End-to-End Model for Offline Handwritten Mongolian Word Recognition

Hongxi Wei[1, 2], Cong Liu[1, 2], Hui Zhang[1, 2], Feilong Bao[1, 2] and Guanglai Gao[1, 2]

[1] School of Computer Science, Inner Mongolia University, Hohhot 010021, China
[2] Provincial Key Laboratory of Mongolian Information Processing Technology, Hohhot, China
cswhx@imu.edu.cn

Abstract. This paper proposed an end-to-end model for recognizing offline handwritten Mongolian words. To be specific, a sequence to sequence architecture with attention mechanism is used to perform the task of generating a target sequence from a source sequence. The proposed model consists of two LSTMs and one attention network. The first LSTM is an encoder which consumes a frame sequence of one word image. The second LSTM is a decoder which can generate a sequence of letters. The attention network is added between encoder and decoder, which allow the decoder to focus on different positions in a sequence of frames during the procedure of decoding. In this study, we have attempted two schemes for generating frames from word images. In the first scheme, frames are generated with overlapping. Each adjacent two frames overlap half a frame. In the second scheme, frames are generated without overlapping. In addition, the height of the frame is also taken into consideration in our study. By comparison, the better scheme for generating frames has been determined. Experimental results demonstrate that the proposed end-to-end model outperforms the state-of-the-art method.

Keywords: Offline handwritten recognition, Traditional Mongolian, Segmentation-free, Sequence to sequence, Attention.

1 Introduction

Mongolian language is mainly used in China (e.g. Inner Mongolia Autonomous Region), Republic of Mongolia, Russia and their neighboring areas. The Mongolian language used in China is called *traditional Mongolian*. Correspondingly, the Mongolian language used in Republic of Mongolia and Russia is called *Cyrillic Mongolian*, in which its letters are the same as alphabets of Russian. In this study, we focused on the problem of offline handwritten word recognition for the traditional Mongolian. In the rest of this paper, the traditional Mongolian is called Mongolian without particularly emphasizing.

In recent years, a large number of Mongolian documents in handwriting format have been scanned into images. In general, these scanned handwriting documents can be converted into texts by utilizing the technology of offline handwritten recognition. However, offline handwritten Mongolian word recognition (HMWR) is a challenging

task due to the huge number of vocabularies, special word formations, and various handwriting styles.

Mongolian is an alphabetic language. All letters of one Mongolian word are conglutinated together in the vertical direction to form a backbone, and letters have initial, medial or final visual forms according to their positions within a word [1]. A blank space is used to separate two words. The Mongolian language is also a kind of agglutinative language. Its word formation and inflection are built by connecting different suffixes to the roots or stems. Hence, the number of vocabularies of the Mongolian is huge, and frequently used vocabulary is about one million. Moreover, the Mongolian language has a very special writing system, which is quite different from Arabic, English and other Latin languages. Its writing order is vertical from top to bottom and the column order is from left to right. A fragment of one handwritten Mongolian document and an example of one Mongolian word are illustrated in Fig. 1.

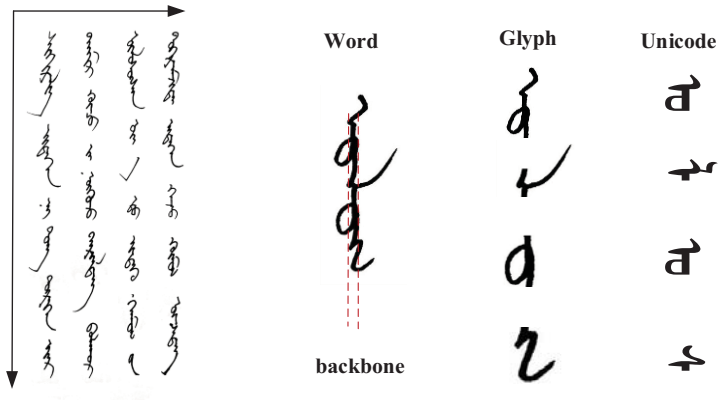

Fig. 1. A sample of handwritten Mongolian text.

In Fig. 1, we can see that it is so difficult to segment handwritten Mongolian words into individual glyphs. Therefore, the conventional character segmentation based schemes [2-6] are infeasible to HMWR. Holistic word recognition is an alternative solution. With the success of deep learning, convolutional neural network (CNN) has been applied successfully in offline handwritten word recognition by the manner of segmentation-free [7-11]. However, CNN based methods classify its input into a certain class among the vocabularies. So, they often suffer from the problem of out-of-vocabulary (OOV), especially for Mongolian language with large vocabularies. More recently, sequence to sequence model becomes popular for the task of offline handwritten recognition [12-14]. Therein, word or text images are fed into the encoder of a sequence to sequence model, and then sequences of letters are generated by a decoder as recognition results. As long as the decoder is able to generate the whole alphabets, sequence to sequence model can solve the problem of OOV.

As far as we know, there is little literature about offline handwritten Mongolian word recognition. In [15] and [16], Fan et al. proposed a DNN-HMM hybrid model for realizing HMWR. Specifically, each Mongolian word image is divided into a

number of sub-characters, and then every sub-character was modeled by an HMM. The observing sequences in HMMs are obtained by using a DNN. Through evaluating on an offline handwritten Mongolian (MHW) dataset, experimental result demonstrates that this model is the state-of-the-art so far.

In this paper, we proposed a novel end-to-end model for recognizing offline handwritten Mongolian words. To be specific, a sequence to sequence architecture with attention mechanism is used to perform the task of generating a target sequence (i.e. a sequence of letters) from a source sequence (i.e. a frame-sequence of one word image). The proposed model consists of two LSTMs and one attention network. The first LSTM is considered as an *encoder* which consumes frame-sequences of one word image. The second LSTM is taken as a *decoder* which can generate a sequence of letters. The attention network is added between encoder and decoder, which can not only improve the effect of parallelism but also decrease training time. This kind of model has been extensively employed for neural machine translation. In this study, our model incorporates a set of improvements for accomplishing the aim of offline handwritten Mongolian word recognition.

The rest of the paper is organized as follows. The proposed model is given in Section 2. Experimental results are shown in Section 3. Section 4 provides the conclusions.

2 The Proposed Model

The proposed model is designed as a sequence to sequence architecture with attention mechanism, which is composed of a BiLSTM based encoder, a LSTM based decoder, and a DNN based attention network is adopted to connect between the encoder and the decoder. Frames of each word image should be inputted into a DNN based feature extractor before being fed into the BiLSTM based encoder. The detailed architecture is shown in Fig. 2.

2.1 Details of the proposed model

In this model, the original inputs are handwritten Mongolian word images. So, all word images should be transformed into a certain kind of sequences in advance. In order to attain this aim, each word image is normalized, and then divided into multiple frames with equal size. After that, the sequence of frames is passed to the encoder and converted into hidden states of the corresponding encoder. Then, the hidden states are fed into the attention network, in which a part of the hidden states are enhanced and the rest are faded. Next, the processed hidden states are put into the decoder. Finally, the decoder is able to generate a sequence of letters as the output of the model.

The decoder of the proposed model is also an LSTM which requires the attention network to choose the most relevant states for the current output letter and filter out irrelevant states. The attention network allows the decoder to focus on different positions in a sequence of frames during the procedure of decoding. The attention network plays an important part in the proposed model.

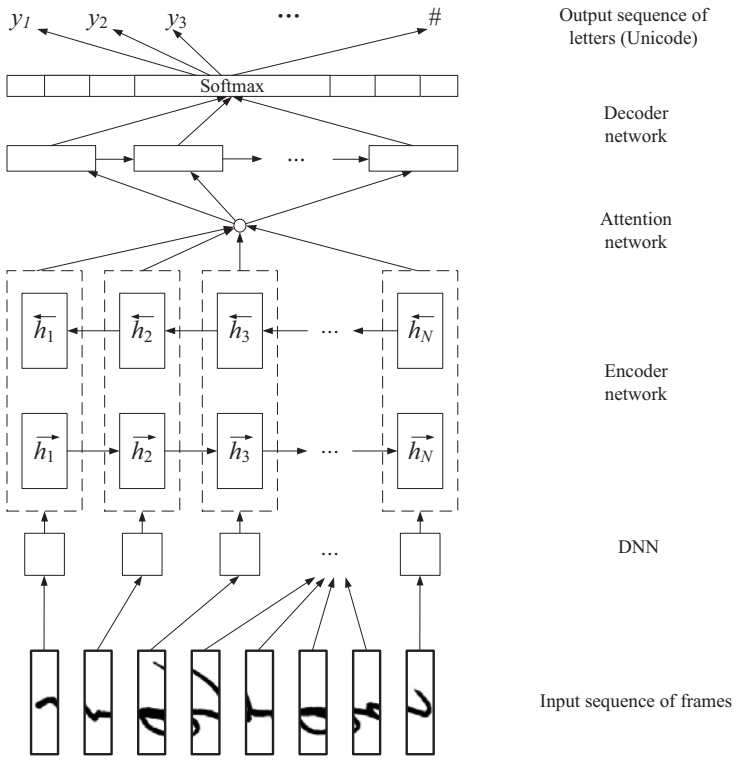

Fig. 2. The architecture of the proposed model.

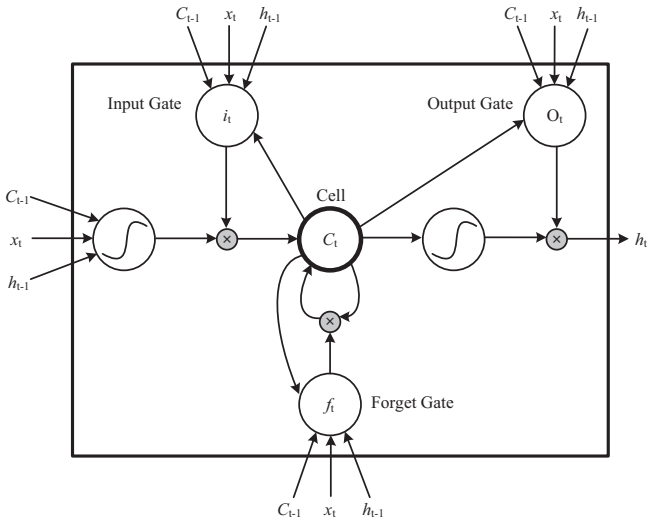

Fig. 3. The structure of a memory cell.

LSTM network is able to learn long-term dependencies, which includes a set of memory cells [17]. The structure of a single memory cell is presented in Fig. 3. A LSTM network transmits the previous information in two ways [18]: the output (or hidden) vector (denoted by h) and the state vector (denoted by c), that combined using three gates, are explicitly designed to store and propagate long-term dependencies.

The gate i is named the input gate. Its value will be updated in the state vector. The gate f is named the forget gate, which can learn the information from the previous state that can be thrown away. With the output of these two gates, the memory cell creates a new state vector. Finally, the gate o is named the output gate, which can generate the output vector of the memory cell. The following equations are used in each memory cell to produce the output vector and the state vector of the moment t, severally.

$$i_t = \sigma(W_{xi} \cdot x_t + W_{hi} \cdot h_{t-1} + W_{ci} \cdot c_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma\left(W_{xf} \cdot x_t + W_{hf} \cdot h_{t-1} + W_{cf} \cdot c_{t-1} + b_f\right) \tag{2}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot tanh(W_{xc} \cdot x_t + W_{hc} \cdot h_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(W_{xo} \cdot x_t + W_{ho} \cdot h_{t-1} + W_{co} \cdot c_t + b_o) \tag{4}$$

$$h_t = o_t \cdot tanh(c_t) \tag{5}$$

Where W (e.g. W_{xi}) and b (e.g. b_i) are trainable parameters, σ is the sigmoid function, and i, f, o, c and h are the input gate, the forget gate, the output gate, the state vector and the output vector, respectively. The BiLSTM of our proposed model contains a forward LSTM and a backward LSTM, and each LSTM consists of 64 memory cells.

2.2 Generating input sequence of frames

The scheme which transforms a word image into a sequence of frames may influence the recognition performance. In this study, we have attempted two schemes for obtaining frames. In the first scheme, frames are generated with overlapping. Each adjacent two frames overlap half a frame, as illustrated in Fig. 4 (a). In the second scheme, frames are generated without overlapping, as illustrated in Fig. 4 (b). The height of the frame is also taken into consideration. Its effect has been tested in our experiment (see Section 3).

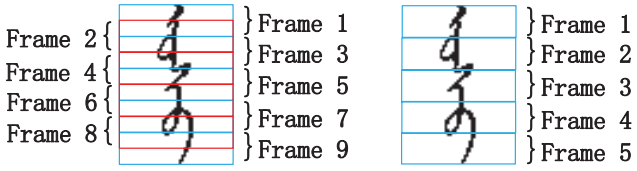

Fig. 4. Two schemes for obtaining frames. (a) Two adjacent frames overlap half a frame. (b) Two adjacent frames without overlapping.

2.3 Data augmentation

In various machine learning tasks, data augmentation schemes can improve performance by generating new data derived from the original data. Here, new samples of word images can be synthesized through a series of simple operations including rotation and scale in horizontal and vertical directions, respectively. The detailed operations are introduced as bellow.

The word images are scaled in the horizontal direction and the scaling factor is sampled from a uniform distribution $U(0.8,1.2)$. The vertical direction is the writing direction of the Mongolian, in which letters are conglutinated together in this direction. We scale the image in the vertical direction to simulate the height variation. To further simulating the variation of each letter, we first segment the image into random number of slices, and then scale each slice with a different scaling factor. The number of slices is random selected based on the whole image height. A taller image is segmented into more slices with high probability. The scaling factor of each slice is sampled from a uniform distribution $U(0.8,1.2)$.

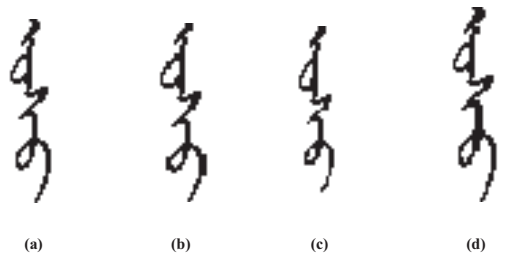

Fig. 5. Samples generated by data augmentation. (a) The original word image; (b)~(d) The generated samples using data augmentation scheme.

Furthermore, we rotate the image in a small angle to simulate the slanted written Mongolian words. The rotating angle is sampled from the Gaussian distribution $\mathrm{N}(0,3)$. These three operations are applied to the original image in a pipeline: (1) Image is scaled in the horizontal direction, randomly. (2) The scaled image is scaled in the vertical direction, randomly. (3) The scaled image is rotated, randomly. In theory, this method can generate a large number of samples to improve the generalization ability of the trained model. Several generated word images are provided in Fig. 5.

3 Experimental Results

3.1 Dataset

In this study, an offline handwritten Mongolian words (MHW) dataset [16] is used to evaluate our proposed model. The training set covers 5,000 vocabularies and each one has 20 samples. The MHW contains two testing sets. The first one (denoted by **Testing Set I**) has 1,000 words randomly selected from the training set and each word has

5 samples. The writers are the same as the training set. The second one (denoted by **Testing Set II**) has 939 words different from the training set. Each word has 15 samples and the writers are different from the training set. In the dataset, each word image has been placed into the center of a bounding box with 48 pixels width.

3.2 Baseline

To evaluate the performance of the proposed model, a state-of-the-art method (i.e. a DNN-HMM hybrid model with Unicode labels) presented in [15-16] is taken as a baseline. The DNN consists of 4 hidden layers and each hidden layer has 1024 nodes. The output of DNN is a triple of Unicode labels, which is formed by the current label concatenated with the previous one and the following one. The HMM is used to model the sequence property. We did not re-implement this model, and just list the original results (**82.16%** and **73.16%**) reported in [16].

3.3 The performance of the proposed model

The height of the frame is an important factor affecting the recognition accuracy. In [16], the authors only used a certain scheme to form the frame sequence, in which the frame height is 11 pixels and the overlap is 10 pixels. In this experiment, we tested various frame heights and overlap sizes.

In Fig. 6, when the frame height is decreased from 8 pixels to 2 pixels with an interval of 2 pixels, the accuracy is increasing. The best accuracies on the two testing sets are **87.68%** and **81.12%**, respectively. It demonstrates that the smaller of the frame height the higher accuracy is obtained. The attention mechanism can determine the most relevant frames associated with a certain output label. In ideal, the relevant frames should only contain the corresponding glyph of the output label. When the frame is taller, a frame may contain more than one glyph, which results in decreasing the accuracy. Therefore, the shorter frame is better.

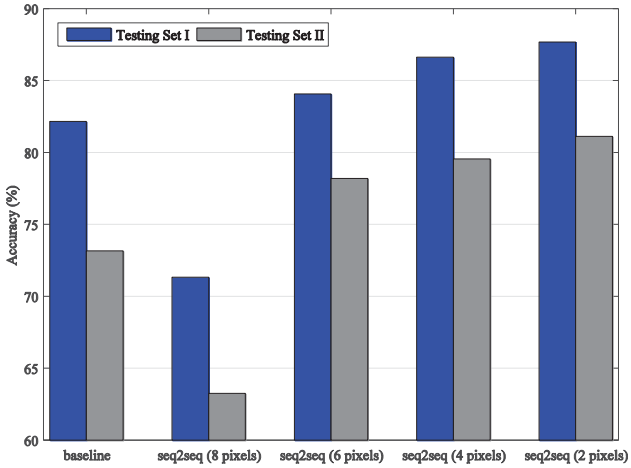

Fig. 6. The performance of the proposed model.

In Fig. 7, when the frame height is set to 2 pixels, the overlapping scheme is better. To improve the accuracy of the non-overlapping scheme, the frame height is set to one pixel. But, the improvement is not enough to beat the overlapping scheme. We can conclude that the overlapping scheme is better than the non-overlapping scheme. Because that the overlapping scheme can provide extra context information of frames to LSTM so as to obtain better performance.

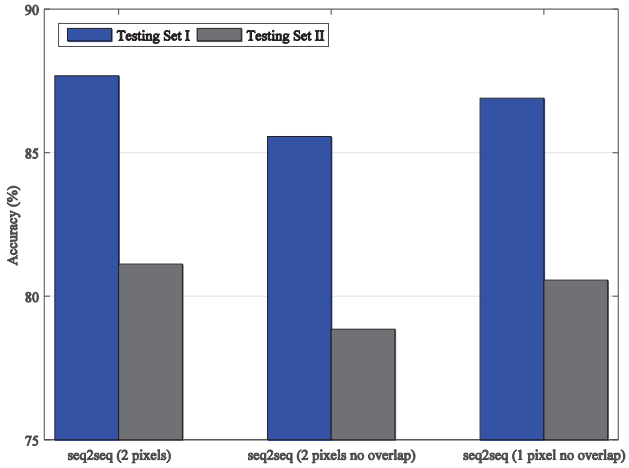

Fig. 7. The comparative results between generated frames with overlapping and non-overlapping.

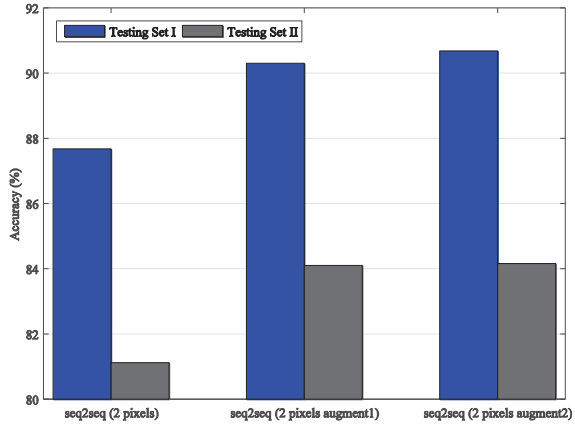

Fig. 8. The performance of data augmentation.

The effect of data augmentation has been tested in our experiment. The results are presented in Fig. 8. We can see that the accuracy is improved by increasing the number of samples. Here, the number of samples for the training set augments twice as much (denoted by **augment1**) and triple as much (denoted by **augment2**), separately. For testing set I, the corresponding accuracies are increased to **90.30%** and **90.68%**,

severally. For testing set II, the accuracies are also increased to **84.10%** and **84.16%**, respectively. Therefore, we can conclude that data augmentation scheme can improve the performance.

4 Conclusions

In this paper, we proposed an end-to-end model for recognizing offline handwritten Mongolian words. Specifically, a sequence to sequence structure with attention mechanism is used to perform the task of generating a target sequence from a source sequence. The proposed model consists of two LSTMs and one attention network. The first LSTM is an encoder which processes a sequence of frames. The second LSTM is a decoder which can generate a sequence of letters as recognition results. The attention network is added between encoder and decoder, which allow the decoder to focus on different positions in a sequence of frames during the procedure of decoding.

In this work, we have attempted two schemes for generating frames from handwritten Mongolian word images. In the first scheme, frames are generated with overlapping that each adjacent two frames overlap half a frame. In the second scheme, frames are generated without overlapping. In addition, the height of the frame is also taken into consideration. By comparison, our proposed model is superior to the state-of-the-art DNN-HMM hybrid model on a dataset of offline handwritten Mongolian words. Moreover, we have compared several heights of frames. The frame heights are set to 8 pixels, 6 pixels, 4 pixels and 2 pixels, separately. When the frame height is set to 2 pixels, the best performance can be attained. It demonstrates that the smaller of the frame height the higher accuracy can be obtained. It is in line with the working mechanism of attention. The architecture of sequence to sequence can handle the problem of OOV. Therefore, the proposed model is especially suited for realizing large vocabulary HMWR.

Acknowledgement

This paper is supported by the National Natural Science Foundation of China under Grant 61463038.

References

1. H. Wei and G. Gao, “A keyword retrieval system for historical Mongolian document images,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, no. 1, pp. 33–45, 2014.
2. H. Wei and G. Gao, “Machine-printed traditional Mongolian characters recognition using BP neural networks,” in *Proceeding of 2009 International Conference on Computational Intelligence and Software Engineering (CiSE'09)*, IEEE, 2009, pp. 1–7.
3. H. Hu, H. Wei, and Z. Liu, “The CNN based machine-printed traditional Mongolian characters recognition,” in *Proceedings of the 36th Chinese Control Conference (CCC'17)*, IEEE, 2017, pp. 3937–3941.

4. G. Gao, X. Su, H. Wei, and Y. Gong, “Classical Mongolian words recognition in historical document,” in *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR'11)*, IEEE, 2011, pp. 692–697.
5. X. Su, G. Gao, W. Wang, F. Bao, and H. Wei, “Character segmentation for classical Mongolian words in historical documents,” in *Proceedings of the 6th Chinese Conference on Pattern Recognition (CCPR'14)*, Springer, 2014, Part II, pp. 464–473.
6. X. Su, G. Gao, H. Wei, and F. Bao, “A knowledge-based recognition system for historical Mongolian documents,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 19, no. 4, pp. 221–235, 2016.
7. A. Yuan, G. Bai, P. Yang, Y. Guo, and X. Zhao, “Handwritten English word recognition based on convolutional neural networks,” in *Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR'12)*, IEEE, 2012, pp. 207–212.
8. 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.
9. M. Elleuch, N. Tagougui, and M. Kherallah, “Towards unsupervised learning for Arabic handwritten recognition using deep architectures,” in *Proceedings of the 22nd International Conference on Neural Information Processing (ICONIP'15)*, Springer, 2015, Part I, pp. 363–372.
10. I. Kim and X. Xie, “Handwritten Hangul recognition using deep convolutional neural networks,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 18, no. 1, pp. 1–13, 2015.
11. C. Adak, B. B. Chaudhuri, and M. Blumenstein, “Offline cursive Bengali word recognition using CNNs with a recurrent model,” in *Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR'16)*, IEEE, 2016, pp. 429–434.
12. X. Zhang and C. L. Tan, “Unconstrained handwritten word recognition based on trigrams using BLSTM,” in *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR'14)*, IEEE, 2014, pp. 2914–2919.
13. R. Messina and J. Louradour, “Segmentation-free handwritten Chinese text recognition with LSTM-RNN,” in *Proceedings of 13th International Conference on Document Analysis and Recognition (ICDAR'15)*, IEEE, 2015, pp. 171–175.
14. P. Voigtlaender, P. Doetsch, and H. Ney, “Handwriting recognition with large multidimensional long short-term memory recurrent neural networks,” in *Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR'16)*, IEEE, 2016, pp. 228–233.
15. D. Fan and G. Gao, “DNN-HMM for large vocabulary Mongolian offline handwriting recognition,” in *Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition (ICFHR'16)*, IEEE, 2016, pp. 72–77.
16. D. Fan, G. Gao, and H. Wu, “MHW Mongolian offline handwritten dataset and its application,” *Journal of Chinese Information Processing*, vol. 32, no. 1, pp. 89–95, 2018.
17. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no.8, pp.1735-1780, 1997.
18. A. Graves, A. Mohamed, and G. E. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proceedings of the 38th International Conference on Acoustics, Speech and Signal Processing (ICASSP'13)*, 2013, pp. 6645-6649.