Retrieval-Based Language Model Adaptation for Handwritten Chinese Text Recognition

Shuying Hu[1], Qiufeng Wang[2*], Kaizhu Huang[3], Min Wen[1] and Frans Coenen[4]

[1]Department of Applied Mathematics, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou, 215123, Jiangsu, China.

[2*]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou, 215123, Jiangsu, China.

[3]Data Science Research Center, Duke Kunshan University, No.8 Duke Avenue, Kunshan, 215316, Jiangsu, China.

[4]Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, United Kingdom.

*Corresponding author(s). E-mail(s): Qiufeng.Wang@xjtlu.edu.cn;
Contributing authors: Shuying.Hu20@student.xjtlu.edu.cn;
kaizhu.huang@dukekunshan.edu.cn; Min.Wen@xjtlu.edu.cn; Coenen@liverpool.ac.uk;

Abstract

In handwritten text recognition, compared to human, computers are far short of linguistic context knowledge, especially domain-matched knowledge. In this paper, we present a novel retrieval-based method to obtain an adaptive language model for offline recognition of unconstrained handwritten Chinese texts. The content of handwritten texts to be recognized is varied and usually unknown a priori. Therefore we adopt a two-pass recognition strategy. In the first pass, we utilize a common language model to obtain initial recognition results, which are used to retrieve the related contents from Internet. In the content retrieval, we evaluate different types of semantic representation from BERT output and the traditional TF-IDF representation. Then, we dynamically generate an adaptive language model from these related contents, which will consequently be combined with the common language model and applied in the second-pass recognition. We evaluate the proposed method on two benchmark unconstrained handwriting datasets, namely CASIA-HWDB and ICDAR-2013. Experimental results show that the proposed retrieval-based language model adaptation yields improvements in recognition performance, despite the reduced Internet contents hereby employed.

Keywords: Recognition, Handwritten Chinese Text Recognition, Internet Content, Information Retrieval, Language Model Adaptation

1 Introduction

Documents comprising handwritten or printed characters are one of the most popular tools for our communication and archiving [1]. To digitize these documents, optical character recognition (OCR) has been widely researched and applied [1, 2]. Solid progress has been made in many areas, e.g. from isolated character recognition to character string recognition, from printed character recognition to unconstrained handwriting recognition, and from documents with clear background to scene text recognition with complex background. While related tasks are getting more and more complicated, recent advancement in OCR has lead to great success in real applications. Chinese handwriting recognition has been an important branch of OCR since 1970s [3, 4]. Powered by deep learning, handwritten isolated Chinese character recognition has achieved tremendous advance [4, 5, 6, 7, 8]. Remarkably, the reported accuracy rate can even be higher than that of human recognition: 97.64% was reported in [7] whilst human only get the accuracy of 96.13%. Nevertheless, automated unconstrained handwritten Chinese text recognition still remains unsatisfactory and actually far behind human recognition, since human can effectively leverage sufficient linguistic context knowledge [4, 5, 6, 9, 10]. Concretely, there are huge challenges in automated unconstrained handwritten Chinese text recognition including the low-quality of text images, flexible handwriting styles, and unusual topics with possible specific terminologies, and shortage of linguistic context. Figure 1 illustrates one example from the benchmark dataset of CASIA-HWDB with the cursive handwriting style and specific terminologies, where the airport name '白' (white) is often incorrectly recognized as '自' (self).

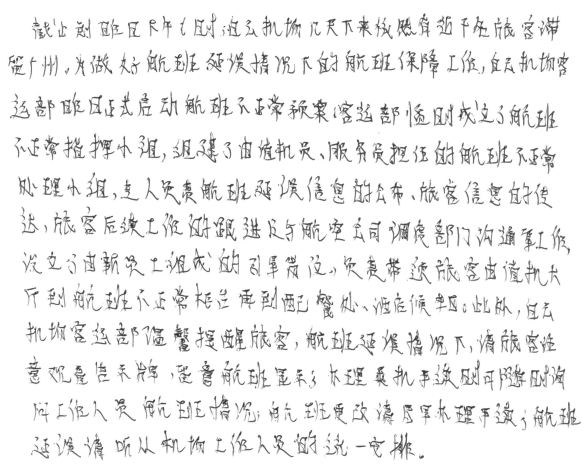

Fig. 1: Example of a handwritten cursive Chinese text page where the airport name is often incorrectly recognized.

Generally, handwritten Chinese text recognition (HCTR) is a sequence pattern recognition problem, which can be translated to searching for the optimal path in a complicated candidate lattice by over-segmentation under certain path evaluation criterion [9, 10]. Inspired by how humans read texts, handwritten text recognition usually

utilizes language models to represent linguistic context knowledge, which characterizes the statistical dependency between characters and assigns the prior probability of a sequence of characters. Unfortunately, in lack of sufficient linguistic context knowledge, current automated methods are still far away from humans' ability, thus limiting the ceil point especially for HCTR.

Language models (LMs) play a very important role in HCTR. There are a fast-growing body of methods which have explored how to apply LMs in HCTR recently [9, 13, 14]. In some literature, LM is merely exploited as post-processing for correcting recognition errors [23, 24]. Despite its simplicity and plug-and-play property, such practice may actually limit LM's full play in the learning process. Instead, as indicated in many investigations, LM needs be seamlessly integrated in the path search so that the learning process can be well guided to the optimum [9, 22].

On the other hand, there are sufficient and diverse contents on Internet, where linguistic context might be mined. Indeed, Internet contents have been exploited in many fields, e.g. never-ending machine learning [15, 16], image recognition [17], speech recognition [18], postcards recognition [19], and scene text recognition [20, 21]. However, to the best of our knowledge, all of these works either directly utilize the contents or simply correct the recognized result as post-processing. There have rare studies investigating how to integrate Internet contents with LM for HCTR.

However, if the domain of LM does not match the handwritten texts to be recognized, its effect could be limited. In fact, one can usually observe such phenomenon, since characters or words usually enjoy various statistics in the text corpus of different domains [26]. To deal with this mismatch problem, language model adaption has been widely investigated. This is particularly the case in unsupervised adaptation where no prior information exists about the domain of handwritten texts [14]. In [14], a large text corpus was collected with different domains downloaded from the Internet in advance. A set of LMs were then trained for those pre-defined domains. However, these LMs are unchanged for all handwritten texts, which is not flexible.

In this paper, we propose a method to dynamically retrieve the related content from Internet for HCTR, then train an adaptive LM from this content, which is integrated in the whole recognition process. For the retrieval of the related content, we adopt the Transformer-based language model BERT [58], which maps each word to a semantic space with the context information. Since we do not have any prior information for the handwritten texts to be recognized, we utilize a two-pass recognition strategy. In the first-pass recognition, we apply a common language model to get an initial text result. We then retrieve the related

contents based on this transcription and build a domain-matched N-gram model, which will be integrated with the common LM in the second-pass recognition. The retrieval of the related contents is based on a criterion that measures the similarity between the recognized text and contents from the Internet corpus.

The rest of this paper is organized as follows: Section 2 reviews some related work on handwriting recognition and language models; Section 3 gives an overview of our HCTR system; Section 4 describes in details the proposed retrieval-based language model; Section 5 presents the experimental results, and Section 6 concludes this paper with final remarks.

2 Related works

In the context of retrieval-based handwritten Chinese text recognition from internet contents, much work has contributed to the related issues, including internet-based pattern recognition, language model, and handwritten Chinese text recognition. In the following, we will give an overview of these previous work.

2.1 Internet-based pattern recognition

An LM is generally developed using a text corpus of millions or even billions of sentences, so crawling sources of online text data is a simple way for building LMs. Given the popularization of the Internet in recent decades, some researchers have considered to utilize the idea of web search on Internet contents to improve the performance of pattern recognition tasks. These methods can be roughly categorized into three groups according to how they utilize the Internet: (i) collecting and labelling the data set, (ii) training the classifier, and (iii) integrating in the recognition system directly. In the first category, people have used the Internet to work on the data set cooperatively. For example, Russel et al. designed a web-based tool for Image annotation [27], and the famous large-scale ImageNet image dataset was collected and labelled using Internet crowdsourcing [28, 29]. The second category of methods utilizes the Internet to get a large set of related data, which is then used to train the classifier. Fergus et al. [17] automatically learned the categories of objects from the images retrieved on Google, and Hankz et al. [30] utilized Internet contents in the transfer learning between different domains for action-model learning. The last category is to use the Internet content as the linguistic context in the recognition. For example, Nishizaki et al. [18] utilized Internet content to correct the errors in the post-processing of speech recognition, Chen et al. [31] applied Internet content to language model adaptation, and White law et al. [32] explored Internet content for spell checking and autocorrection. All of these

work has promoted the development of pattern recognition, and improved the performance.

In the area of OCR, much work has also been reported on exploring Internet content to improve recognition performance, and most of these work used the idea of the aforementioned third category. Clemens et al. [19] utilized Internet content to verify and correct text recognition for postcards. Bassil et al. [33] applied the autocorrection function of Google in the post-processing of OCR. Donoser et al. [20, 21] re-scored candidate characters based on the Internet retrieval results in scene text recognition, where the authors assumed that the web search results of correct recognition text were many more than those of the wrong results. However, all of these work only leveraged the Internet contents in the post-processing of OCR instead of being integrated in the whole recognition process. Recently, Oprean et al. [34, 35] utilized the contents of Wikepedia to construct the dictionary dynamically to overcome the Out-of-Vocabulary (OOV) issue in handwritten English word recognition, and successfully extended it to the handwritten English text recognition using a deep learning framework [36].

2.2 Language model

LMs have been widely used in speech recognition, handwriting recognition, and machine translation [25, 37]. The N-gram model is the most popular LM in handwriting recognition, which characterizes the statistical dependency between the neighbour N characters or words [9, 22, 23, 24, 38, 39]. However, the N-gram model usually has two issues: zero-probability for the unseen N-gram sequences (though various smoothing methods have been proposed) and the local context limit (considering the moderate model size and decoding time where n is usually 2 or 3). In recent decades, the neural network based language model was developed, quickly to overcome the zero-probability issue due to the distributed representation of all words [40, 41, 42, 43, 44]. Some of these models have been successfully applied in the HCTR [10]. For the issue of local context limit, many topic model based language models have been proposed [45, 46, 47], and Xie et al. [48] developed an implicit LM to integrate the global linguistic context in the online HCTR.

Language model adaptation (LMA) is an important technique used to adapt a common language model to match the domain of each recognition task, which can be categorized into supervised LMA and unsupervised LMA [26]. Supervised LMA assumes that the domain of the recognition task is known in advance. Therefore, a large set of related texts can be obtained to train a domain-matched LM [49]. However, the domain information is usually unknown a priori, which requires unsupervised LMA. The basic idea is to use a common LM to get an initial recognition result, which is then used to either search related

text to train an adaptive LM [31] or learn the weights to combine various LMs to get a balanced LM [50]. Most work on LMA has been conducted in speech recognition and natural language processing (NLP), while there are few examples where LMA has been used for HCTR. Recently, our previous work [14] has investigated three unsupervised LMA methods based on a pre-defined LM set, validating the effectiveness of LMA in HCTR.

2.3 Handwritten Chinese text recognition

HCTR has attracted lots of attention since two databases (HIT-MW [51] and CASIA-HWDB [12]) were released and two competitions were organized at ICDAR 2011 [5] and ICDAR 2013 [52]. HCTR has achieved great progress in recent years [9, 10, 38, 39, 48, 53, 54, 63, 64]. The approaches adopted can be divided into two categories: over-segmentation based and segmentation-free.

In the over-segmentation based approaches, the text line is over-segmented into a sequence of primitive segments, each corresponding to a character or a part of a character. Then a candidate segmentation-recognition lattice is constructed by combining neighboring segments to be recognized via the character classifier, where each candidate path represents one segmentation-recognition result. This approach can take advantage of the character shape and overlapping characteristic to better separate the characters at their boundaries. Most reported HCTR work has been based on this framework [9, 10, 38, 39]. However, a high recall over-segmentation algorithm is usually necessary, which might be however difficult for touching characters [55]. To improve the recall of segmentation, Wu et al. [10] proposed an over-segmentation algorithm based on the convolutional neural network (CNN). To avoid difficulties of finding exact boundaries in character-level annotation, Wang et al. [11] proposed a weakly supervised learning method to optimize the character classifier by string-level training such that strong annotation is not required. The recent works [62, 63] formulated the character segmentation as a character detection based on a fully convolutional network, which demonstrated great potentials for the segmentation-based HCTR.

The segmentation-free approach is also called the implicit segmentation approach[1], which first uses the technique of a sliding window to split a text image into a sequence, then uses the Hidden Markov Model (HMM) or Recurrent Neural Networks (RNN) based models with Connectionist temporal classification (CTC) or attention strategies to get the recognition result. Most reported work of English text recognition is based on this approach [56]. Su et al. [53] was the first

[1] Accordingly, the over-segmentation is also called as explicit segmentation.

to apply this approach to HCTR, and Wang et al. [54] estimated the observation probability by using a deep CNN model instead of the Gaussian mixture model (GMM). Recently, deep neural network based approaches have been widely used in the segmentation-free recognition, namely, RNN, especially Long-Short Term Memory (LSTM) and Bidirectional LSTM. The work in [57] proposed a multi-dimensional LSTM and CTC framework for end-to-end HCTR. Addtionally, the work reported in [48] utilized the fully connected RNN model for the online text recognition. Wang et al. [64] further proposed a writer-aware CNN based on parsimonious HMM to address the issues of large vocabulary and diversity of writing styles in offline HCTR. Under the powerful seq-to-seq framework, the segmentation-free approach has shown the great progress in the HCTR. However, this approach is usually difficult to obtain the character boundary information.

3 System overview

In this paper, we design the over-segmentation based framework of [11] as the baseline recognition system, then propose to integrate a retrieval-based language model to improve the recognition performance. As the text to recognize is usually not known in advance, we adopt the two-pass recognition strategy. In the first-pass, we utilize the baseline system to obtain an initial recognition result, which will be used to search the related contents from an Internet corpus and build a retrieval-based language model integrated in the path search as shown in the dashed-line box in Figure 2.

In the general over-segmentation based system, we usually regard the text recognition problem as searching the optimal path in a candidate lattice. In the following sections, we will describe the lattice generation and path search, respectively. Further details are provided in the following two sections.

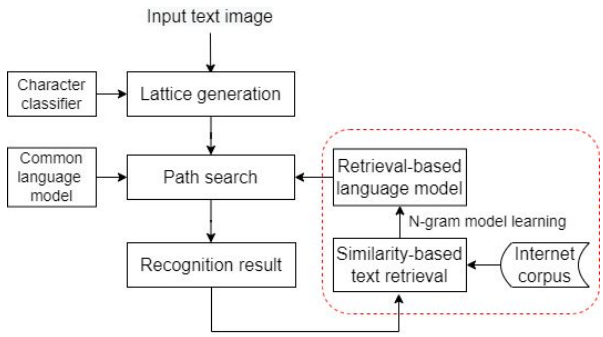

Fig. 2: System diagram for handwritten Chinese text recognition with the retrieval-based language model.

3.1 Lattice generation

Given an input text line image, we first over-segment it to a sequence of primitive segments (corresponding to a character or a part of a character), then combine several consecutive segments to form a candidate character pattern. Finally, we

engage a pre-trained CNN-based character classifier to output the top K character classes for each character pattern. By combining all the candidate character patterns and the corresponding character classes, we obtain a candidate segmentation-recognition lattice, where each path represents a candidate recognition result. Since we focus on the language model in this paper, we will omit an exhaustive description of the lattice generation and refer readers to the work [10].

3.2 Path search

Once the candidate lattice is generated, our target is to find the optimal path under a path evaluation function. The lattice is very complicated as it contains the uncertainties of both segmentation and recognition. Therefore, it is not possible to utilize the exhaustive search method. To improve the efficiency, we adopt the refined beam search algorithm [9], where the beam width is the same as the setting in the baseline work [9].

In the path evaluation function, we consider both the character recognition score and the linguistic context score as shown in Eqn. 1, where the weight λ_{LM} is used to balance two scores,

$$\mathcal{L} = \mathcal{L}_{\text{char}} + \lambda_{\text{LM}} \cdot \mathcal{L}_{\text{LM}}. \tag{1}$$

In our system, the character recognition score $\mathcal{L}_{\text{char}}$ is calculated by a deep CNN on the corresponding character pattern as shown in Eqn. 2, where the coefficient w_i represents the normalized character width to overcome the bias issue of the short-length path [10]. In our framework, we refer to the same CNN structure in the baseline work [10, 11].

$$\mathcal{L}_{\text{char}} = -w_i \cdot \log[\mathbf{CNN}(\cdot)]. \tag{2}$$

In the evaluation function, the LM plays an important role to provide the linguistic context information, and we utilize the popular N-gram model as shown in Eqn. 3.

$$\mathcal{L}_{\text{LM}} = -\log[\mathbf{Ngram}(\cdot)]. \tag{3}$$

Once we obtain the scores of both character recognition and language model, we aim to minimize the total function value to get the optimal path in the lattice.

4 Retrieval-based language model

In general, one common LM is used to evaluate the linguistic context expressed by Eqn. 1, however, if this LM does not match the document to be recognized, it may give a wrong score. To overcome this issue, we consider integrating an adaptive language model in Eqn. 1, which is constructed from the texts matched to the document to be recognized.

As the document to be recognized is variable, we first utilize the information retrieval techniques to obtain the matched texts from a large corpus (e.g., Internet resources), then learn an LM from these retrieved texts. This is called a retrieval-based language model. Finally, we integrate the loss function of the retrieval-based language model, $\mathcal{L}_{\text{rLM}}$, into the path evaluate function to obtain a more accurate score with a balanced weight λ_{rLM} as shown in Eqn. 4:

$$\mathcal{L} = \mathcal{L}_{\text{char}} + \lambda_{\text{LM}} \cdot \mathcal{L}_{\text{LM}} + \lambda_{\text{rLM}} \cdot \mathcal{L}_{\text{rLM}}. \tag{4}$$

In information retrieval, the similarity metric plays an important role, and we adopt the widely used Cosine similarity as shown in Eqn. 5:

$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}. \tag{5}$$

In the above, the symbols $\mathbf{u}$ and $\mathbf{v}$ denote the vector representation of query document (i.e., the document to be recognized) and retrieved content in the prepared corpus, respectively. In the following, we will describe two methods to obtain the vector representation of a document in our system.

4.1 TF-IDF based content retrieval

The first approach to generate vectors for documents is based on the Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a technique for vectorizing documents based on the Bag of words (BoW) model. It measures how important a term is within a document relative to a collection of documents. Term Frequency (TF) measures how frequently a term occurs in a document. It is the ratio of the number of times a term w appear in the document to the document length (total number of terms in the document):

$$\text{TF}(w) = \frac{N_w}{N}, \tag{6}$$

where N_w represents the number of times for the item w in a document, and N represents the total number of terms in the document. Inverse Data Frequency (IDF) measures how important a term is. It is defined as the log of the ratio of total number of documents in a collection to number of documents that contain a particular word:

$$\text{IDF}(w) = \log(\frac{D}{D_w}), \tag{7}$$

where the variable D represents the total number of documents in the corpus, and the variable D_w represents the number of documents that contain w. As a result, it weighs down the frequent terms and scales up the rare ones. Finally the TF-IDF value is the product of TF and IDF:

$$\text{TF-IDF}(w) = \text{TF}(w) \cdot \text{IDF}(w), \tag{8}$$

To obtain the vector representation of a document, we first segment all documents in the corpus into

word levels using the Jieba toolkit[2] with a dictionary, then calculate the TF-IDF value of each word in the dictionary for each document (if the word does not show in this document, we set the TF-IDF value as zero), and finally we concatenate all TF-IDF values to be the vector of each document (i.e., the dimension of each vector is the size of the dictionary).

4.2 BERT based content retrieval

In our system, we also utilize the BERT model [58] to extract the vector of each document. BERT is a method for pre-training language representations, which is able to capture the contextual information of each word. For NLP tasks like semantic textual similarity (STS), BERT has achieved new state-of-the-art performance.

In our experiment, we adopt the pre-trained Bert-Base-Chinese model[3], which contains 12 layers with 68 hidden states for each layer. We first segment each document into tokens using the BERT tokenizer as the input of the multi-layer Transformer, then obtain the document vector by four different methods. Specifically, **BE** stands for Bert-based embedding, and these four methods are (1) **BE1**: the representation of [CLS] of the last layer; (2) **BE2**: the average of the sequence of token vectors from the last layer; (3) **BE3**: the average of the sequence of token vectors from the second last layer; (4) **BE4**: the average of the sequence of token vectors from the last two layers.

[2]https://pypi.org/project/jieba/
[3]https://github.com/google-research/bert

4.3 Language model construction

After retrieving the top-N related news articles, the SRI Language Modeling Toolkit (SRILM) [59] is utilized to build the retrieval-based language model. SRILM can provide frequency counts for N-grams after processing the related news corpus. It is expected that some phrases, terminologies, and names of places or celebrities can be better recognized with the dynamically generated retrieval-based language model. For example, if the document is related to tennis, then in the small Internet corpus, tennis-related terminologies and names of famous players may appear multiple times, increasing their possibility to be chosen during the search algorithm.

4.4 Additional improvements on HCTR

In addition to the retrieval-based model, we also implemented two extra improvements for HCTR.

Concatenating Adjacent Lines. Since the text line image is recognized one by one, the character at the beginning and the end of a line is recognized without context information from its adjacent lines in the same document. As a result, the language model has incomplete stencils at the beginning and the end of a text line, which will

lead to less accurate recognition. A simple idea is to concatenate two adjacent text lines for recognition. Thus, both previous and current text lines can benefit from it.

Adjusting Weights of Punctuation Marks. It is observed that some punctuation marks are often incorrectly recognized, and sometimes their neighboring Chinese characters are affected and not correctly recognized as well. This is partially because that some punctuation marks, such as commas and enumeration commas, are relatively simple shapes and resemble the strokes of Chinese characters. The framework may confuse them with some of the over-segmented parts of a Chinese character, or vice versa. For these scenarios, we proposed to adjust the weight for the loss function of the language model such that the importance of the context over the shape of the character pattern can be modified accordingly. When the top candidate character pattern belongs to a certain punctuation mark, the weight of the language model will be scaled by a factor of α_{mark},

$$\lambda_{\text{LM}} = \alpha_{\text{mark}} \cdot \lambda_{\text{LM}}. \tag{9}$$

5 Experiments

5.1 Dataset and experimental setting

In this paper, we evaluate the proposed method on two benchmark Chinese handwriting recognition datasets: CASIA-HWDB [12] and ICDAR-2013 competition dataset [6]. The CASIA-HWDB database contains both isolated characters and unconstrained handwritten texts, where the training set contains 3,118,447 isolated character samples of 7,356 classes and 4076 pages of handwritten pages (including 41,781 textline samples). We tested our system on the test set containing 1,015 pages (including 10,449 textline samples). The dataset ICDAR-2013 contains 300 pages (including 3,432 textline samples) for testing only.

The values of the hyper-parameters in this paper are set by following the baseline system, which make a trade-off between the accuracy and efficiency [10]. Specifically, we set the maximum number of concatenated segments as 4 (e.g., a character pattern can contain at most 4 segments) in the candidate lattice generation, and the candidate number of character classification as 20 (i.e., the top 20 character classes with the high classification scores). In the path search, we set the beam width as 10. The common language model $\mathcal{L}_{\text{LM}}$ is a character-level 2-gram language model, and the weight λ_{LM} is set as 0.1 by the trial-and-error method.

For the retrieval-based language model, we take the top 200 documents for each recognized document from the retrieval in the Internet corpus. For each document to be recognized, the articles that have similar content are selected from a large Internet corpus from Sogou [4] which consists of more than 70,000 news articles. The selection is based on the sentence similarity and the top-N most related news articles that are stored to form a small Internet corpus. In Eqn. 4, we set the weight λ_{rLM} as 0.05 by the trial-and-error method.

Furthermore, we concatenate two adjacent text lines during the path search to add the linguistic context for the beginning characters. In addition, we adjust the language model weight λ_{LM} on the punctuation marks as they have weak linguistic context. Both tricks are effective for the improvement of the recognition performance.

We evaluate the recognition performance in terms of two character-level metrics, i.e., Correct Rate (CR) and Accurate Rate (AR):

$$\mathrm{CR} = \frac{N_{\mathrm{t}} - N_{\mathrm{de}} - N_{\mathrm{se}}}{N_{\mathrm{t}}}, \tag{10}$$

$$\mathrm{AR} = \frac{N_{\mathrm{t}} - N_{\mathrm{de}} - N_{\mathrm{se}} - N_{\mathrm{ie}}}{N_{\mathrm{t}}}. \tag{11}$$

In the above equations, N_t is the total number of characters in the transcript of test documents. The numbers of substitution errors N_{se}, deletion errors N_{de} and insertion errors N_{ie} are calculated by aligning the recognition result string with the transcript using dynamic programming.

[4] http://www.sogou.com/labs/resource/list_news.php

5.2 Experimental results

For the evaluation, we mainly evaluate the effectiveness of the proposed retrieval-based language model on the benchmark datasets of both CASIA-HWDB and ICDAR-2013, and the results are shown in Tables 1 and 2, respectively. In our baseline recognition system, we only use a common language model (i.e., character bi-gram) as shown in the first row of each table. In the retrieval-based language model, we also utilize the character bi-gram and take top 200 relevant texts retrieved by the TF-IDF method. The last line in each table shows the proposed method with the proposed two further improvement tricks (i.e., concatenating adjacent lines and adjusting language model weights on punctuation marks).

As shown in Tables 1 and 2, we can see that the performance of only using the retrieved-based model decreases significantly in comparison to the baseline recognition. The reason is that the retrieved-based model is only built on a small size of related contents, resulting in serious sparsity in the N-grams. However, we can see that the recognition performance is increased by the combination of the common language model and retrieval-based model. Since the retrieval-based model can provide complimentary linguistic information from the relevant contents, it increases

Table 1: Comparison of different models on CASIA-HWDB. (The bold values indicate the highest performance)

Experiment	CR(%)	AR(%)	Ch(%)	Sb(%)	Dg(%)	Lt(%)
Common LM	94.85	94.09	96.79	85.01	89.20	60.37
Retrieval LM	92.70	90.78	94.96	81.65	83.06	62.92
Common + Retrieval LM	94.86	94.16	96.78	84.72	90.93	67.07
Common + Retrieval LM+Improvements	**94.91**	**94.25**	**96.85**	**84.78**	**91.15**	**67.43**

Table 2: Comparison of different models on ICDAR-2013. (The bold values indicate the highest performance)

Experiment	CR(%)	AR(%)	Ch(%)	Sb(%)	Dg(%)	Lt(%)
Common LM	94.28	93.28	96.19	81.90	85.30	42.79
Retrieval LM	90.65	87.87	92.85	76.63	80.16	45.15
Common+Retrieval LM	94.68	93.79	96.50	82.55	87.64	47.04
Common+Retrieval LM+Improvements	**94.87**	**94.00**	**96.57**	**83.92**	**88.15**	**48.46**

the potential that characters or phrases that often appear in certain fields are recognized correctly. Based on this, we also implemented two further improvement tricks and found they increase the recognition accuracy to the CR 94.91% and the AR 94.25% on CASIA-HWDB, the CR 94.87% and the AR 94.00% on ICDAR-2013. In summary, we can see that both CR and AR values are improved by using the retrieval-based model, connecting two adjacent textlines and adjusting the language model weights of punctuation marks.

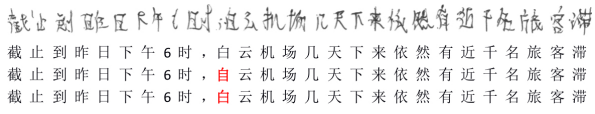

Fig. 3: An example of recognition with/without the retrieval-based language model. The first row is the text line image; the second row is the transcript (ground-truth); the third row is the result without the retrieval-based language model; the last row is result with the retrieval-based language model.

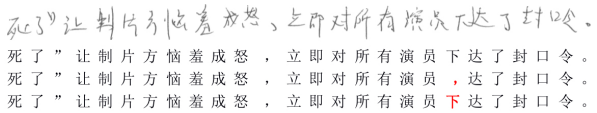

Fig. 4: An example of recognition with/without adjusted weights for punctuation marks. The first row is the text line image; the second row is the transcript (ground-truth); the third row is the result without adjusted weights for punctuation marks; the last row is result with adjusted weights for punctuation marks.

In order to demonstrate the effectiveness of the proposed method intuitively, we show some recognition examples in Figures 3, 4 and 5. In Figure 3, we show that the airport name is recognized correctly after the retrieval-based language model is integrated. The possible reason is that this airport name frequently appeared in the retrieved corpus, which increases the estimation of language model. As shown in Figure 4, the recognition was improved by adjusting the language model weight

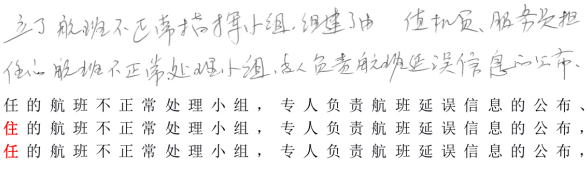

Fig. 5: An example of recognition with/without concatenating adjacent lines. The first two rows are two adjacent text line images; the third row is the transcript (ground-truth) for the second text line image; the fourth row is the result for the second text line image without concatenating two adjacent lines; the last row is result for the second text line image with concatenating two adjacent lines.

for punctuation marks, where the original recognition is misclassified by a comma. In Figure 5, we show one example of recognition by concatenating adjacent lines. We can see that the character at the beginning of the current line is misrecognized to '住' due to the high similarity to that character image, which is corrected by concatenating the previous line recognition as the last character provides more contexts for this correction.

Taking the paragraph in Figure 3 as an example, we examined how many the retrieved news are related to the target text. As shown in Table 3, 158 out of the top 200 retrieved news hit the target of air transportation, while the remaining 42 news also had strong overlapping with air transportation. In 19 sports news and 4 entertainment news, sports teams and celebrities encountering flight delay are often reported. It can be expected that news of other types of transportation may also be retrieved such as railway/road/boat transportation. The 6 social news are mainly using air transportation for natural disaster rescue. The airport name is widely used in these retrieved texts, that's why the recognition error in the Figure 3 is corrected by adding this retrival LM.

Table 3: Statiscts on the top 200 news retrieved for a paragraph related to air transportation.

Theme	Number of news
Air transportation	158
Sports	19
Other transportation	13
Entertainment	4
Society	6

5.3 Comparison of different Bert-based embedding

Four different approaches to represent the document vector were compared: **BE1**, **BE2**, **BE3** and **BE4**. As shown in Tables 4 and 5, we can see that the performance of **BE1** was the worst because [CLS] token appears at the start of the text for classification tasks, and [CLS] token embedding does not convey much semantic information as a sentence representation. On the other hand, the difference of performance among **BE2**, **BE3** and **BE4** was negligible, and the reason is that the size of the news corpus is not large enough. As a consequence, we adopt **BE2** to obtain the document embedding in the following experiments.

Table 4: Comparison of different Bert-based embedding on CASIA-HWDB.

Experiment	CR(%)	AR(%)
BE1	94.82	94.15
BE2	94.89	94.24
BE3	94.89	94.23
BE4	94.89	94.23

Table 5: Comparison of different Bert-based embedding on ICDAR-2013.

Experiment	CR(%)	AR(%)
BE1	94.74	93.88
BE2	94.85	93.99
BE3	94.86	94.01
BE4	94.86	94.01

5.4 Comparison of different retrieval-based models

In this section, we evaluate different patterns in the retrieval-based language model, including two retrieval methods and different orders of N-gram models. The results are shown in Table 6. By comparison of different N-gram models, we can see that the 2-gram model performs the worst for both the TF-IDF and the BERT retrieval methods on both datasets, due to capturing very short contexts. After increasing to the 3-gram model, the accuracy is increased considerably, especially on the ICDAR-2013 dataset, where the AR is increased from 94.00% to 94.18%, and from 93.99% to 94.18% for TF-IDF and BERT methods respectively. However, the 4-gram model does not boost the accuracy further because the number of retrieved text is very limited and leads to very sparse 4-gram items. In other words, most of the 4-gram context scores will be learned by the back-off to the 3-gram estimation [59]. Comparing TF-IDF with BERT, we find that their performance is very similar on the ICDAR-2013 dataset, and note only a little improvement by BERT for the 3-gram model on CASIA-HWDB. The possible reason is that our query (i.e., transcript from the first-pass recognition) is not reliable due to some recognition errors.

5.5 Comparison with existing methods

Table 7 shows the comparison of existing methods and ours on the ICDAR-2013 dataset. In ours apporach, we apply common LM, retrival LM and the two additional improvements (see Sec.4.4). For the language model, we utilize a 5-gram Common LM and 4-gram Retrieval LM. As the number of retrieved text is very limited leading to very sparse 4-gram (or higher order) items, it is not necessary to utilize higher order LM (see Sec.5.4). For the character model, we apply a VGG-style CNN model which is the same as the baseline system [10, 11]. As shown in Sec. 3, we adopt the recognition system of [11] as the baseline, which is a weakly supervised learning method with only transcript level annotation under the over-segmentation framework. From the table, we can oberve that our method achieves a much higher accuracy than the baseline work [11],

Table 6: Results of different retrieval-based models on CASIA-HWDB and ICDAR-2013.

Retrival methods	N-gram	CASIA-HWDB		ICDAR-2013	
		CR(%)	AR(%)	CR(%)	AR(%)
TF-IDF	2-gram	94.91	94.28	94.87	94.00
	3-gram	94.93	94.27	95.02	94.18
	4-gram	94.92	94.26	95.03	94.19
BERT	2-gram	94.89	94.24	94.85	93.99
	3-gram	94.95	94.29	95.02	94.18
	4-gram	94.94	94.28	95.02	94.18

Table 7: Comparison with existing methods on the ICDAR-2013 dataset.

Methods	CR(%)	AR(%)
Wang et al. [60]	95.53	94.02
Wu et al. [10]	96.32	96.20
Xie et al. [61]	96.70	96.22
Peng et al. [62]	95.51	94.88
Wang et al. [11]	95.73	95.11
Peng et al. [63]	97.32	96.79
Ours (Common+Retrieval LM +Improvements)	96.13	95.48

which demonstrates the effectiveness of the proposed retrieval-based language model adaptation. By comparison with the other methods, our performance is even competitive to the accuracy in the work [10]. Note that [10] adopted the same over-segmentation based recognition framework, but optimized the recognition model under strong supervision with character-level annotation and elaborately integrated geometric context models. Although the recent work [63] optimized the model under the segmentation-free framework with transcript level annotation only, it utilized a large set of synthetic data to boost the accuracy.

6 Conclusion

In this paper, we proposed a retrieval-based language model for handwritten Chinese text recognition, which obtains the adaptive linguistic context during the recognition. Since the document to be recognized is unknown a priori, we engaged a two-pass recognition strategy. In the first-pass recognition, we take a common language model to recognize the document to output an initial transcript, which was used to retrieve related contents from a large text corpus. For the retrieval method, we evaluated both TF-IDF and four BERT-based embedding methods in our experiments. Finally, we built an adaptive language model from the retrieved contents, and combined with a common language model in the second-pass recognition to obtain the final transcript. We evaluated the proposed method on two benchmark datasets, and the extensive experimental results demonstrated the effectiveness of the proposed retrieval-based language model. In the future, we will consider an online Internet-based retrieval method to obtain

the related common sense knowledge to build the adaptive language model.

Acknowledgments. The work was funded by National Natural Science Foundation of China under no.61876154 and no.61876155; Jiangsu Science and Technology Programme (Natural Science Foundation of Jiangsu Province) under no.BE2020006-4.

References

[1] G. Nagy. Disruptive developments in document recognition. Pattern Recognition Letters, vol. 79, pp. 106–112, 2016.

[2] H. Fujisawa. Forty Years of Research in Character and Document Recognition—An Industrial Perspective. Pattern Recognition, vol. 41, pp. 2435-2446, 2008.

[3] R.-W. Dai, C.-L. Liu, B.-H. Xiao. Chinese Character Recognition: History, Status and Prospects. Frontiers of Computer Science in China, vol. 1(2), pp. 126-136, 2007.

[4] C.-L Liu, Y. Lu, editors. Advances in Chinese Document and Text Processing. book in Series on Language Processing, Pattern Recognition, and Intelligent Systems, vol. 2, World Scientific, 2017.

[5] Cheng-Lin Liu, Fei Yin, Qiu-Feng Wang, Da-Han Wang. ICDAR 2011 Chinese Handwriting Recognition Competition. 11th International Conference on Document Analysis and Recognition (ICDAR), pp. 1464-1469, 2011.

[6] Fei Yin, Qiu-Feng Wang, Xu-Yao Zhang, Cheng-Lin Liu. ICDAR 2013 Chinese Handwriting Recognition Competition. 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 1464-1470, 2013.

[7] Cheng Cheng, Xu-Yao Zhang, Xiaohu Shao, Xiang-Dong Zhou. Handwritten Chinese Character Recognition by Joint Classification and Similarity Ranking. 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 507-511, 2016.

[8] Xu-Yao Zhang, Yoshua Bengio, Cheng-Lin Liu. Online and offline handwritten Chinese character recognition: A comprehensive study and new benchmark. Pattern Recognition, vol. 61, pp. 348-360, 2017.

[9] Qiu-Feng Wang, Fei Yin, Cheng-Lin Liu. Handwritten Chinese Text Recognition by Integrating Multiple Contexts. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), vol. 34(8), pp. 1469-1481, 2012.

[10] Yi-Chao Wu, Fei Yin, Cheng-Lin 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.

[11] Zhen-Xing Wang, Qiu-Feng Wang, Fei Yin,

Cheng-Lin Liu. Weakly Supervised Learning for Over-Segmentation Based Handwritten Chinese Text Recognition. 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 157-162, 2020.

[12] C.-L. Liu, F. Yin, D.-H. Wang, Q.-F. Wang. CASIA Online and Offline Chinese Handwriting Databases. 11th International Conference on Document Analysis and Recognition (ICDAR), pp. 37-41, 2011.

[13] Qiu-Feng Wang, Erik Cambria, Cheng-Lin Liu, Amir Hussain. Common Sense Knowledge for Handwritten Chinese Text Recognition. Cognitive Computation, vol. 5 (2), pp. 234-242, 2013.

[14] Qiu-Feng Wang, Fei Yin, Cheng-Lin Liu. Unsupervised Language Model Adaptation for Handwritten Chinese Text Recognition. Pattern Recognition, vol. 47, pp. 1202-1216, 2014.

[15] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr. and T.M. Mitchell. Toward an Architecture for Never-Ending Language Learning. In Proceedings of the 24th Conference on Artificial Intelligence (AAAI), 2010.

[16] T. Mitchell, W. Cohen, E. Hruschka, et al. Never-Ending Learning. In Proceedings of the 29th Conference on Artificial Intelligence (AAAI), pp. 2302-2310, 2015.

[17] Fergus R, Fei-Fei L, Perona P, et al. Learning object categories from Google's image search. 10th IEEE International Conference on Computer Vision (ICCV), pp. 1816-1823, 2005.

[18] Nishizaki H., Sekiguchi Y. Word Error Correction of Continuous Speech Recognition Using WEB Documents for Spoken Document Indexing. In: Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead (ICCPOL), vol. 4285, pp. 213-221, 2006.

[19] Clemens Oertel, Shauna O'Shea, Adam Bodnar, D. Blostein. Using the web to validate document recognition results: experiments with business cards. In Proceedings of SPIE, Document Recognition and Retrieval XII, vol. 5676, pp. 17-27, 2005.

[20] Donoser M, Bischof H, Wagner S. Using web search engines to improve text recognition. International Conference on Pattern Recognition (ICPR), pp. 1-4, 2008.

[21] Donoser M., Wagner S., Bischof H.. Context information from search engines for document recognition. Pattern Recognition Letters, vol. 31, pp. 750-754, 2010.

[22] Qiu-Feng Wang, Fei Yin, Cheng-Lin Liu. Integrating language model in handwriting Chinese text recognition. 10th International Conference on Document Analysis and

Recognition (ICDAR), pp. 1036-1040, 2009.

[23] Y.X. Li, C.L. Tan, X.Q. Ding. A hybrid postprocessing system for offline handwritten Chinese Script recognition. Pattern Analysis and Applications, vol. 8, pp. 272-286, 2005.

[24] R.F. Xu, D.S. Yeung, D.M. Shi. A hybrid postprocessing system for offline handwritten Chinese character recognition based on a statistical language model. International Journal of Pattern Recognition and Artificial Intelligence, vol. 19(3), pp. 415-428, 2005.

[25] Rosenfeld R. Two decades of statistical language modeling: where do we go from here? IEEE, vol. 88(8), pp. 1270–8, 2000.

[26] J.R.Bellegarda. Statistical language model adaptation: review and perspectives. Speech Communication, vol. 42(1), pp. 93-108, 2004.

[27] B.C. Russell, A. Torralba, K.P. Murphy, et al. LabelMe: A Database and Web-Based Tool for Image Annotation. International Journal of Computer Vision, vol. 77(1-3), pp. 157–173, 2008.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), pp. 248-255, 2009.

[29] L. Fei-Fei. ImageNet: crowdsourcing, benchmarking & other cool things. CMU VASC Seminar, March, 2010.

[30] Hankz Hankui Zhuo, Qiang yang, Rong Pan, Lei Li. Cross-Domain Action-Model Acquisition for Planning Via Web Search. 21th International Conference on Automated Planning and Scheduling, pp. 298-305, 2011.

[31] Chen L, Lamel L, Gauvain J L, et al. Dynamic language modeling for broadcast news. 8th International Conference on Spoken Language Processing, 2004.

[32] Whitelaw C, Hutchinson B, Chung G Y, et al. Using the web for language independent spellchecking and autocorrection. Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 890-899, 2009.

[33] Bassil Y, Alwani M. OCR Post-Processing Error Correction Algorithm Using Google's Online Spelling Suggestion. Emerging Trends in Computing and Information Sciences, vol. 3(1), pp. 90-99, 2012.

[34] Oprean C, Likforman-Sulem L, Popescu A, et al. Using the Web to Create Dynamic Dictionaries in Handwritten Out-of-Vocabulary Word Recognition. 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 989-993, 2013.

[35] Oprean C, Popescu A, Popescu A, et al. Handwritten word recognition using Web resources and recurrent neural networks.

International Journal on Document Analysis and Recognition (IJDAR), vol. 18(4), pp. 287-301, 2015.

[36] Oprean C, Likformansulem L, Mokbel C, et al. BLSTM-based handwritten text recognition using Web resources. 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 466-470, 2015.

[37] Marti U. V., Bunke H. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition systems. International Journal of Pattern Recognition and Artificial Intelligence, vol. 15(01), pp. 65-90, 2001.

[38] N.-X. Li, L.-W. Jin. A Bayesian-Based Probabilistic Model for Unconstrained Handwritten Offline Chinese Text Line Recognition. IEEE International Conference on Systems, Man and Cybernetics, pp. 3664-3668, 2010.

[39] X.-D. Zhou, D.-H. Wang, F. Tian, C.-L. Liu, M. Nakagawa. Handwritten chinese/japanese text recognition using semi-markov conditional random fields. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), vol. 35(10), pp. 2413–2426, 2013.

[40] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin. A Neural Probabilistic Language Model. Journal of Machine Learning Research, vol. 3, pp. 1137–1155, 2003.

[41] T. Mikolov, M. Karafiat, L. Burget, J. H. Cernocky, S. Khudanpur. Recurrent neural network based language model. Interspeech 2010, 11th Annual Conference of the International Speech Communication Association, pp. 1045-1048, 2010.

[42] Yann N. Dauphin, Angela Fan, Michael Auli, David Grangier. Language Modeling with Gated Convolutional Networks. arXiv preprint arXiv:1612.08083v3, 2017.

[43] Irie K, Tüske Z, Alkhouli T, et al. LSTM, GRU, Highway and a Bit of Attention: An Empirical Overview for Language Modeling in Speech Recognition. Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, pp. 3519-3523,2016.

[44] Luong T, Kayser M, Manning C D. Deep Neural Language Models for Machine Translation. 19th Conference on Computational Natural Language Learning, pp. 305-309, 2015.

[45] J.R.Bellegarda. Exploiting latent semantic information in statistical language modeling. IEEE, vol. 88(8), pp. 1279-1296, 2000.

[46] Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning, vol. 42(1-2), pp. 177-196, 2001.

[47] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[48] Zecheng Xie, Zenghui Sun, Lianwen Jin, Hao Ni, Terry Lyons. Learning Spatial-Semantic Context with Fully Convolutional Recurrent Network for Online Handwritten Chinese Text Recognition. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), vol. 40(8), pp. 1903-1917, 2018.

[49] Gao J, Suzuki H, Yuan W. An empirical study on language model adaptation. ACM Transactions on Asian Language Information Processing, vol. 5(3), pp. 209-227, 2006.

[50] Liu X, Gales M J F, Woodland P C. Use of contexts in language model interpolation and adaptation. Computer Speech & Language, vol. 27(1), pp. 301-321, 2013.

[51] T.-H. Su, T.-W. Zhang, and D.-J. Guan. Corpus-Based HIT-MW Database for Offline Recognition of General-Purpose Chinese Handwritten Text. International Journal on Document Analysis and Recognition (IJDAR), vol. 10(1), pp. 27-38, 2007.

[52] Fei Yin, Qiu-Feng Wang, Xu-Yao Zhang, Cheng-Lin Liu. ICDAR 2013 Chinese Handwriting Recognition Competition. 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 1464-1470, 2013.

[53] Tong-Hua Su, Tianwen Zhang, De-Jun Guan, Hu-Jie Huang. Off-line recognition of realistic Chinese handwriting using segmentation-free strategy. Pattern Recognition, vol. 42(1), pp. 167-182, 2009.

[54] Z.-R. Wang, Jun Du, J.-S. Hu, and Yu-Long Hu. Deep Convolutional Neural Network Based Hidden Markov Model for Offline Handwritten Chinese Text Recognition. 4th Asian Conference on Pattern Recognition (ACPR), pp. 816-821, 2017.

[55] L. Xu, F. Yin, Q.-F. Wang, C.-L. Liu. An over-segmentation method for single touching Chinese handwriting with learning-based filtering. International Journal on Document Analysis and Recognition (IJDAR), vol. 17(1), pp. 91–104, 2014.

[56] Fink G A. Markov models for offline handwriting recognition: a survey. International Journal on Document Analysis and Recognition (IJDAR), vol. 12(4), pp. 269-298, 2009.

[57] Ronaldo Messina, J. Louradour. Segmentation-free handwritten Chinese text recognition with LSTM-RNN. 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 171-175, 2015.

[58] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805, 2019.

[59] Andreas Stolcke. SRILM-An extensible language modeling toolkit. 7th International Conference on Spoken Language Processing

(ICSLP), pp. 901-904, 2002.

[60] S. Wang, L. Chen, L. Xu, W. Fan, J. Sun, and S. Naoi. Deep knowledge training and heterogeneous cnn for handwritten chinese text recognition. 15th International Conference on Frontiers of Handwriting Recognition (ICFHR), pp. 84-89, 2016.

[61] Z.-C. Xie, Y.-X. Huang, Y.-Z. Zhu, L.-W. Jin, Y.-L. Liu, and L.-L. Xie. Aggregation cross-entropy for sequence recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6538-6547, 2019.

[62] D.-Z. Peng, L.-W. Jin, Y.-Q. Wu, Z.-P. Wang, and M.-X. Cai. A Fast and Accurate Fully Convolutional Network for End-to-End Handwritten Chinese Text Segmentation and Recognition. 15th International Conference on Document Analysis and Recognition (ICDAR), pp. 25-30, 2019.

[63] Peng, Dezhi and Jin, L. and Ma, Weihong and Xie, Canyu and Zhang, Hesuo and Zhu, Shenggao and Li, Jing. Recognition of Handwritten Chinese Text by Segmentation: A Segment-annotation-free Approach. IEEE Trans. on Multimedia, 2022.

[64] Z.-R. Wang, J. Du, and J.-M. Wang. Writer-aware CNN for parsimonious HMM-based offline handwritten Chinese text recognition. Pattern Recognition, vol. 100, pp. 107-102, 2020.