Learning Spatial-Semantic Context with Fully Convolutional Recurrent Network for Online Handwritten Chinese Text Recognition

Zecheng Xie, Zenghui Sun, Lianwen Jin*, Hao Ni and Terry Lyons

Abstract—Online handwritten Chinese text recognition (OHCTR) is a challenging problem as it involves a large-scale character set, ambiguous segmentation, and variable-length input sequences. In this paper, we exploit the outstanding capability of path signature to translate online pen-tip trajectories into informative signature feature maps, successfully capturing the analytic and geometric properties of pen strokes with strong local invariance and robustness. A multi-spatial-context fully convolutional recurrent network (MC-FCRN) is proposed to exploit the multiple spatial contexts from the signature feature maps and generate a prediction sequence while completely avoiding the difficult segmentation problem. Furthermore, an implicit language model is developed to make predictions based on semantic context within a predicting feature sequence, providing a new perspective for incorporating lexicon constraints and prior knowledge about a certain language in the recognition procedure. Experiments on two standard benchmarks, Dataset-CASIA and Dataset-ICDAR, yielded outstanding results, with correct rates of 97.50% and 96.58%, respectively, which are significantly better than the best result reported thus far in the literature.

Index Terms—Handwritten Chinese text recognition, path signature, residual recurrent network, multiple spatial contexts, implicit language model

1 Introduction

In recent years, increasingly in-depth studies have led to significant developments in the field of handwritten text recognition. Various methods have been proposed by the research community, including integrated segmentation-recognition methods [1], [2], [3], [4], [5], hidden Markov models (HMMs) and their hybrid variants [6], [7], segmentation-free methods [8], [9], [10] with long short-term memory (LSTM) and multi-dimensional long short-term memory (MDLSTM), and integrated convolutional neural network (CNN)-LSTM methods [11], [12], [13], [14]. In this paper, we investigate the most recently developed methods for online handwritten Chinese text recognition (OHCTR), which is an interesting research topic presenting the following challenges: a large character set, ambiguous segmentation, and variable-length input sequences.

Segmentation is the fundamental component of handwritten text recognition, and it has attracted the attention of numerous researchers [1], [2], [3], [4], [5], [15], [16]. Among the above-mentioned methods, over-segmentation [1], [2], [3], [4], [5], i.e., an integrated segmentation-recognition method, is the most efficient method and still plays a crucial role in OHCTR. The basic concept underlying over-segmentation is to slice the input string into sequential character segments whose candidate classes can be used to construct the segmentation-recognition lattice [2]. Based on the lattice, path evaluation, which integrates the recognition scores, geometry information, and semantic context, is conducted to search for the optimal path and generate the recognition result. In practice, segmentation inevitably leads to mis-segmentation, which is barely rectifiable through post-processing and thus degrades the overall performance.

Segmentation-free methods are flexible alternative methods that completely avoid the segmentation procedure. HMMs and their hybrid variants [6], [7] have been widely used in handwritten text recognition. In general, the input string is converted into slices by sliding windows, followed by feature extraction and frame-wise prediction using an HMM. Finally, the Viterbi algorithm is applied to search for the best character string with maximum a posteriori probability. However, HMMs are limited not only by the assumption that their observation depends only on the current state but also by their generative nature that generally leads to poor performance in labeling and classification tasks, as compared to discriminative models. Even though hybrid models that combine HMMs with other network architectures, including recurrent neural networks [17] and multilayer perceptrons [18], have been proposed to alleviate the above-mentioned limitations by introducing context into HMMs, they still suffer from the drawbacks of HMMs.

The recent development of recurrent neural networks, especially LSTM [8], [9], [19] and MDLSTM [10], [19], has provided a revolutionary segmentation-free perspective to the problem of handwritten text recognition. In general, LSTM is directly fed with a point-wise feature vector that consists of the (x, y)-coordinate and relative features, while it recurrently updates its hidden state and generates per-frame predictions for each time step. Then, it applies connectionist temporal classification (CTC) to perform transcription. It is worth noting that LSTM and MDLSTM have

- *Z. Xie, Z. Sun, and L. Jin are with College of Electronic and Information Engineering, South China University of Technology, Guangzhou, China. E-mail: {zcheng.xie, sunfreding, lianwen.jin}@gmail.com*
- *H. Ni is with Oxford-Man Institute for Quantitative Finance, University of Oxford, Oxford, UK. E-mail: hao.ni@maths.ox.ac.uk*
- *T. Lyons is with Mathematical Institute, University of Oxford, Oxford, UK. E-mail: tlyons@maths.ox.ac.uk*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2732978, IEEE Transactions on Pattern Analysis and Machine Intelligence

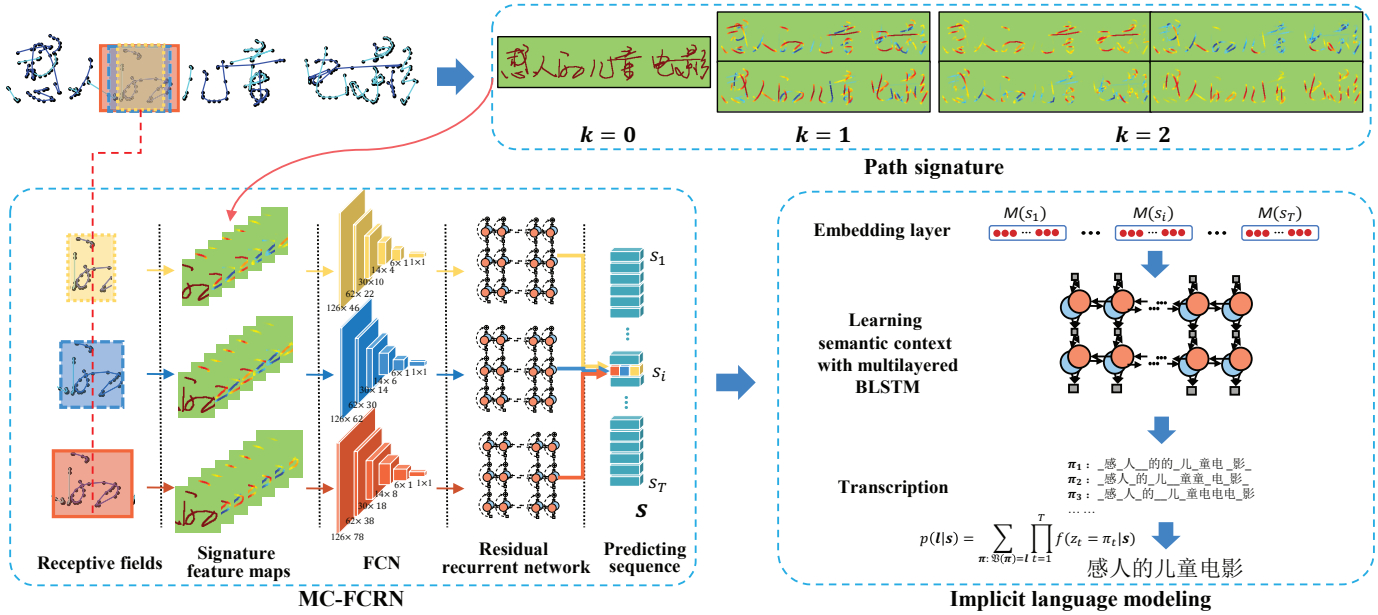


Fig. 1. Overview of the proposed method. Variable-length pen-tip trajectories are first translated into offline signature feature maps that preserve the essential online information. Then, a multi-spatial-context fully convolutional recurrent network (MC-FCRN) take input of the signature feature maps with receptive fields of different scales in a sliding window manner and generate a predicting sequence. Finally, an implicit LM is proposed to derive the final label sequence by exploiting the semantic context of embedding vectors that are transformed from the predicting sequence.

been successfully applied to handwritten text recognition in Western languages, where the character set is relatively small (e.g., for English, there are only 52 classes; therefore it is easy to train the network). However, to the best of our knowledge, very few studies have attempted to address the problem of large-scale (where, e.g., the text lines may be represented by more than 7,000 basic classes of characters and sum up to more than 1 million character samples) handwritten text recognition problems such as OHCTR.

Architectures that integrate CNN and LSTM exhibit excellent performance in terms of visual recognition and description [20], [21], scene text recognition [12], [13], [14], and handwritten text recognition [11]. In text recognition problems, deep CNNs generate highly abstract feature sequences from input sequential data. LSTM is fed with such feature sequences and generates corresponding character strings. Jointly training LSTM with CNN is straightforward and can improve the overall performance significantly. However, in the above-mentioned methods, the CNNs, specifically fully convolutional networks (FCNs), process the input string with only a fixed-size receptive field in a sliding window manner, which we claim is inflexible for unconstrained written characters in OHCTR. Moreover, a deep integrated CNN-LSTM network is usually accompanied by degradation problem [22] that slows down the convergence procedure and affects the system optimization.

In this paper, we propose a novel solution (see Fig. 1) that integrates path signature, a multi-spatial-context fully convolutional recurrent network (MC-FCRN), and an implicit language model (implicit LM) to address the problem of unconstrained online handwritten text recognition. Path signature, a recent development in the field of the rough path theory [23], [24], [25], is a promising approach for translating variable-length pen-tip trajectories into offline signature feature maps in our system, because it effectively preserves the online information that characterizes the analytic and geometric properties of the path. Encouraged by recent advances in deep CNNs and LSTMs, we propose the MC-FCRN for robust recognition of signature feature maps. MC-FCRN leverages the multiple spatial contexts that correspond to multiple receptive fields in each time step to achieve strong robustness and high accuracy. Furthermore, we propose an implicit LM, which incorporates semantic context within the entire predicting feature sequence from both forward and reverse directions, to enhance the prediction for each time step. The contributions of this paper can be summarized as follows:

- We develop a novel segmentation-free MC-FCRN to effectively capture the variable spatial contextual dynamics as well as the character information for high-performance recognition. With a series of receptive fields of different scales, MC-FCRN is able to model the complicate spatial context with strong robustness and high accuracy.
- The residual recurrent network, a basic component of MC-FCRN, not only accelerates the convergence process but also promotes the optimization result, while adding neither extra parameter nor computational burden to the system, as compared to ordinary stacked recurrent network
- We propose an implicit LM that learns to model the output distribution given the entire predicting feature sequence. Unlike the statistical language model that predicts the next word given only a few previous words, our implicit LM exploits the semantic context not only from the forward and reverse directions of the text but also with arbitrary text length.
- Path signature, a novel mathematical feature set, brought from the rough path theory [23], [24], [25]

as a non-linear generalization of classical theory of controlled differential equations, is successfully applied to capture essential online information for long pen-tip trajectories. Moreover, we investigate path signature for learning the variable online knowledge of the input string with different iterated integrals.

The remainder of this paper is organized as follows. Section 2 reviews the related studies. Section 3 formally introduces path signature. Section 4 details the network architecture of FCRN and its extended version, namely MC-FCRN. Section 5 describes the proposed implicit LM and discusses the corresponding training strategy. Section 6 presents the experimental results. Finally, Section 7 concludes the paper.

2 Related Work

Feature extraction [26], [27], [28], [29], [30], [31] plays a crucial role in traditional online handwritten text recognition. The 8-directional feature [26], [29] is widely used in OHCTR owing to its excellent ability to express stroke directions. The projection of each trajectory point in eight directions is calculated in a 2-D manner and eight pattern images are generated accordingly. For further sophistication, Grave et al. [8] considered not only the (x, y)-coordinate and its relationship with its neighbors in the time series but also the spatial information from an offline perspective, thus obtaining 25 features for each point. However, the above-mentioned techniques have been developed empirically. Inspired by the theoretical work of Lyons and his colleagues [23], [24], [25], we applied path signature to translate the online pen-tip trajectories into offline signature feature maps that maintain the essential features for characterizing the online information of the trajectories. Furthermore, we can use truncated path signature in practical applications to achieve a trade-off between complexity and precision.

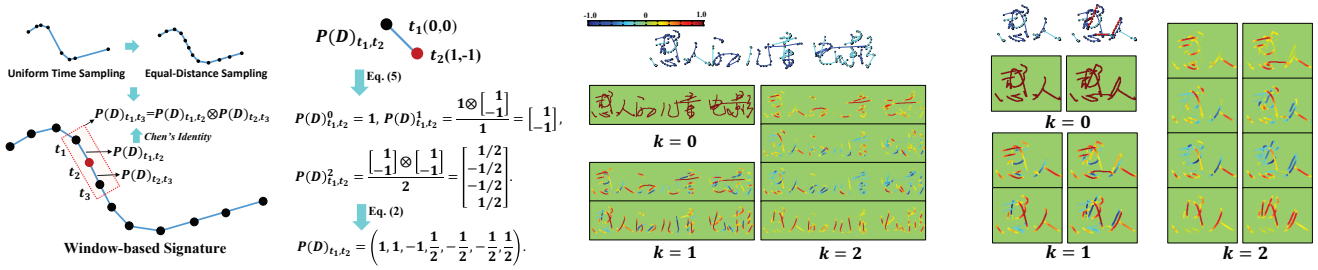
Yang et al. [32], [33] showed that the domain-specific information extracted by the aforementioned methods can improve the recognition performance with deep CNN (DCNN). However, DCNN-based networks are unable to handle input sequences of variable length in OHCTR. On the contrary, LSTM- and MDLSTM-based networks have an inherent advantage in dealing with such input sequences and demonstrate excellent performance in unconstrained handwritten text recognition [8], [9], [34], [35]. Recently, deep learning methods that integrate LSTM and CNN have demonstrated outstanding capability in the field of visual captioning [21], [36] and scene text recognition [12], [13]. However, in this paper, we show that the simple combination of CNN and LSTM cannot utilize their full potential, which is probably due to the degradation problem [22]. On the other hand, highway network [37] [38] and residual connection [22] [39] were advocated to solve the degradation problem [22] in training very deep networks. Therefore, we take inspiration from them and present the residual recurrent network to realize faster and better optimization of the system. Furthermore, our MC-FCRN also differs from these methods in that it uses multiple receptive fields of different scales to capture highly informative contextual features in each time step. Such a multi-scale strategy originates from traditional methods. The pyramid match kernel [40] maps features to multi-dimensional multi-resolution histograms that help to capture co-occurring features. The SIFT vectors [41] search for stable features across all possible scales and construct a high-dimensional vector for the key points. Further, spatial pyramid pooling [42] allows images of varying size or scale to be fed during training and enhances the network performance significantly. GoogLeNet [43] introduced the concept of "inception" whereby multi-scale convolution kernels are integrated to boost performance. We have drawn inspiration from these multi-scale methods to design our MC-FCRN.

In general, language modeling is applied after feature extraction and recognition in order to improve the overall performance of the system [1], [2], [3], [4], [31], [44]. The concept of 'embedding' plays a critical role in computational linguistics. Traditionally, one character is strictly represented with one 'embedding' [45]. However, as emphasized by Vilnis et al [46], representing an object as a single point in space carries limitations. Instead, a density-based distributed embeddings can provide much more information of each word, e.g. capturing uncertainty about a representation and its relationship. Recently, Mukherjee [47] further verified that a visual-linguistic mapping where words and visual categories are both represented by distribution can improve result at the intersection of language and vision, due to the better exploiting of intra-concept variability in each modality. In this paper, we take inspiration from their works and take the predicting feature sequence, instead of one-hot vectors, as input of the implicit LM to maintain the intra-concept variability, which reflects recognition confidence information in our problem. The recent development of neural networks, especially LSTM, in the field of language translation [48] and visual captioning [20], [21] has provided us with a new perspective of language models. To the best of our knowledge, neural networks were first applied to language modeling by Bengio et al. [45]. Subsequently, Mikolov et al. [49] used recurrent neural network, and Sundermeyer et al. [50] used LSTM for language modeling. For language translation, Sutskever et al. [48] used multilayered LSTM to encode the input text into a vector of fixed dimensionality and then applied another deep LSTM to decode the text in a different language. For visual captioning, Venugopalan et al. [21] and Pan et al. [20] extracted deep visual CNN representations from image or video data and then used an LSTM as a sequence decoder to generate a description for the representations. Partially inspired by these methods, we developed our implicit LM to incorporate semantic context for recognition. However, unlike the above-mentioned methods, which only derive context information from the past predicted text, our implicit LM learns to make predictions given the entire predicting feature sequence in both forward and reverse directions.

3 Path Signature

Proper translation of online data into offline feature maps while retaining most, or hopefully all, of the online knowledge within the pen-tip trajectory plays an essential role in online handwritten recognition. To this end, we investigate path signature, which was pioneered by Chen [51] in the form of iterated integrals and developed by Lyons and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2732978, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) (b) (c) (d)

Fig. 2. (a) Illustration of feature extraction of path signature. (b) A simple example of calculation of path signature features. (c) Path signature of one typical online handwritten text example. (d) Left: path signature of the original pen-tip trajectories; Right: path signature of the pen-tip trajectories with randomly added connections between adjacent strokes. It is notable that excepting for the additional connections, the original part of the sequential data has the same path signature (same color).

his colleagues as a fundamental component of rough path theory [23], [24], [25]. Path-signature was first introduced into handwritten Chinese character recognition by Benjamin Graham [52], and followed by Yang et al [32], [33], but only at the character level. We go further by applying path signature to extremely long sequential data that usually consist of hundreds of thousands of points and prove its effectiveness in OHCTR problem. In the following, we first briefly introduce path signature theoretically, and then technically for sake of implementation and application.

Consider the pen strokes of the online handwritten text collected from a writing plane $H \subset R^2$. Then, a pen stroke can be expressed as a continuous mapping denoted by $D : [a, b] \rightarrow H$ with $D = (D_t^1, D_t^2)$ and $t \in [a, b]$. For $k \geq 1$ and a collection of indexes $i_1, \cdots, i_k \in \{1, 2\}$, the k-th fold *iterated integral* of D along the index $i_1, \cdots, i_k$ can be defined by

$$P(D)_{a,b}^{i_1,\cdots,i_k} = \int_{a<t_1<\cdots<t_k<b} dD_{t_1}^{i_1}, \cdots, dD_{t_k}^{i_k}. \quad (1)$$

The signature of the path is a collection of all the iterated integrals of D:

$$\begin{aligned} P(D)_{a,b} =& (1, P(D)_{a,b}^1, P(D)_{a,b}^2, P(D)_{a,b}^{1,1}, \\ & P(D)_{a,b}^{1,2}, P(D)_{a,b}^{2,1}, P(D)_{a,b}^{2,2}, \cdots), \end{aligned} \quad (2)$$

where the superscripts of the terms $P(X)_{a,b}^{i_1,\cdots,i_k}$ run over the set of all *multi-indexes*

$$G = \{(i_1, ..., i_k) | i_1, \cdots, i_k \in \{1, 2\}, k \geq 1\}. \quad (3)$$

Then, the k-th iterated integral of the signature $P(D)_{a,b}^{(k)}$ is the finite collection of terms $P(D)_{a,b}^{i_1,\cdots,i_k}$ with multi-indexes of length k. More specifically, $P(D)_{a,b}^{(k)}$ is the 2^k-dimensional vector defined by

$$P(D)_{a,b}^{(k)} = (P(X)_{a,b}^{i_1,\cdots,i_k} | i_1, \cdots, i_k \in \{1, 2\}). \quad (4)$$

In [25], it is proved that the whole signature of a path determines the path up to time re-parameterization; i.e., path signature can not only characterize the path displacement and its further derivative as the classical directional features do, but also provide more detailed analytic and geometric properties of the path. In practice, we have to use the truncated signature feature, which can capture the global information on the path. Increasing the degree of truncated signature results in the exponential growth of dimension but may not always lead to significant gain.

Next, we describe the practical calculation of path signature in OHCTR from the implementation and application point of view. As illustrated in Fig. 2a, the pen-tip trajectories of the online handwritten text samples are represented by a sequence of uniform-time sampling points. First, the uniform-time sampling trajectory is translated into equal-distance sampling style. Then, to calculate the signature feature for a specific point, e.g., the red point in Fig. 2a, we estimate the window-based signature $P(D)_{t_1,t_3}$ that takes this point as the midpoint. In order to calculate $P(D)_{t_1,t_3}$, we first compute point-wise signature $P(D)_{t_1,t_2}$ and $P(D)_{t_2,t_3}$; then combine them according to *Chen's identity* [51].

As adjacent sampling points of text samples are connected by a straight line $D = (D_t^1, D_t^2)$ with $t \in [a, b]$, the iterated integrals $P(D)_{a,b}^{(k)}$ can be calculated iteratively as follows:

$$P(D)_{a,b}^{(k)} = \begin{cases} 1 & , k = 0, \\ (P(D)_{a,b}^{(k-1)} \otimes \triangle_{a,b})/k & , k \geq 1, \end{cases} \quad (5)$$

where $\triangle_{a,b} := D_b - D_a$ denotes the path displacement and $\otimes$ represents the tensor product. In Fig. 2b, we provide a simple example to explain the calculation of path signature according to Eq. (5) and Eq. (2). Suppose we have two adjacent straight lines $D = (D_t^1, D_t^2)$ with $t \in [t_1, t_2]$ and $D = (D_t^1, D_t^2)$ with $t \in [t_2, t_3]$, as shown in Fig. 2a. Then, following *Chen's identity* [51], we can calculate the path signature for the concatenation of these two paths as

$$P(D)_{t_1,t_3}^{(k)} = \sum_{i=0}^{k} P(D)_{t_1,t_2}^{i} \otimes P(D)_{t_2,t_3}^{k-i}. \quad (6)$$

Given the pen-tip trajectories of online handwritten text, for each sequential stroke point, we first compute the path signature within a sliding window according to Eq. (2) and Eq. (6). Then the path signature features of certain level (k) along all the stroke points will form the corresponding feature maps. Specifically, the path signature feature vector of each stroke point spreads over $2^{(k+1)}-1$ two-dimensional matrices, according to the coordinates of the stroke point of the handwritten text data. The $0, 1, 2$-th iterated integral signature feature maps, i.e., the above-mentioned two-dimensional matrices, of one typical online handwritten text example are visualized in Fig. 2c.

In Fig. 2a, $P(D)_{t_1,t_3}$ is computed with window size 3. In practice, we set the window size as 9 to keep strong local invariance and robustness. Fig. 2d shows that, although connections are randomly added between adjacent strokes within a character or between characters, their impact on the path signature of the original input string is not significant, which proves that path signature based on sliding window has excellent local invariance and robustness.

4 Multi-Spatial-Context FCRN

Unlike character recognition, where it is easy to normalize characters to a fixed size, text recognition is complicated because it involves input sequences of variable length, such as feature maps and online pen-tip trajectories. We propose a new fully convolutional recurrent network (FCRN) for spatial context learning to overcome this problem by leveraging a fully convolutional network, a residual recurrent network, and connectionist temporal classification, all of which naturally take inputs of arbitrary size or length. Furthermore, we extend our FCRN to multi-spatial-context FCRN (MC-FCRN), as shown in Fig. 1, to learn multi-spatial-context knowledge from complicated signature feature maps. In the following subsections, we briefly introduce the basic components of FCRN and explain their roles in the architecture. Then, we demonstrate how MC-FCRN performs multi-spatial-context learning for the OHCTR problem.

4.1 Fully Convolutional Recurrent Network

4.1.1 Fully Convolutional Network

DCNNs exhibit excellent performance in computer vision applications such as image classification [39], [42], scene text recognition [12], [13], and visual description [20], [21]. Following the approach of Long et al. [53], we remove the original last fully connected classification layer from DCNNs to construct a fully convolutional network. Fully convolutional networks not only inherit the ability of DCNNs to learn powerful and interpretable image features but also adapt to variable input image size and generate corresponding-size feature maps. It is worth noting that such CNN feature maps contain strong spatial order information from the overlap regions (known as receptive fields) of the original feature maps. Such spatial order information is very important and can be leveraged to learn spatial context to enhance the overall performance of the system. Furthermore, unlike image cropping or sliding window-based approaches, FCNs eliminate redundant computations by sharing convolutional response maps layer by layer to achieve efficient inference and backpropagation.

4.1.2 The Residual Recurrent Network

Recurrent neural networks (RNNs), which are well known for the self-connected hidden layer that recurrently transfers information from output to input, have been widely adopted to learn continuous sequential features. Recently, long short-term memory (LSTM) [54], a variant of RNN that overcomes the gradient vanishing and exploding problem, has demonstrated excellent performance in terms of learning complex and long-term temporal dynamics in applications such as language translation [55], visual description [20], [21], and text recognition [12], [13]. Bidirectional LSTM (BLSTM) facilitates the learning of complex context dynamics in both forward and reverse directions, thereby outperforming unidirectional networks significantly. Stacked LSTM is also popular for sequence learning for accessing higher-level abstract information in temporal dimensions.

Integrated CNN-LSTM systems demonstrate their outstanding capability in visual recognition and description [20], [21] and scene text recognition [12], [13], [14]. However, the degradation problem [39] usually accompanies deep integrated CNN-LSTM networks and slows down the convergence process. Driven by the significance of deep residual learning [22], [39] for optimization of very deep networks, we presented the residual recurrent network to accelerate the convergence of our FCRN and obtain better optimization result. Theoretically, we explicitly reformulate the LSTM\BLSTM layer (denoted by h with parameter $\boldsymbol{\omega}_h$) as learning the spatial contextual information with reference to the input. Denoting the l-th LSTM\BLSTM layer output as $q_l(\boldsymbol{x})$, we have

$$q_l(\boldsymbol{x}) = h(q_{l-1}(\boldsymbol{x})) + q_{l-1}(\boldsymbol{x}). \tag{7}$$

Iteratively applying $q_l(\boldsymbol{x}) = h(q_{l-1}(\boldsymbol{x})) + q_{l-1}(\boldsymbol{x}) = h(q_{l-1}(\boldsymbol{x})) + h(q_{l-2}(\boldsymbol{x})) + q_{l-2}(\boldsymbol{x})$ to $q_L(\boldsymbol{x})$, we get

$$q_L(\boldsymbol{x}) = q_0(\boldsymbol{x}) + \sum_{l=0}^{L-1} h(q_l(\boldsymbol{x})), \tag{8}$$

where L is the total number of layers of the residual multi-layered LSTM\BLSTM. Residual recurrent network has the following advantages in jointly learning with deep CNN. First, gradient information can easily pass through the complex residual recurrent network through the identity mapping $q_L(\boldsymbol{x}) = q_0(\boldsymbol{x})$ according to Eq. (8), as the term $\sum_{l=1}^{L-1} h(q_l(\boldsymbol{x}))$ for the residual spatial learning is very small and has not yet functioned in the early training stage. Therefore, the system gains rapid growth in the nascent period (as illustrated in Fig. 6). Furthermore, by gradually occupying a greater proportion in Eq. (8), the term $\sum_{l=1}^{L-1} h(q_l(\boldsymbol{x}))$ plays an increasingly important role in spatial context learning. As a result, our residual recurrent network captures the contextual information from a sequence through the term $\sum_{l=1}^{L-1} h(q_l(\boldsymbol{x}))$ in an elegant manner, making the text recognition process more efficient and reliable than processing each character independently. Finally, the residual recurrent network significantly promotes system performance while not adding extra parameter or computational burden to the system.

4.1.3 Transcription

Connectionist temporal classification (CTC), which facilitates the use of FCN and LSTM for sequential training without requiring any prior alignment between input images and their corresponding label sequences, is adopted as the transcription layer in our framework. Let C represent all the characters used in this problem and let "blank" represent the null emission. Then, the character set can be denoted as $C' = C \cup \{blank\}$. Given input sequences $\boldsymbol{u} = (u_1, u_2, \cdots, u_T)$ of length T, where $u_t \in R^{|C'|}$, we can obtain an exponentially large number of label sequences of length T, refered to as alignments $\boldsymbol{\pi}$, by assigning a label to

each time step and concatenating the labels to form a label sequence. The probability of alignments is given by

$$p(\boldsymbol{\pi}|\boldsymbol{u}) = \prod_{t=1}^{T} p(\pi_t, t|\boldsymbol{u}). \tag{9}$$

Alignments can be mapped onto a transcription $\boldsymbol{l}$ by applying a sequence-to-sequence operation $\mathscr{B}$, which first removes the repeated labels and then removes the blanks. For example, "tree" can be obtained by $\mathscr{B}$ from "_tt_r_ee_e" or "_t_rr_e_eee_". The total probability of a transcription can be calculated by summing the probabilities of all alignments that correspond to it:

$$p(\boldsymbol{l}|\boldsymbol{u}) = \sum_{\boldsymbol{\pi}:\mathscr{B}(\boldsymbol{\pi})=\boldsymbol{l}} p(\boldsymbol{\pi}|\boldsymbol{u}). \tag{10}$$

As suggested by Graves and Jaitly [56], since the exact position of the labels within a particular transcription cannot be determined, we consider all the locations where they could occur, thereby allowing a network to be trained via CTC without pre-segmented data. A detailed forward-backward algorithm to efficiently calculate the probability in Eq. (10) was proposed by Graves [19].

4.2 Learning Multiple Spatial Contexts with Multi-Spatial-Context FCRN (MC-FCRN)

In recent years, the concept of introducing contextual information within sequential data by using recurrent neural network is gaining popularity [12], [13], [21], [36]. However, very few studies have focused on incorporating multi-scale context information into sequential recognition problem. Given the fact that multi-scale strategies, such as SIFT [41], pyramid match kernel [40], SPP [42] and GoogLeNet [43], bring significant improvement in different area, we consider it of great advantage to introduce the multi-scale strategy into the sequential problems, such as OHCTR. Therefore we propose a novel architecture, which we refer to as multi-spatial context FCRN, to learn to model the multi-spatial context within the sequential online handwritten text data. In the following section, we will introduce MC-FCRN progressively from a basic model, spatial context learning to multi-spatial context learning.

4.2.1 Basic Model

First, we introduce a simple model having three components: FCN (denoted by f with parameters $\boldsymbol{\omega}_f$), fully connected layers (denoted by g with parameters $\boldsymbol{\omega}_g$), and CTC. Given signature feature maps $\boldsymbol{x} = (x_1, x_2, \cdots, x_T)$, where x_t represents the receptive field of the t-th time step, the objective of this model is to learn to make a prediction for each receptive field:

$$\begin{aligned} &o(z_t, x_t) = g(z_t, f(x_t)) \\ &s.t. \quad \begin{cases} \sum_{i=1}^{|C'|} g(z_t = C'_i, f(x_t)) = 1, \\ g(z_t, f(x_t)) > 0. \end{cases} \end{aligned} \tag{11}$$

where z_t represents the prediction of the t-th time step and $o(z_t, x_t)$ models the probability distribution over all the words in C' given the receptive field x_t in the t-th time step. Since each frame in the FCN feature sequence represents a distribution over the character set without considering any other feature vector, this simple model can hardly incorporate any spatial context for recognition.

4.2.2 Learning Spatial Context

LSTM inherently possesses the advantage of processing sequential data; thus, it is a good choice for capturing spatial contextual information within an FCN feature sequence. Specially, we use the proposed residual recurrent network for spatial context learning and develop the proposed FCRN by stacking the residual recurrent network right after FCN. More importantly, the input of the residual recurrent network is now the output of FCN, i.e. $q_0(\boldsymbol{x}) = (f(x_1), f(x_2) \dots f(x_T))$. Now based on Eq. (8), the objective of the FCRN is to learn to make a prediction for each time step given the entire input signature feature maps:

$$\begin{aligned} &o(z_t, \boldsymbol{x}) = g(z_t, q_L^t(\boldsymbol{x})) \\ &s.t. \quad \begin{cases} \sum_{i=1}^{|C'|} g(z_t = C'_i, q_L^t(\boldsymbol{x})) = 1, \\ g(z_t, q_L^t(\boldsymbol{x})) > 0. \end{cases} \end{aligned} \tag{12}$$

where the overall system has parameters $\boldsymbol{\omega} = (\boldsymbol{\omega}_f, \boldsymbol{\omega}_h, \boldsymbol{\omega}_g)$, $q_L^t(\boldsymbol{x})$ is the t-th time step of the output feature sequence of the residual recurrent network, and $o(z_t, \boldsymbol{x})$ models the probability distribution over all the words in C' in the t-th time step given the entire input signature feature maps $\boldsymbol{x}$. When we remove the term $\sum_{l=1}^{L-1} h(q_l(\boldsymbol{x}))$ from Eq. (8), Eq. (12) simply reduce to Eq. (11). Therefore, the residual recurrent network plays a key role in learning spatial context information in FCRN.

4.2.3 Learning Multiple Spatial Contexts

Before moving toward learning multiple spatial contexts, we should first introduce the concept of receptive field and describe its role in learning multi-spatial context. A receptive field is a rectangular local region of input images that can be properly represented by a highly abstract feature vector in the output feature sequence of FCN. Let r_l represent the local region size (width/height) of the l-th layer, and let the (x_l, y_l)-coordinate denote the center position of this local region. Then, the relationship of r_l and (x_l, y_l)-coordinate between adjacent layers can be formulated as follows:

$$\begin{aligned} r_l &= (r_{l+1} - 1) \times m_l + k_l, \\ x_l &= m_l \times x_{l+1} + (\frac{k_l - 1}{2} - p_l), \\ y_l &= m_l \times y_{l+1} + (\frac{k_l - 1}{2} - p_l), \end{aligned} \tag{13}$$

where k is the kernel size, m is the stride size, and p is the padding size of a particular layer. Recursively applying Eq. (13) to adjacent layers in the FCN from the last response maps down to the original image should yield the region size and the center coordinate of the receptive field that corresponds to the related feature vector of the FCN feature sequence.

Technically, the receptive fields of an ordinary fully convolutional network are of the same scale, except for some specially designed networks, such as GooLeNet. However,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2732978, IEEE Transactions on Pattern Analysis and Machine Intelligence

unconstrained handwritten Chinese text has severe recognition problem owing to its large character set, variable writing style and ambiguous segmentation problem. Therefore, in OHCTR problem, it is significantly important to observe a specific local region of the original input image with a series of receptive fields of different scales.

In the following, we explain how to generate receptive fields with different scales for each time step. We observe that the size of the receptive field is sensitive to the kernel size. Assume that the kernel size is increased from k_l to $k_l+\Delta k_l$. We can derive the following mapping from Eq. (13): $r_{l-1} = r'_{l-1} + \Delta k_l \times m_{l-1}$, where $r'_{l-1}$is the original region size of the $(l-1)$-th layer. Thus, we have

$$r_0 = r'_0 + \Delta k_l \times \prod_{i=1}^{l-1} m_i. \tag{14}$$

In other words, if we increase the kernel size of the l-th layer by Δk_l, then the receptive field will be enlarged by $\Delta k_l \times \prod_{i=1}^{l-1} m_i$. As shown in Fig. 1 and Fig. 5, our MC-FCRN split into three subnetworks after sharing the first four convolutional layer. These three subnetworks have increasing numbers of convolutional layers, leading to increasingly larger scale of receptive field (see Fig. 3). Further, note that when $k_l = 2p_l + 1$ and $m_l = 1$, the center position (i.e., (x_l, y_l)-coordinate) of the receptive field does not change from higher layers to lower layers. Therefore, receptive fields with different scales in the same time step have the same center position, which ensures that multiple spatial contexts are incorporated while confusion is avoided.

There are different ways to fuse the corresponding feature vectors of these multi-scale receptive fields. Typically, we can simply concatenate or add the vectors before the residual recurrent network. However, we found that the system can better benefit from multiple spatial contexts when fusing after the residual recurrent network. Let $\boldsymbol{q}(\boldsymbol{x})$ represent the concatenation of the output feature sequences $(\{q_L(\boldsymbol{x})\}_1, \{q_L(\boldsymbol{x})\}_2, \{q_L(\boldsymbol{x})\}_3, \cdots)$ of the residual recurrent network with receptive fields of different scales. As shown in Fig. 3, at the t-th time step, $\boldsymbol{q}^t(\boldsymbol{x}) = (\{q_L^t(\boldsymbol{x})\}_1, \{q_L^t(\boldsymbol{x})\}_2, \{q_L^t(\boldsymbol{x})\}_3, \cdots)$ represents the extracted features of receptive fields of increasing scales; this is where the multi-spatial context comes. Formally, the objective of our MC-FCRN is to learn to make a prediction for each time step given the entire input signature feature maps:

$$o(z_t, \boldsymbol{x}) = g(z_t, \boldsymbol{q}^t(\boldsymbol{x})) \tag{15}$$
$$s.t. \quad \begin{cases} \sum_{i=1}^{|C'|} g(z_t = C'_i, \boldsymbol{q}^t(\boldsymbol{x})) = 1, \\ g(z_t, \boldsymbol{q}^t(\boldsymbol{x})) > 0. \end{cases}$$

where $o(z_t, \boldsymbol{x})$ models the probability distribution over all the words in C' in the t-th time step given the entire input signature feature maps $\boldsymbol{x}$.

Now, suppose $p(\pi_t, t|\boldsymbol{x}) = g(z_t = \pi_t, \boldsymbol{q}^t(\boldsymbol{x}))$. By Eq. (9), the probability of alignments can be represented by:

$$p(\boldsymbol{\pi}|\boldsymbol{x}) = \prod_{t=1}^{T} g(z_t = \pi_t, \boldsymbol{q}^t(\boldsymbol{x})). \tag{16}$$

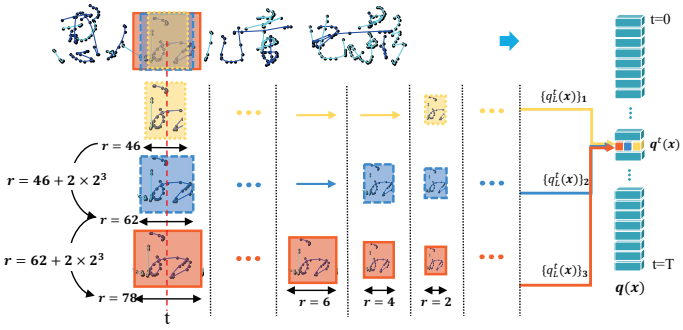


Fig. 3. Illustration of multiple spatial contexts. Different receptive fields in the same time step have the same center position, and their region sizes should satisfy Eq. (14).

Then, the total probability of a transcription can be calculated by applying Eq. (16) to Eq. (10):

$$p(\boldsymbol{l}|\boldsymbol{x}) = \sum_{\boldsymbol{\pi}:\mathscr{B}(\boldsymbol{\pi})=\boldsymbol{l}} \prod_{t=1}^{T} g(z_t = \pi_t, \boldsymbol{q}^t(\boldsymbol{x})). \tag{17}$$

The training is achieved by searching for $\boldsymbol{\omega}$ that minimizes the negative penalized log-likelihood:

$$L(Q) = -\sum_{(\boldsymbol{x},\boldsymbol{l})\subset Q} ln\{\sum_{\boldsymbol{\pi}:\mathscr{B}(\boldsymbol{\pi})=\boldsymbol{l}} \prod_{t=1}^{T} g(z_t = \pi_t, \boldsymbol{q}^t(\boldsymbol{x}); \boldsymbol{\omega})\} + R(\boldsymbol{\omega}) \tag{18}$$

where $\boldsymbol{l}$ is the label sequence, Q represents the training set, and $R(\boldsymbol{\omega})$ denotes the regularization term. In our experiment, R is a weight decay penalty implemented with $L2$ regularization.

4.2.4 Discussion on MC-FCRN

Multi-scale strategy has been widely used in computer vision, as it encourages analyzing the image from coarser to finer levels and aggregates local feature in them [40], [41], [42], [43]. Nevertheless, it has not been considered for sequence labeling problem, like OHCTR. The proposed MC-FCRN is a novel solution that introduces multi-scale strategy into OHCTR problem and has the following remarkable properties:

- For handwritten text recognition problem, normalization on the text level easily results in characters of variable scales, no mentioning the large amount of characters, cursive writing styles and ambiguous segmentation problem. With hierarchical-scale receptive fields, MC-FCRN can generate informative feature that is robust and insensitive to the complex and cursive handwritten Chinese handwritten text.
- MC-FCRN explicitly keeps different receptive fields of the same time step aligning at the same center position, as illustrated in Fig. 3, thus it introduces the multi-scale spatial context elegantly while avoiding confusion between adjacent time steps, which consequently avoids oscillation problem from gradient explosion during optimization.
- MC-FCRN is the first attempt to introduce multi-scale spatial context for OHCTR sequence labeling problem. It provides a simple yet effective way to deal with the

complicate handwritten Chinese text by enabling the system to 'observe' the signature feature maps from multi-scale perspectives. Besides, it is very easy to extend or eliminate one specific 'scale' in MC-FCRN to keep the trade-off between complexity and efficiency.

- The residual recurrent network of MC-FCRN not only substantially accelerates the convergence procedure but also promotes the performance significantly, while adding neither extra parameter nor computational burden to the system.

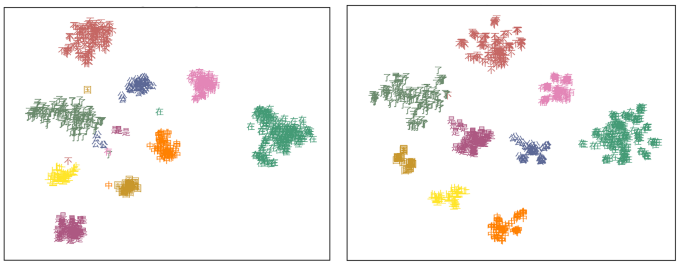
5 Implicit Language Modeling

We say that a system is an implicit language model (implicit LM) if it does not directly learn the conditional probabilities of the next word given previous words, but implicitly incorporates lexical constraints and prior knowledge about the language to improve the system performance. The network architecture of our implicit LM consists of three components: the embedding layer, the language modeling layer, and the transcription layer. Given the predicting feature sequence $\boldsymbol{s} = (s_1, s_2, \cdots, s_T)$ from the multi-spatial-context FCRN, the objective of the implicit LM is to learn to make a prediction for each time step given the entire input sequence:

$$f(z_t, \boldsymbol{s}) = U(z_t, (M(s_1), M(s_2), \cdots, M(s_T))) \tag{19}$$
$$s.t. \quad \begin{cases} \sum_{i=1}^{|C'|} f(z_t = C'_i|\boldsymbol{s}) = 1, \\ f(z_t|\boldsymbol{s}) > 0. \end{cases}$$

where $f(z_t, \boldsymbol{s})$ models the probability distribution over all the words in C' in the t-th time step given the entire predicting feature sequence $\boldsymbol{s}$, while the mapping M and the probability function U represent two successive processing stages that constitute the prediction procedure of the implicit LM.

In the first stage, the mapping M, implemented by the embedding layer, translates the input sequence $\boldsymbol{s} = (s_1, s_2, \cdots, s_T)$ into real vectors $(M(s_1), M(s_2), \cdots, M(s_T))$, where $M(s_t) \in R^m$. Note that the mapping M differs from the mapping C [45] in traditional neural language models, because the embedding mapping used here takes the input of a predicting feature vector, not just a one-hot vector. It is noteworthy to mention that the predicting feature vector here is highly related to character, playing a very similar role to a one-hot vector. In Fig. 5, we visualize 10 categories of characters with their predicting feature vector extracted from the predicting sequence of MC-FCRN. From Fig. 5, we can observe the following properties of the predicting feature vectors: (1) Although the predicting features of the same character are different, they cluster together in the feature space, maintaining distinct distance from other characters. Such a phenomenon suggests that the predicting feature can represent which character it probably is. (2) There should exits a statistical central position for each character in the feature space. The predicting features of the same character do not have to distribute around this central position strictly, so as to represent the intra-concept variability [47]. Their distances to the central position of a specific character reflect the confidence information of being that character. More importantly, since our predicting feature vector do not neglect such confidence information (compared to the one-hot vector), our implicit LM can take advantage of it as well as semantic context to rectify the misclassified characters, which has the same purpose when we decode with statistical language model traditionally.



(a) (b)

Fig. 4. t-SNE visualization of the predicting feature vector of some typical characters extracted from the output sequence of MC-FCRN (left) and implicit LM (right). After applying implicit LM, the system manages to rectify most of the unreasonably distributed feature vector by incorporating semantic context information.

In the second stage, the probability function U, maps the embedding vectors for words in context $(M(s_1), M(s_2), \cdots, M(s_T))$ to a conditional probability distribution over all the words in C', i.e., the i-th element of the output vector of U estimates the probability $p(z_t = C'_i|(M(s_1), M(s_2), \cdots, M(s_T)))$. Then, we can represent the function $f(z_t, \boldsymbol{s})$ by the composition of mappings M and U as follows:

$$f(z_t = C'_i, \boldsymbol{s}) = U(C'_i, (M(s_1), M(s_2), \cdots, M(s_T))) \tag{20}$$

In this paper, we implemented the embedding layer as a fully connected layer that can be represented by a $|C'| \times m$ matrix $\boldsymbol{\theta}_M$. The function U with parameters $\boldsymbol{\theta}_U$ is implemented by multilayered BLSTM for learning long-term context from the forward and reverse directions. The overall parameters for the implicit LM f are given by $\boldsymbol{\theta} = (\boldsymbol{\theta}_M, \boldsymbol{\theta}_U)$.

Now, suppose $p(\pi_t, t|\boldsymbol{s}) = f(z_t = \pi_t|\boldsymbol{s})$. By Eq. (9), the probability of alignments can be represented by:

$$p(\boldsymbol{\pi}|\boldsymbol{s}) = \prod_{t=1}^{T} f(z_t = \pi_t|\boldsymbol{s}). \tag{21}$$

Then, the total probability of a transcription can be calculated by applying Eq. (21) to Eq. (10):

$$p(\boldsymbol{l}|\boldsymbol{s}) = \sum_{\boldsymbol{\pi}:\mathscr{B}(\boldsymbol{\pi})=\boldsymbol{l}} \prod_{t=1}^{T} f(z_t = \pi_t|\boldsymbol{s}). \tag{22}$$

The training is achieved by searching for $\boldsymbol{\theta}$ that minimizes the negative penalized log-likelihood:

$$L(Q) = -\sum_{(\boldsymbol{s},\boldsymbol{l})\subset Q} ln\{\sum_{\boldsymbol{\pi}:\mathscr{B}(\boldsymbol{\pi})=\boldsymbol{l}} \prod_{t=1}^{T} f(z_t = \pi_t|\boldsymbol{s};\boldsymbol{\theta})\} + R(\boldsymbol{\theta}) \tag{23}$$

where $R(\boldsymbol{\theta})$ is the regularization term.

5.1 Interesting Properties of Implicit Language Model

- Our implicit LM offers the unique advantage of leveraging semantic context from both directions of the text,

significantly outperforming language models that only predict the conditional probability of the next word given previous words in one direction. Furthermore, because LSTM can capture long-term complicated dynamics in the sequence, our implicit LM has the potential to learn semantic context from the entire sequence to enhance recognition performance.

- The predicting feature sequence contains information that indicates not only the predicted labels but also the confidence about their prediction, providing much more information than a simple one-hot vector [45]. Thus, the implicit LM is able to improve network performance by exploiting the confidence information of the predicted labels in addition to their semantic context knowledge.
- As shown in the experiments, implicit LM has significant advantage over statistical LM in both decoding speed and prediction accuracy, exhibiting great potential in practical applications. Furthermore, unlike statistical LM whose RAM size increases as the corpus grows, the implicit LM has a fixed size that is determined by its network architecture, making implicit LM an ideal alternative option for the statistical LM in the case of a large corpus.

5.2 Training Strategy

In the process of implicit LM training, we do not update the parameters of MC-FCRN. Given a training instance $(\boldsymbol{x}, \boldsymbol{l})$, we feed the fixed-parameter MC-FCRN with the signature feature maps $\boldsymbol{x}$ to obtain a predicting feature sequence of length T. Then, the feature sequence with label $\boldsymbol{l}$ is used to train the implicit LM. Note that the training set for the implicit LM should contain semantic information, i.e., the characters should be understandable in context. In fact, we synthesized the text samples based on the corpora and isolated characters in CASIA1.0-1.2 [57] to train implicit LM and completely ignored the real training examples in CASIA2.0-2.2 [57]. However, our training set for MC-FCRN training does not contain any semantic knowledge. Actually, we shuffle the order of the characters in each training instance to achieve this effect. There are two main reasons for using different training strategies during the training procedure of MC-FCRN and the implicit LM. First, our MC-FCRN and implicit LM can concentrate on learning spatial context and semantic context, respectively. If we directly learn spatial-semantic context in a unified network, then such a network may heavily overfit the context information of the training set. Second, because the training set of CASIA2.0-2.2 has the same corpus as the test set, we should not use the samples from the training set directly for training, as it may lead to unfair comparison with the results of other methods.

5.3 Statistical Language Model

In the post-processing procedure, the language model plays a significant role in decoding the prediction sequence [2], [3], [4]. The decoding algorithm proposed by Graves et al. [56] is adopted in our experiments to incorporate the traditional statistical language model.

6 Experiments

To evaluate the effectiveness of the proposed system, we conducted experiments on the standard benchmark dataset CASIA-OLHWDB [57] and the ICDAR2013 Chinese handwriting recognition competition dataset [58] for unconstrained online handwritten Chinese text recognition.

6.1 Databases

In the following experiments, we used the training set of CASIA-OLHWDB [57], including both unconstrained text lines and isolated characters, as our training data. The training set of CASIA2.0-2.2 (one subset of CASIA-OLHWDB for OHCTR problem) contains 4072 pages of handwritten texts, with 41,710 text lines, including 1,082,220 characters of 2650 classes. We randomly split the training set into two groups, with approximately 90% for training and the remainder for validation and further parameter learning for language modeling. The isolated training character samples (totally 3,129,496 character samples) of CASIA1.0-1.2 (one subset of CASIA-OLHWDB) were also used to construct synthetic text data for system optimization. Two popular benchmark datasets for unconstrained online handwritten Chinese text recognition were used for performance evaluation, i.e., the test set of CASIA2.0-2.2 (Dataset-CASIA) and the test set of the online handwritten Chinese text recognition task of the ICDAR 2013 Chinese handwriting recognition competition [58] (Dataset-ICDAR). Dataset-CASIA contains 1020 text pages, including 268,924 characters of 2626 classes, while Dataset-ICDAR contains 3432 text lines, including 91,576 characters of 1375 classes. Note that for general-pupose recognition and fair comparison with previous work [2], [3], [4], our system had 7356 classes and was trained using not only handwritten texts of CASIA2.0-2.2 but also synthetic text data based on isolated characters from CASIA-OLHWDB.

For language modeling, we conducted experiments using both the implicit LM and the statistical language model. Three corpora were used in this paper: the PFR corpus [59], which contains news text of 2,199,492 characters from the 1998 People's Daily corpus; the PH [60] corpus, which contains news text of 3,697,028 characters from the People's Republic of China's Xinhua news recorded between January 1990 and March 1991; and the CLDC corpus [61], which contains contemporary corpus of approximately 50 million characters in 7,356 classes collected by the Institute of Applied Linguistics. For statistical language modeling, we used the SRILM toolkit [62] to build our language model.

6.2 Experimental Setting

The detailed architecture of our MC-FCRN and implicit LM is shown in Fig. 5. Batch Normalization [63] was applied after all but the first two convolutional layers in order to achieve faster convergence and avoid over-fitting. As the recognition system consists of more class categories (7356) than that of the training set of CASIA2.0-2.2 (2650), the training procedure of MC-FCRN is divided into two stages. In the first stage, synthetic text data based on the isolated characters are taken for network optimization. When the network reaches convergence, it is finetuned on the real

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2732978, IEEE Transactions on Pattern Analysis and Machine Intelligence

samples from CASIA2.0-2.2. Note that, the order of characters in each text sample was randomly shuffled to discard the semantic context for fairness. To accelerate the training process, our network was trained with shorter texts segmented from text lines in the training data, which could be normalized to the same height of 126 pixels while retaining the width at fewer than 576 pixels. In the test phase, we maintained the same height but increased the width to 2400 pixels in order to include the text lines from the test set.

We constructed our experiments within the CAFFE [64] deep learning framework, in which LSTM is implemented following the approach of Venugopalan et al. [21] while the other processes are contributed by ourselves. Further, we used AdaDelta as the optimization algorithm with $\rho = 0.9$. The experiments were conducted using GeForce Titan-X GPUs. For performance evaluation, we used the correct rate (CR) and accuracy rate (AR) as performance indicators, as specified in the ICDAR 2013 Chinese handwriting recognition competition [65].

6.3 Experimental Results

6.3.1 Effect of Path Signature

Table 1 summarizes the recognition results of FCRN with path signature for different truncated levels (Sig0, Sig1, Sig2, and Sig3). Sig0 implies that only the $k = 0$ iterated integral is considered in the experiments, Sig1 implies that the $k = 0$ and $k = 1$ iterated integrals are considered in the experiments, and so on for Sig2 and Sig3. The experiments showed that the system performance improves monotonically from 85.14% to 87.94% on Dataset-ICDAR and from 89.58% to 92.22% on Dataset-CASIA as the path signature increases from Sig0 to Sig3. This proves the effectiveness of applying path signature to the OHCTR problem. Such results are obtained because the path signature captures more essential information from the pen-tip trajectories with higher iterated integrals. We also observed that the performance improvement slows down as the iterated integrals increase, because the iterated integral of the path increases rapidly in dimension with severe computational burden while carrying very little information. We also compare the path signature with previous state-of-the-art feature, i.e., the 8-directional feature [29]. As listed in Table 1, it can be seen that all Sig1-Sig3 outperform the 8-Dir feature. As Sig2 achieves a reasonable trade-off between efficiency and complexity, we selected it for feature extraction in the following experiments.

6.3.2 Effect of Multiple Spatial Contexts

In our system, the residual recurrent network acts as the basic component for spatial context learning. As listed in Table 2, the residual recurrent network clearly has advantages over the conventional recurrent network for both unidirectional and bidirectional LSTM. Specifically, we drew the curves of the correct rate for all these four networks for the first training stage (see Section 6.2) on the validation set in Fig. 6. The comparison of the curves conveys two vital messages. First, the residual recurrent network substantially accelerates the convergence procedure from the very beginning, e.g., the first five epoches. This is because the recurrent network had not yet functioned during that period; thus FCN network can be optimized directly with CTC loss function through the residual connection. Second, as time progressed, the recurrent network gradually augments its impact on the recognition result by increasingly incorporating more spatial contextual information in an elegant manner. Therefore, the FCRN with the residual recurrent network tends to achieve superior results, and

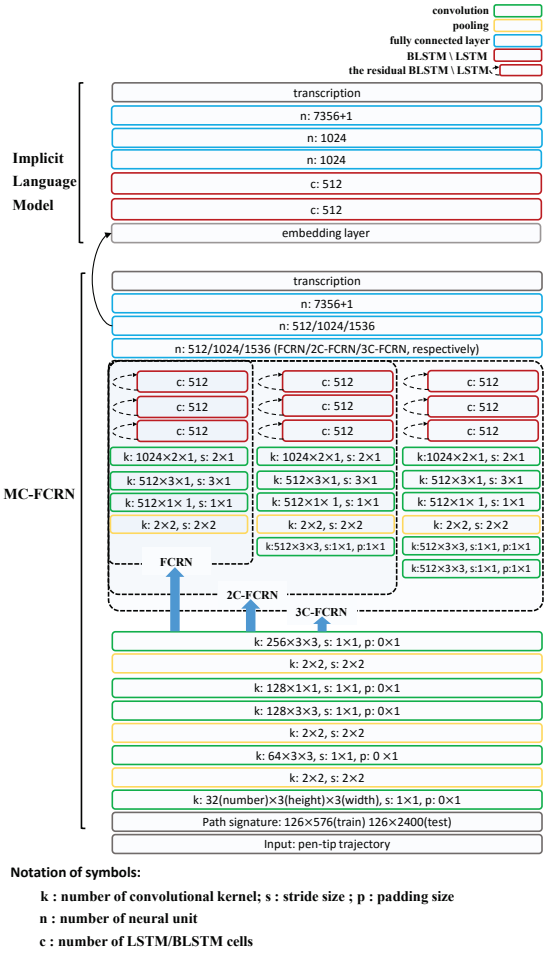


Fig. 5. Illustration of network architecture of MC-FCRN.

TABLE 1
Effect of Path Signature (Percent)

Path signatures	Feature maps	Dataset-ICDAR CR	Dataset-ICDAR AR	Dataset-CASIA CR	Dataset-CASIA AR
Sig0	1	85.14	83.60	89.58	87.67
Sig1	3	86.83	85.82	91.47	90.54
Sig2	7	87.82	86.85	92.29	91.38
Sig3	15	87.94	87.20	92.22	91.58
8-Dir [29]	8	85.71	83.32	90.48	87.82
8-Dir+Sig0	9	87.67	86.52	92.05	91.00

TABLE 2
Effect of Residual Recurrent Network (Percent)

Architecture	Dataset-ICDAR CR	Dataset-ICDAR AR	Dataset-CASIA CR	Dataset-CASIA AR
LSTM	85.67	84.60	90.14	89.17
Residual LSTM	87.82	86.85	92.29	91.38
BLSTM	87.96	87.28	92.69	92.15
Residual BLSTM	89.24	88.32	93.66	92.96

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2732978, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3
Effect of Spatial Context (Percent)

System	Foot print MB (RAM)	Dataset-ICDAR			Dataset-CASIA		
		runtime	CR	AR	runtime	CR	AR
FCRN	49.8	108s	87.82	86.85	163s	92.29	91.38
2C-FCRN	109.5	211s	90.17	88.88	341s	94.47	93.31
3C-FCRN	181.8	299s	90.76	89.52	458s	94.72	93.74
FCRN-2	109.5	207s	89.26	88.37	342s	93.61	92.83
FCRN-3	181.8	302s	89.79	88.76	455s	94.02	93.23
GoogLeNet (6 inception)	60.5	165s	80.72	79.27	290s	83.08	81.86
GoogLeNet (9 inception)	81.0	179s	72.20	71.39	311s	76.81	75.94

TABLE 4
Effect of Semantic Context (Percent)

System		Foot print MB (RAM)	Dataset-ICDAR			Dataset-CASIA		
			runtime	CR	AR	runtime	CR	AR
FCRN(Baseline)	w.o. LM	49.8	1.8min	87.82	86.85	2.7min	92.29	91.38
implicit LM	PH	65.4	1.7 ±0.5min	93.92	93.34	5 ±0.5min	94.26	93.74
	PFR			94.21	93.69		94.43	93.89
	CLDC			95.33	94.52		96.39	95.80
statistical LM	PH	11.1	130 ±5min	91.78	90.70	360 ±10min	94.22	93.11
	PFR	6.6		91.93	90.88		94.27	93.10
	CLDC	576.3		92.67	91.81		96.09	95.50
implicit LM + statistical LM	CLDC	641.7	132±5min	95.79	95.04	365±10min	96.81	96.27

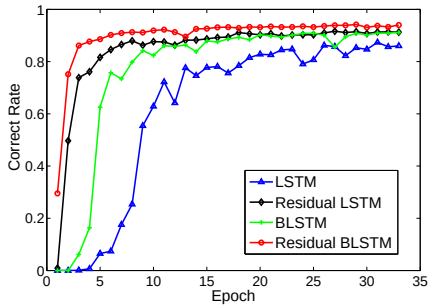


Fig. 6. Curves of correct rate of (residual) recurrent network on validation set for the first training stage (refer to Section 6.2).

more importantly, with less oscillation.

The effects of multiple spatial contexts are summarized in Table 3, and the network architectures of FCRN, 2C-FCRN, and 3C-FCRN are shown in Fig. 5. From Fig. 5, we can see that FCRN, 2C-FCRN, and 3C-FCRN have one, two, and three receptive fields of different scales for each time step, respectively. The experiments showed that the system performance improved monotonically for both Dataset-ICDAR and Dataset-CASIA in the order of FCRN, 2C-FCRN, and 3C-FCRN, suggesting that we successfully leveraged the multiple spatial contexts by using multiple receptive fields and improved the system performance. Furthermore, we designed FCRN-2 and FCRN-3 such that their architectures and sizes were similar to those of 2C-FCRN and 3C-FCRN, except that their receptive fields for each time step were of the same scale. As listed in Table 3, although FCRN-3 does benefit from increased parameter number, its performance is even lower than that of 2C-FCRN, which further verifies the significance of the additional spatial context.

The inception mechanism of GoogLeNet, a similar but different way to leverage multiple spatial contexts, demonstrated outstanding ability in image classification and object detection task. Therefore, we replaced the fully convolutional network with GoogLeNet using 6 inception layers and 9 inception layers, respectively, with proper customization to maintain a reasonable length of the output feature sequence. However, as listed in Table 3, incorporating the inception does not yield sufficiently good results. There are two reasons to explain this phenomenon. First, in MC-FCRN, we carefully maintain different receptive fields of the same time step at the same center position (as illustrated in Fig.3), while the inception mechanism of GoogLeNet encourages different scales of convolution kernel to fuse together, and stack upon one another. The stacked inception layers did perform well in the image-based classification problem, but caused confusion between adjacent time steps, thus failing in the sequence labeling problem, like OHCTR. Second, the gradient explosion problem frequently occurs during the training of integrated GoogLeNet-LSTM network. Note that in this paper, we have already used multiple ways to avoid the loss oscillation problem, such as using Bath Normalization strategy and residual recurrent network, but oscillation still occurs with GoogLeNet-LSTM during optimization.

6.3.3 *Effect of Semantic Context*

To evaluate the effectiveness of the implicit LM, we conducted experiments based on FCRN text line recognizer and three different corpora, i.e., the PFR, PH and CLDC corpus. As listed in Table 4, both the implicit LM and statistical language model substantially improved the system performance on both Dataset-ICDAR and Dataset-CASIA. In the table, we also observe the superiority of the implicit LM over the statistical language model, especially on Dataset-ICDAR. We attribute this superiority to the potential ability of the implicit LM to learn semantic context from the entire sequence as well as the confidence information of the predicted words. Furthermore, the evaluation time for Dataset-ICDAR and Dataset-CASIA testing data is provided as runtime term on the table. With the maximum number of the prefix paths for beam search on the statistical language model set to 100, the proposed implicit LM is approximately 70 times faster than the statistical language model, thus exhibiting a significant advantage for practical applications. Furthermore, unlike the statistical language model where its RAM size increases as the corpus grows, the implicit LM has

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2732978, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 5
Comparison with Previous Methods Based on Correct Rate and Accuracy Rate (Percent) for Dataset-ICDAR and Dataset-CASIA

Method	Dataset-ICDAR				Dataset-CASIA			
	w.o. LM		with LM		w.o. LM		with LM	
	CR	AR	CR	AR	CR	AR	CR	AR
Shi et al., 2016 [12]	85.14	83.60	-	-	89.58	87.67	-	-
Wang et al., 2012 [1]	-	-	-	-	-	-	92.76	91.97
Zhou et al., 2013 [3]	-	-	94.62	94.06	87.93	85.92	94.34	93.75
Zhou et al., 2014 [4]	-	-	94.76	94.22	-	-	95.32	94.69
VO-3 [58]	-	-	95.03	94.49	-	-	-	-
2C-FCRN (residual LSTM) + CLDC (implicit LM)	90.17	88.88	96.01	95.46	94.47	93.31	97.07	96.72
2C-FCRN (residual LSTM) + CLDC (statistical LM)	90.17	88.88	94.51	93.45	94.47	93.31	97.01	96.49
2C-FCRN (residual LSTM) + CLDC (implicit LM & statistical LM)	90.17	88.88	**96.58**	**96.09**	94.47	93.31	**97.50**	**97.23**
Sun et al., 2016 [66] (2765 classes)	90.18	89.12	94.43	93.40	95.82	95.30	97.55	97.05
3C-FCRN (2651 classes)	93.53	92.86	**97.15**	**96.50**	95.50	94.73	**97.75**	**97.31**

Upper: 7356 classes; Lower: less than 3000 classes.

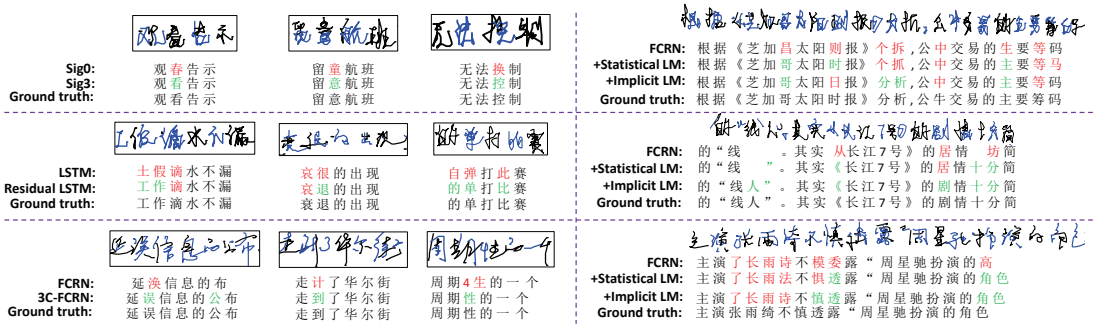


Fig. 7. Unconstrained handwritten Chinese text lines with their corresponding labels and the result predicted by the systems.

a fixed size that is determined by its network architecture. Therefore, the implicit LM is an ideal alternative option for the statistical language model in the case of a large corpus, e.g. CLDC.

It should be noted that the implicit LM requires a significant amount of time for training in the experiments. With the small corpus of PH and PFR, it takes about 10 days to achieve good convergence with the simple FCRN recognizer. As for CLDC corpus that has at least 12 times more corpus than PH and PFR, we spent approximately 9 weeks for optimizing implicit LM to outperform statistical LM. Although its training process is time-consuming, in return, implicit LM provides more accurate and much faster predictions in the test phase, exhibiting great potential in practical applications. As listed in Table 4, the system performance can be further improved by jointly applying both the implicit LM and the statistical language model, which further verifies the complementarity between them.

6.3.4 Comparison with Published State-of-the-art Methods

The methods of Wang et al. [1], Zhou et al. [3] [4] and VO-3 [58] were all based on the segmentation strategy, which are different from our segmentation-free MC-FCRN that incorporate the recently developed FCN, LSTM, and CTC. As listed in Table 5, our proposed method, 2C-FCRN with the implicit LM trained on corpus CLDC, demonstrates superior performance to state-of-the-art results on both Dataset-ICDAR and Dataset-CASIA. Furthermore, we integrate the implicit LM with the statistical language model to leverage the complementarity of them and the results are listed in Table 5. It can be observed that our method significantly outperforms the best results on Dataset-ICDAR and Dataset-CASIA, with a relative error reduction of 29.04% and 47.83% on accuracy rate, respectively. Unlike the above-mentioned systems that have 7356 classes, Sun et al. [66] applied deep stacked LSTM with only 2675 classes to tackle the OHCTR problem. For a relatively fair comparison, we reduced our network category so that it had 2650 classes, similar to the training set of CASIA2.0-CASIA2.2. The experiments shows that our network have a much more balanced and better results on both Dataset-ICDAR and Dataset-CASIA compared to Sun et al. [66].

6.4 Error Analysis

In this section, we offer typical examples, as shown in Fig. 7, to show the effectiveness of our method. The left-top examples shown in Fig. 7 demonstrate that sig3 takes advantage of online information to recognize ambiguous characters. The left-middle examples show that the residual recurrent network improves the optimization of the network. The left-bottom examples exhibit that MC-FCRN has a strong capability of capturing spatial context information for recognition. Finally, as shown in the right panel of the figure, the implicit LM exhibit better capability of leveraging semantic context information as well as confidence information of predicted words for recognition, as compared to the statistical language model.

7 Conclusion

In this paper, we addressed the challenging problem of unconstrained online handwritten Chinese text recognition by proposing a novel system that incorporates path signature, a multi-spatial-context fully convolutional recurrent network (MC-FCRN), and an implicit language model. We exploited the spatial structure and online information of pen-tip trajectories with a powerful path signature. Experiments showed

that the path signature truncated at level two achieves a reasonable trade-off between efficiency and complexity for OHCTR problem. For spatial context learning, we presented the residual recurrent network to accelerate the convergence process and improve the optimization results without introducing extra parameters or computational burden to the system. For multi-spatial context learning, we demonstrated that our MC-FCRN successfully exploits multiple spatial contexts from receptive fields with multiple scales to robustly recognize the input signature feature maps. For semantic context learning, an implicit LM was developed to learn to make predictions conditioned on the entire predicting feature sequence, significantly improving the system performance. In the experiments, our best result significantly outperformed all other existing method on two standard benchmarks Dataset-ICDAR and Dataset-CASIA.

One limitation of the proposed implicit LM is that it involves too much training time in the case of large corpus. How to accelerate the training procedure of the RNN-based implicit LM is still a challenging problem and remains for further study.

REFERENCES

[1] D.-H. Wang, C.-L. Liu, and X.-D. Zhou, "An approach for real-time recognition of online chinese handwritten sentences," *Pattern Recognition*, vol. 45, no. 10, pp. 3661–3675, 2012.

[2] 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, 2012.

[3] X.-D. Zhou, D.-H. Wang, F. Tian, C.-L. Liu, and M. Nakagawa, "Handwritten chinese/japanese text recognition using semi-markov conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2413–2426, 2013.

[4] X.-D. Zhou, Y.-M. Zhang, F. Tian, H.-A. Wang, and C.-L. Liu, "Minimum-risk training for semi-markov conditional random fields with application to handwritten chinese/japanese text recognition," *Pattern Recognition*, vol. 47, no. 5, pp. 1904–1916, 2014.

[5] Y.-C. Wu, F. Yin, and C.-L. Liu, "Improving handwritten chinese text recognition using neural network language models and convolutional neural network shape models," *Pattern Recognition*, vol. 65, pp. 251–264, 2017.

[6] Y. Bengio, "Markovian models for sequential data," *Neural computing surveys*, vol. 2, no. 1049, pp. 129–162, 1999.

[7] T.-H. Su, T.-W. Zhang, D.-J. Guan, and H.-J. Huang, "Off-line recognition of realistic chinese handwriting using segmentation-free strategy," *Pattern Recognition*, vol. 42, no. 1, pp. 167–182, 2009.

[8] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[9] M. Liwicki, A. Graves, H. Bunke, and J. Schmidhuber, "A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks," *Proc. 9th Int. Conf. on Document Analysis and Recognition*, vol. 1, pp. 367–371, 2007.

[10] R. Messina and J. Louradour, "Segmentation-free handwritten chinese text recognition with lstm-rnn," *13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 171–175, 2015.

[11] Z. Xie, Z. Sun, L. Jin, Z. Feng, and S. Zhang, "Fully convolutional recurrent network for handwritten chinese text recognition," *CoRR*, vol. abs/1604.04953, 2016.

[12] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[13] P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang, "Reading scene text in deep convolutional sequences," 2016.

[14] D. K. Sahu and M. Sukhwani, "Sequence to sequence learning for optical character recognition," *CoRR*, vol. abs/1511.04176, 2015.

[15] G. Seni and E. Cohen, "External word segmentation of off-line handwritten text lines," *Pattern Recognition*, vol. 27, no. 1, pp. 41–52, 1994.

[16] M. Cheriet, N. Kharma, C.-L. Liu, and C. Suen, *Character recognition systems: a guide for students and practitioners*. John Wiley & Sons, 2007.

[17] J. Schenk and G. Rigoll, "Novel hybrid nn/hmm modelling techniques for on-line handwriting recognition," in *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006.

[18] S. Marukatat, T. Artières, P. Gallinari, and B. Dorizzi, "Sentence recognition through hybrid neuro-markovian modeling," *Sixth International Conference on Document Analysis and Recognition*, pp. 731–735, 2001.

[19] A. Graves, *Supervised sequence labelling*. Springer, 2012.

[20] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4594–4602.

[21] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," pp. 4534–4542, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," pp. 630–645, 2016.

[23] T. Lyons and Z. Qian, "System control and rough paths,(2002)."

[24] T. Lyons, "Rough paths, signatures and the modelling of functions on streams," *CoRR*, vol. abs/1405.4537, 2014.

[25] B. Hambly and T. Lyons, "Uniqueness for the signature of a path of bounded variation and the reduced path group," *Annals of Mathematics*, pp. 109–167, 2010.

[26] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake, "Modified quadratic discriminant functions and the application to chinese character recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 1, pp. 149–153, 1987.

[27] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[28] B. Verma, J. Lu, M. Ghosh, and R. Ghosh, "A feature extraction technique for online handwriting recognition," vol. 2, pp. 1337–1341, 2004.

[29] Z.-L. Bai and Q. Huo, "A study on the use of 8-directional features for online handwritten chinese character recognition," *Eighth International Conference on Document Analysis and Recognition*, pp. 262–266, 2005.

[30] F. Biadsy, J. El-Sana, and N. Y. Habash, "Online arabic handwriting recognition using hidden markov models," in *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006.

[31] C.-L. Liu and X.-D. Zhou, "Online japanese character recognition using trajectory-based normalization and direction feature extraction," in *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006.

[32] W. Yang, L. Jin, Z. Xie, and Z. Feng, "Improved deep convolutional neural network for online handwritten chinese character recognition using domain-specific knowledge," *13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 551–555, 2015.

[33] W. Yang, L. Jin, D. Tao, Z. Xie, and Z. Feng, "Dropsample: a new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten chinese character recognition," *Pattern Recognition*, vol. 58, pp. 190–203, 2016.

[34] T. Bluche and R. Messina, "Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention," *CoRR*, vol. abs/1604.03286, 2016.

[35] T. Bluche, "Joint line segmentation and transcription for end-to-end handwritten paragraph recognition," pp. 838–846, 2016.

[36] 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," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, 2015.

[37] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, 2015, pp. 2377–2385.

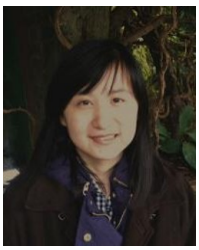
[38] ——, "Highway networks," *CoRR*, vol. abs/1505.00387, 2015.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[40] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," *Tenth IEEE International Conference on Computer Vision*, vol. 2, pp. 1458–1465, 2005.
[41] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.
[44] Q.-F. Wang, F. Yin, and C.-L. Liu, "Integrating language model in handwritten chinese text recognition," *10th International Conference on Document Analysis and Recognition*, pp. 1036–1040, 2009.
[45] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," *Innovations in Machine Learning*, pp. 137–186, 2006.
[46] L. Vilnis and A. McCallum, "Word representations via gaussian embedding," *International Conference on Learning Representations (ICLR)*, 2014.
[47] T. Mukherjee and T. Hospedales, "Gaussian visual-linguistic embedding for zero-shot recognition." EMNLP, 2016.
[48] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, pp. 3104–3112, 2014.
[49] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." *INTERSPEECH*, vol. 2, p. 3, 2010.
[50] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling." *INTERSPEECH*, pp. 194–197, 2012.
[51] K.-T. Chen, "Integration of paths–a faithful representation of paths by noncommutative formal power series," *Transactions of the American Mathematical Society*, vol. 89, no. 2, pp. 395–407, 1958.
[52] B. Graham, "Sparse arrays of signatures for online character recognition," *CoRR*, vol. abs/1308.0371, 2013.
[53] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.
[54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[55] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, 2014.
[56] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1764–1772, 2014.
[57] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "Casia online and offline chinese handwriting databases," *2011 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 37–41, 2011.
[58] F. Yin, Q.-F. Wang, X.-Y. Zhang, and C.-L. Liu, "ICDAR 2013 chinese handwriting recognition competition," *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1464–1470, 2013.
[59] "the people' s daily corpus." http://icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp, [Online] the People' s Daily News and Information Center, the Peking University Institute of Computational Linguistics and Fujitsu Research and Development Center Limited. Accessed March 25, 2016.
[60] G. Jin., "The ph corpus." ftp://ftp.cogsci.ed.ac.uk/pub/chinese, [Online] Accessed March 25, 2016.
[61] "Chinese linguistic data consortium." http://www.chineseldc.org, [Online] the Contemporary Corpus developed by State Language Commission P.R.China, Institute of Applied Linguistics, 2009 Accessed October 22, 2016.
[62] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit." *INTERSPEECH*, vol. 2002, p. 2002, 2002.
[63] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.
[64] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678, 2014.
[65] C.-L. Liu, F. Yin, Q.-F. Wang, and D.-H. Wang, "ICDAR 2011 chinese handwriting recognition competition (2011)."
[66] L. Sun, T. Su, C. Liu, and R. Wang, "Deep lstm networks for online chinese handwriting recognition," in *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*. IEEE, 2016, pp. 271–276.

Zecheng Xie is a PhD student in information and communication engineering at the South China University of Technology. He received a BS in electronics and information engineering from South China University of Technology in 2014. His research interests include machine learning, document analysis and recognition, computer vision, and human-computer interaction.

Zenghui Sun is a master student in communication and information system at the South China University of Technology. He received a BS in electronics and information engineering from South China University of Technology. His research interests include machine learning and computer vision.

Lianwen Jin (M'98) received the B.S. degree from the University of Science and Technology of China, Anhui, China, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 1991 and 1996, respectively. He is a professor in the College of Electronic and Information Engineering at the South China University of Technology. His research interests include handwriting analysis and recognition, image processing, machine learning, and intelligent systems. He has authored over 100 scientific papers. He has received the New Century Excellent Talent Program of MOE Award and the Guangdong Pearl River Distinguished Professor Award, and is a member of the IEEE Computational Intelligence Society, IEEE Signal Processing Society, and IEEE Computer Society.



Hao Ni is a senior lecturer in financial mathematics at UCL since September 2016. Prior to this she was a visiting postdoctoral researcher at ICERM and Department of Applied Mathematics at Brown University from 2012/09 to 2013/05 and continued her postdoctoral research at the Oxford-Man Institute of Quantitative Finance until 2016. She finished her D.Phil. in mathematics in 2012 under the supervision of Professor Terry Lyons at University of Oxford.

Terry Lyons is the Wallis Professor of Mathematics of Oxford university. He was a founding member (2007) of, and then Director (2011-2015) of, the Oxford Man Institute of Quantitative Finance. He was the Director of the Wales Institute of Mathematical and Computational Sciences (WIMCS; 2008-2011). Lyons came to Oxford in 2000 having previously been Professor of Mathematics at Imperial College London (1993-2000), and before that he held the Colin Maclaurin Chair at Edinburgh (1985-93). His research interests are focused on Rough Paths, Stochastic Analysis, and Applications. He is also interested in developing mathematical tools that can be used to effectively model and describe high dimensional systems that exhibit randomness. He was President of the UK Learned Society for Mathematics, the London Mathematical Society (2013-2015).