International Journal on Document Analysis and Recognition (IJDAR) (2023) 26:109–119
https://doi.org/10.1007/s10032-022-00419-2

ORIGINAL PAPER

Retrieval-based language model adaptation for handwritten Chinese text recognition

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

Received: 15 July 2022 / Revised: 5 September 2022 / Accepted: 1 October 2022 / Published online: 27 October 2022

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 [4–8]. Remarkably, the reported accuracy rate can even be higher than that of human recognition: 97.64% was reported in [7], while human only gets the accuracy of 96.13%. Nevertheless, automated unconstrained handwritten Chinese text recognition still remains unsatisfactory and actually far behind

✉ Qiufeng Wang
Qiufeng.Wang@xjtlu.edu.cn

Shuying Hu
Shuying.Hu20@student.xjtlu.edu.cn

Kaizhu Huang
kaizhu.huang@dukekunshan.edu.cn

Min Wen
Min.Wen@xjtlu.edu.cn

Frans Coenen
Coenen@liverpool.ac.uk

[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, UK