs are composed of 18 convolutional layers and 3 max pooling layers. We used filter sizes of 1×1, 3×3, and 5×5 and 32 or 64 feature maps. The convolutional layers with 1×1 filters are used to reduce the number of parameters by reducing the feature maps. In order to fit our input into the network of the main experiment, which uses image slices of size 8×32, we applied a max pooling operation with asymmetric filter size 1×2 at the fifth CNN layer. The remaining of the max pooling layers use pooling size 2×2. The details of our CNN architecture in the preliminary experiment are depicted in Figure 2, and Figure 3 visualizes its outputs with 留 characters as the example. In the main experiment, we merge the isolated character data and text lines data by interpreting every single character as a short text line for the input, which is processed by a sliding window technique in order to obtain 8×32 frames. Due to the memory constraints, we applied only five CNN layers in contrast to 18 as in the preliminary experiment. The architecture with five convolutional layers and five max pooling layers is illustrated in Figure 4. Furthermore, we replace the fully connected layer by three BLSTM layers (see next section), where each BLSTM layer contains 512 memory cells for each direction.

B. Long short-term memory

Recurrent neural networks (RNNs) are a kind of neural network that consider not only the current input but also

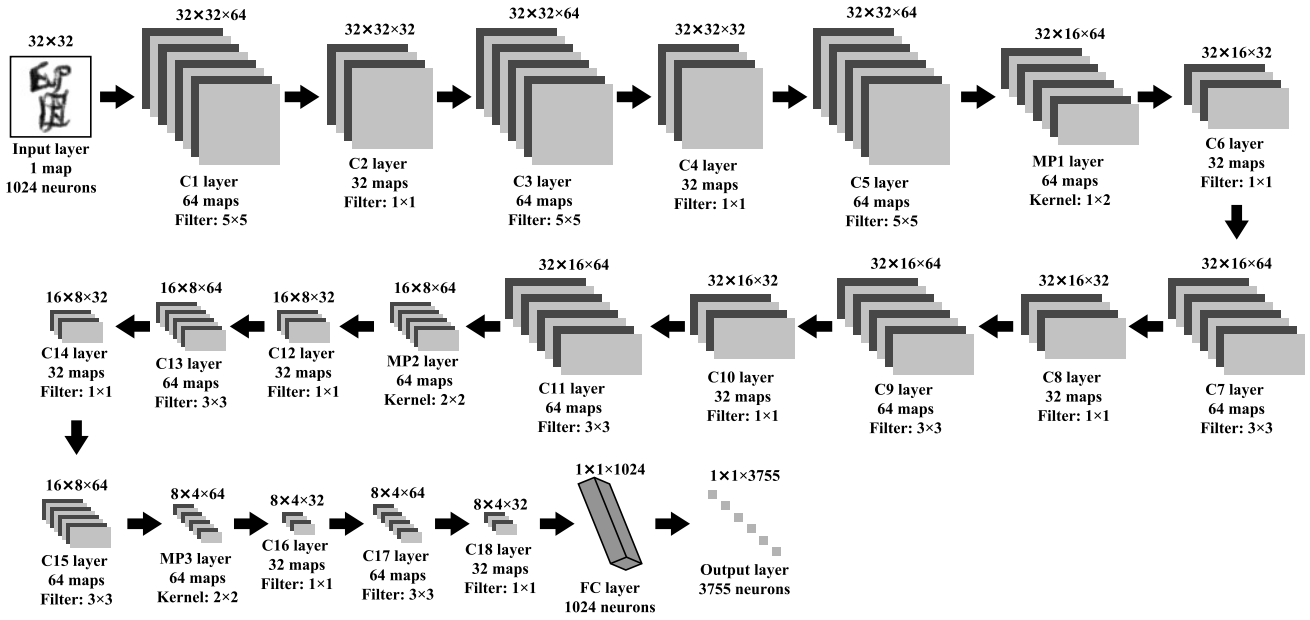

Figure 2. The CNN architecture of preliminary experiment that was conducted on the isolated character data, where the C1, C2, ..., C18 layers are the convolutional layers. The max pooling layers are defined as MP1, MP2, and MP3. Lastly, a fully connected layer combines the filter which are then classified using a softmax layer.

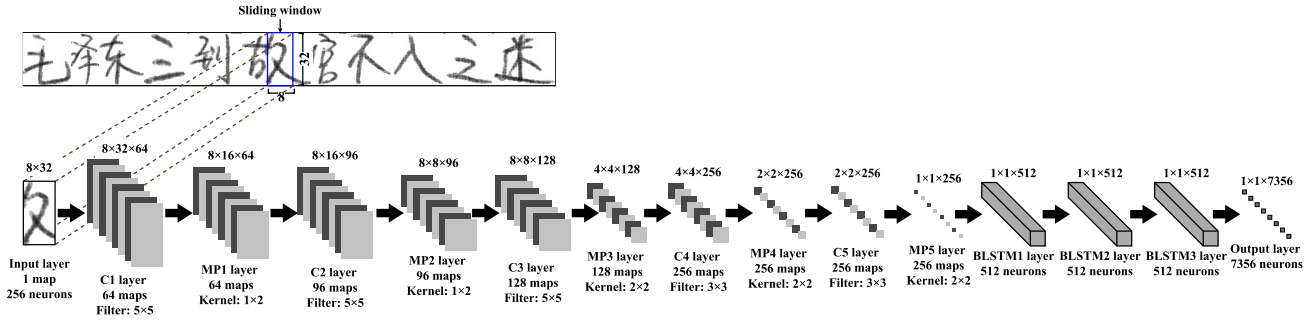

Figure 4. The CNN architecture of the main experiment. An input slice of 8×32 pixels is processed by five consecutive CNN layers followed by three BLSTM layers and finally classified into characters using a softmax layer.

the own internal hidden state of the previous time step when computing the output activation. Unlike a feed-forward neural network, the sequential information of each neuron can be saved and reused. Even though the RNNs perform well, a main obstacle of RNN training had appeared that lead to an unstable gradient computation, which became well-known as the *vanishing gradient* problem. The analysis in [15] lead to the invention of the long short-term memory recurrent neural networks (LSTMs), which greatly improve gradient propagation during optimization. This is achieved by three additional gates that protect the gradient information within the neural network, name an input, a forget, and an output gate. The input gate protects the neuron by filtering every input. To decide whether the neuron information should be maintained or removed is the task of a forget gate. Then the output gate limits the memory capacity that goes out as a hidden state. Each gate uses a logistic sigmoid as activation function, which ranges from 0 to 1. In our system we use bidirectional LSTMs (BLSTMs), which are added on top of the CNN layers. BLSTMs are a direct extension to LSTMs, where the input sequence is scanned both in forward and backward direction and subsequently merged into a combined representation. This model was implemented by [16] and is nowadays applied to many sequential learning tasks including handwriting and speech recognition. Our BLSTM has 512 neurons in each layer and direction, and in total three BLSTMs applied in our network.

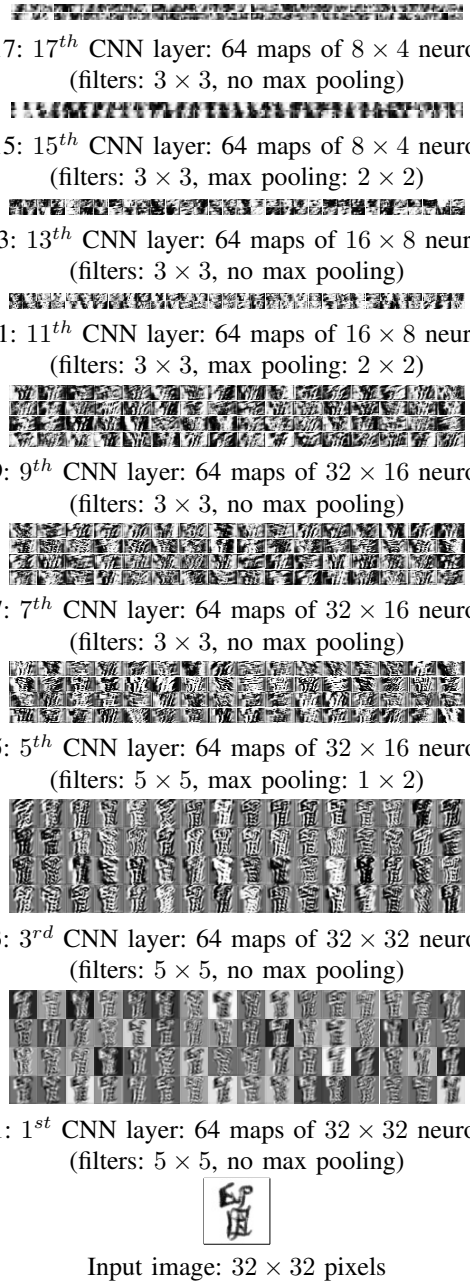

Figure 3. CNN visualization of filter training on isolated characters. For every second CNN layer the resulting feature maps are plotted.

C. Hidden Markov model

The HMM is a state model that is widely used for handling sequential or temporal data, and efficient algorithms for training and inference in HMMs have been developed [17]. An HMM contains transition probabilities $p(s_t|s_{t-1})$, for the transition from a previous state s_{t-1} to the current state s_t, and emission probabilities $p(x_t|s_t) = p(x_1, x_2, ..., x_t|s_1, s_2, ..., s_t)$, which assign a probability to an input x given a state s. It can be formulated as follows:

$$P(x_1, ..., x_T, s_1, ..., s_T) = \prod_{t=1}^{T} p(x_t|s_t)p(s_t|s_{t-1}) \quad (3)$$

In our system the HMM is used in two phases. First, it is trained with Gaussian mixture models as emission probability distributions in order segment the training data using a Viterbi alignment, such that each frame is mapped to a label in the lexicon and a framewise neural network training is possible. After neural network training, the softmax activations are rescaled by the label prior and used as replacement of the Gaussian mixture models when new data is recognized. This decoding technique is commonly referred to as hybrid HMM approach and is implemented as in [3].

IV. Experiment and Result

A. Dataset

In this approach, we used the offline Chinese handwriting datasets, which was created by the Institute of Automation of Chinese Academy of Sciences (CASIA), as described in [5]. The datasets consist of isolated characters (HWDB 1.0-1.2) and handwritten text (HWDB 2.0-2.2) data with 7356 classes in total. Similar to [18], we used 3,118,477 samples of isolated character data and 4,076 pages of handwritten text data as our training set. However, 10% of the training set were used to gauge the convergence of our training.

Moreover we used the handwritten text dataset from the ICDAR 2013 Handwritten Chinese Character Recognition Competition [13] as evaluation dataset, which consists of 300 text pages decomposed into 3,397 lines. The ICDAR 2013 dataset was written by 60 writers, who did not write any samples from other CASIA datasets (HWDB 1.0-2.2), and the text content is completely different from the HWDB 2.0-2.2 datasets. To compare our result with the literature [4], we only processed the text lines that did not contain invalid characters. Therefore, the numbers of our extracted lines is smaller than the reported number from ICDAR 2013, which was 3,432 lines.

B. Language model

The language model was estimated based on two text corpora: from the training data of CASIA itself [5] and from the PH [19] corpus. From CASIA we used the transcriptions of the given training set, which contains 7,356 unique characters and 4,203,395 characters in total. The PH corpus consists of 3,753,291 characters with 4,723 unique characters. A 10-gram LM with Kneser-Ney smoothing was trained on the label data of the combined corpora and used in the recognition of our systems.

C. Preprocessing and experimental settings

First of all, we preprocessed all data similar to [3]. However, the principal components analysis (PCA) step only applied to extract our baseline system. Starting with full

Table I
THE RESULT OF THE PRELIMINARY EXPERIMENT ON THE ISOLATED CHARACTERS. NUMBERS ARE REPORTED AS "CORRECT RATE" (CR) IN PERCENT.

System	HWDB 1.1	ICDAR 2013
Our CNNs architecture	92.00	92.20
CNN [22]	90.46	92.88
MQDF + CNN + MC [22]	92.03	94.44

pages containing multiple lines of text as the input, the first step is to segment the lines using the ground truth that is provided by [5]. This is followed by a moment-based normalization. We then extract 8×32 frames by applying a sliding window and a shift of 3 pixels. HMM training and decoding was done in RASR [20], and the neural networks were trained with our Theano-based RNN training framework RETURNN [21]. We trained and evaluated our approach on GeForce GTX 980 GPUs, and all evaluation results were measured based on the Correct Rate (CR) and the Accuracy Rate (AR) as defined in [13].

D. Isolated Characters

This preliminary experiment was conducted to examine the CNN performance independent of the sequence modeling problem and therefore applied them to isolated character data. In this case, only the CASIA HWDB 1.1 dataset was used to train our CNNs, which contains 3755 classes and 1,121,749 samples The dataset was split into a train and and a validation set based on the writers. For the train set, we used the first 240 writers of HWDB 1.1 (from 1001-f to 1240-f) and the remaining writers were used as validation set. In addition, the isolated characters ICDAR 2013 competition dataset was used as evaluation set. This was done to make a comparison with other groups. All the data was processed as pixel features, which were first converted from the provided binary format to grayscale images form and were resized to 32×32 by a bi-linear scaling algorithm. This scaling method was processed with keeping the ratio by adding the white padding and performed uniformly. Lastly, the data was normalized by its center of gravity. Table I shows our results and those of other groups. We can see that our single deep CNN system is able to achieve competitive performance to a complex system combination scheme.

E. Continuous handwriting recognition

In this experiment, the inputs were 8×32 image slices that were extracted by applying a sliding window to the text line images of the CASIA database. These frames were passed through the CNN layers followed by three BLSTM layers. For comparison we trained a baseline system only composed of the BSTLM layers and which uses a standard preprocessing by reducing the components of the image slices to 24 principal components using PCA. Each layer of BLSTM consists of 512 memory cells in each direction.

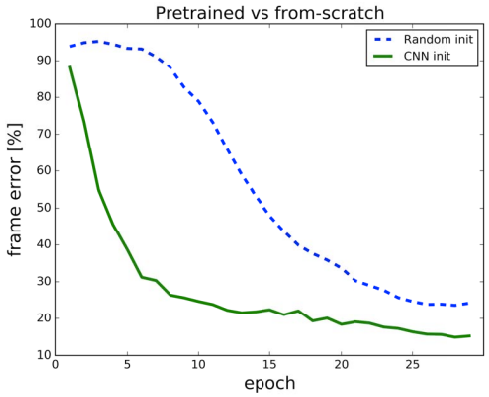

Figure 5. Identical systems were trained, but one with a random weight initialization and another with filters obtained from training on the isolated characters. We can not only observe a faster convergence but also a significantly better overall performance if the filters are initialized from previously trained CNNs. The errors reported are relative numbers of misclassified frames.

Table II
THE RESULT OF OUR MAIN EXPERIMENT COMPARED TO RESULTS OF OTHER GROUPS. THE SECOND COLUMN SHOWS THE ACCURACY RATE OF THE VALIDATION SET, AND THE THIRD COLUMN FOR THE EVALUATION SET OF THE COMPETITION.

System	Valid [%]	Eval [%]
CNN + LSTM (Our system)	84.87	83.50
LSTM (Baseline)	76.24	77.00
HMM + MQDF [13]	–	86.73
Multiple geometric context + LM [18]	–	89.28
MDLSTM + CTC [4]	–	89.4

Table II describes the result of this experiment and a comparison with the baseline and other groups is shown. We can see that the CNN based system outperforms the baseline without CNNs by a large magnitude. Moreover, to proof the CNN-based feature is better than PCA-based feature, we conduct an experiment which only contains a single CNN layer with 64 maps of size 5×5 followed by a 2×2 max pooling for CNN-based features. That experiment achieved the AR 83.74% on valid and 80.60% on eval, which means a 9.8% and 4.7% relative improvement compared to PCA-based features (baseline) for valid and eval, respectively. In order to make use of the isolated characters, we studied the effect of initializing the convolutional layer in the text line system with filters that were trained on the isolated characters only. As shown in Figure 5 we can observe a strong improvement when the system is initialized with such filters, which suggests that filters might also be reusable in tasks with multiple scripts and languages.

V. CONCLUSION AND FUTURE WORK

We presented our approach of combining a CNN and a BLSTM in a framewise hybrid HMM framework. It shows that the initialized filters of a CNN can achieve

better performance than the randomly initialized filters. The presented system shows competitive performance to much more complex systems and the software can be obtain at our institute's website in order to reproduce the experiment. More importantly, in the future we are looking to take even more advantage of pretrained convolutional filters for the specific purpose of offline handwriting recognition, by taking multiple scripts and data sources into account.

REFERENCES

[1] A. Graves and J. Schmidhuber, "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks," in *Advances in Neural Information Processing Systems 21*, 2009, pp. 545–552.

[2] T. Bluche, H. Ney, and C. Kermorvant, "Tandem HMM with convolutional neural network for handwritten word recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 2390–2394.

[3] M. Kozielski, P. Doetsch, and H. Ney, "Improvements in RWTH's System for Off-Line Handwriting Recognition," in *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, Aug. 2013, pp. 935–939.

[4] R. Messina and J. Louradour, "Segmentation-free handwritten Chinese text recognition with LSTM-RNN," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Aug. 2015, pp. 171–175.

[5] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "CASIA Online and Offline Chinese Handwriting Databases," in *2011 International Conference on Document Analysis and Recognition (ICDAR)*, Sep. 2011, pp. 37–41.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[7] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Sep. 2014, arXiv: 1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 1–9.

[9] U.-V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," *International Journal on Document Analysis and Recognition*, vol. 5, no. 1, pp. 39–46, Nov. 2002. [Online]. Available: http://link.springer.com/article/10.1007/s100320200071

[10] E. Augustin, M. Carré, E. Grosicki, J.-M. Brodin, E. Geoffrois, and F. Prêteux, "Rimes evaluation campaign for handwritten mail processing," in *Proceedings of the Workshop on Frontiers in Handwriting Recognition*, vol. 1, 2006.

[11] F. Menasri, J. Louradour, A.-L. Bianne-Bernard, and C. Kermorvant, "The a2ia french handwriting recognition system at the rimes-icdar2011 competition," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2012, pp. 82 970Y–82 970Y.

[12] A.-L. Bianne-Bernard, F. Menasri, R. A.-H. Mohamad, C. Mokbel, C. Kermorvant, and L. Likforman-Sulem, "Dynamic and contextual information in hmm modeling for handwritten word recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 10, pp. 2066–2080, 2011.

[13] F. Yin, Q.-F. Wang, X.-Y. Zhang, and C.-L. Liu, "ICDAR 2013 Chinese Handwriting Recognition Competition," in *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, Aug. 2013, pp. 1464–1470.

[14] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[16] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM networks," in *2005 IEEE International Joint Conference on Neural Networks, 2005. IJCNN '05. Proceedings*, vol. 4, Jul. 2005, pp. 2047–2052.

[17] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[18] Q.-F. Wang, F. Yin, and C.-L. Liu, "Handwritten Chinese Text Recognition by Integrating Multiple Contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1469–1481, Aug. 2012.

[19] G. Jin. (1990-1991) The ph corpus (online). [Online]. Available: ftp://ftp.cogsci.ed.ac.uk/pub/chinese

[20] S. Wiesler, A. Richard, P. Golik, R. Schlüter, and H. Ney, "RASR/NN: The RWTH neural network toolkit for speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 3313–3317.

[21] P. Doetsch, A. Zeyer, P. Voigtlaender, I. Kulikov, R. Schlüter, and H. Ney, "RETURNN: The RWTH extensible training framework for universal recurrent neural networks," *submitted to Interspeech 2016*, 2016.

[22] Y. Wang, X. Li, C. Liu, X. Ding, and Y. Chen, "An MQDF-CNN Hybrid Model for Offline Handwritten Chinese Character Recognition," in *2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Sep. 2014, pp. 246–249.