

基于多重规则和路径评价的在线中英文手写识别方法

付鹏斌, 刘鹏辉, 杨惠荣, 董澳静

(北京工业大学 信息学部, 北京 100124)

摘要: 手写文本识别方法主要应用于文本输入技术, 对人机交互领域的发展起关键作用。针对多数在线输入法无法识别中英文混合手写识别的问题, 提出一种在线中英文混合手写文本识别方法。通过对文本笔画进行基于水平相对位置、垂直重叠率、面积重叠率规则的整合以及连笔切分, 得到一系列字符片段, 同时利用笔画个数、宽高比、中心偏离、平滑度等几何特征和识别置信度, 对字符片段进行中英文分类。在此基础上, 根据分类结果并结合自然语言模型的路径评价及动态规划搜索算法, 分别对候选的中、英文字符片段进行合并处理, 得到待识别的中、英文字符序列, 并将其分别送入卷积神经网络的中、英文识别模型中, 得到手写文本识别结果。实验结果表明, 在线手写中英文混合文本识别正确率达93.67%, 不仅能切分在线手写中文文本行, 而且对包含字符连笔的在线手写中英文文本行也有较好的切分效果。

关键词: 在线手写识别; 中英文混合手写; 中英文分类; 文本行切分; 路径评价

开放科学(资源服务)标志码(OSID):



中文引用格式: 付鹏斌, 刘鹏辉, 杨惠荣, 等. 基于多重规则和路径评价的在线中英文手写识别方法[J]. 计算机工程, 2022, 48(3): 253-262.

英文引用格式: FU P B, LIU P H, YANG H R, et al. Online Chinese and English handwriting recognition method based on multiple rules and path evaluation[J]. Computer Engineering, 2022, 48(3): 253-262.

Online Chinese and English Handwriting Recognition Method Based on Multiple Rules and Path Evaluation

FU Pengbin, LIU Penghui, YANG Huirong, DONG Aoqing

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

[Abstract] Handwritten text recognition is mainly used in text input technology, which plays a key role in the development of human-computer interaction. To address the lack of functionality for Chinese and English mixed handwritten text recognition in most online input methods, an online Chinese and English mixed handwritten text recognition method is proposed. Through the integration of text strokes based on the horizontal relative position, vertical overlap rate, area overlap rate rules, and continuous stroke segmentation, a series of character segments are obtained. In addition, Chinese and English character segments are classified based on the number of strokes, aspect ratio, center deviation, smoothness, and recognition confidence. On this basis, according to the classification results, combined with the path evaluation of the natural-language model and dynamic programming search algorithm, the candidate and English character segments are combined to obtain the Chinese and English character sequences to be recognized, which are, respectively, sent to the Chinese and English recognition models of the Convolutional Neural Network (CNN) to obtain the handwritten text recognition results. The experimental results show that the recognition accuracy of the online handwritten Chinese and English mixed text is 93.67%, the proposed method can segment online handwritten Chinese text lines as well as online handwritten Chinese and English text lines containing characters.

[Key words] online handwriting recognition; mixed Chinese and English handwriting; Chinese and English classification; text line segmentation; path evaluation

DOI: 10.19678/j.issn.1000-3428.0060600

基金项目: 国家自然科学基金(61772048); 北京市自然科学基金(4153058)。

作者简介: 付鹏斌(1967—), 男, 副教授, 主研方向为图像处理、模式识别; 刘鹏辉, 硕士研究生; 杨惠荣(通信作者), 博士; 董澳静, 硕士研究生。

收稿日期: 2021-01-15 **修回日期:** 2021-02-24 **E-mail:** yanghuirong@bjut.edu.cn

0 概述

中英文混合文本识别是一个涉及字符切分、分类和识别的复杂上下文问题。目前,对于印刷体中英文混合文本识别的研究成果较多,且识别率较高^[1-2]。在手写文本识别方面,文献[3]将输入的手写中文文本行切分为字符片段,动态构建候选序列,并通过结合多种上下文信息搜索最佳路径,实时得到识别结果。文献[4]基于半马尔科夫条件随机场构建识别候选序列,自然融合候选片段置信度、几何和语义得分进行路径评价,并提出一种前后向阵列修剪算法,减少使用语言模型训练的计算量。文献[5]提出一种结合三元语言模型紧凑的CNN-BLSTM方法,使用多阶段训练方法实现多感受野机制,该方法达到了业界前沿的效果。文献[6]开发了“谷歌”在线手写识别系统,支持22种脚本和97种语言,实现了快速、高准确度的识别。文献[7]开发了在线手写识别系统,支持102种语言,识别效果较好。但上述在线手写文本识别方法的研究^[8]以及相关识别的研究^[9-11]仅能支持单一语言的文本识别,缺乏对中英文混合手写文本识别的支持。在商业领域,绝大多数国内输入法不支持中英文混合手写识别。法国公司Myscript开发的手写笔记软件nebo支持中英文混合手写识别,且识别效果在业界处于较高水平,但软件收费且核心识别技术不对外公开。因此,亟待研究一种有实用价值的在线中英文混合手写识别技术。

本文提出一种在线中英文混合手写文本识别方法,使用基于多重规则的切分算法得到字符片段,并在分类算法中进行中英文片段分类。在此基础上,结合自然语言模型和动态规划算法得到字符序列,分别送入基于CNN的在线手写识别模型,最终得到中英文混合手写文本识别结果。

1 预处理

联机手写数据通常是通过手写板、手写笔或鼠标得到的按书写笔画排序的点数据序列。在无约束情况下,手写文本常常会出现字符粘连、交错、噪声点以及文本行书写倾斜的情况,影响识别效果。特别是文本行的倾斜,会对后续文本切分和识别带来严重的影响,因此预处理阶段的重要工作除了降噪外就是进行文本行的倾斜矫正。由于文本行的字符中心大致符合直线拟合趋势,因此采用最小二乘法对手写文本行进行倾斜矫正。

令每个笔画的点坐标序列为 $P[(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)]$,则该笔画中心点为 $(x_{n/2}, y_{n/2})$ 。对文本行中所有笔画中心点 (x_i, y_i) , $0 < i \leq n$ 进行拟合,回归直线方程为: $\hat{y} = kx + b$ 。

结合最小二乘法思想:

$$Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - kx_i - b)^2 \quad (1)$$

其中: \hat{y}_i 是 y_i 的估计。在式(1)中分别对 k 、 b 求偏导数得:

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^n -2(y_i - b - kx_i) \quad (2)$$

$$\frac{\partial Q}{\partial k} = \sum_{i=1}^n -2(x_i y_i - bx_i - kx_i^2) \quad (3)$$

$$\frac{\partial Q}{\partial b} = \frac{\partial Q}{\partial k} = 0 \quad (4)$$

由式(4)得到拟合直线的参数值:

$$\begin{cases} k = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\ b = \bar{y} - k\bar{x} \end{cases} \quad (5)$$

其中: \bar{x} 、 \bar{y} 分别是 x 、 y 的平均值。

求得拟合直线后,计算文本行中心点,计算公式如式(6)所示:

$$(x, y)_{\text{center}} = \left(\frac{x_{\min} + x_{\max}}{2}, \frac{kx_{\min} + b + kx_{\max} + b}{2} \right) \quad (6)$$

其中: x_{\min} 、 x_{\max} 分别为文本行点坐标序列中 x 的最小值和最大值。

拟合直线与水平面的夹角为 α ,文本行围绕中心点进行中心旋转 α 度。倾斜矫正效果如图1所示。

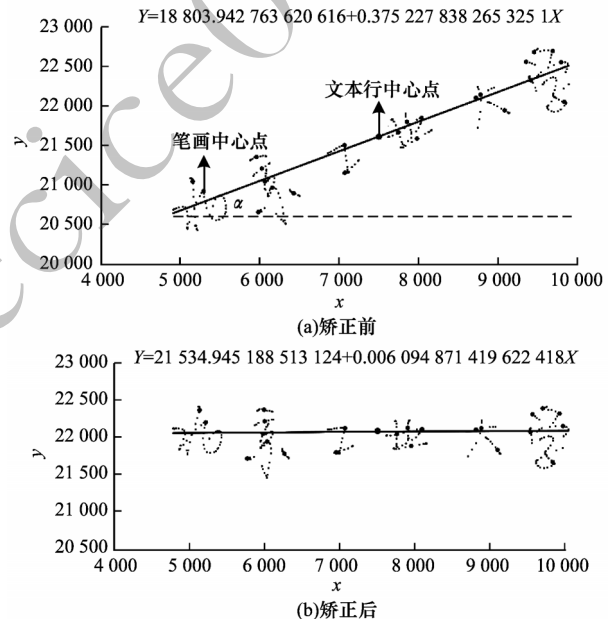


图1 倾斜矫正效果

Fig.1 Tilt correction effect

2 中英文混合文本分割

由于预处理后得到的笔画序列包含中、英文字符且可能存在字符重叠、粘连问题,因此需要进行字符切分,且字符切分算法的好坏将直接影响文本识别结果。欠切分方法得到的字符片段可能包含多个字符,会导致识别错误,而过切分方法得到的字符片

段通常包含单个字符或单个字符的子片段,可通过合并算法获得正确字符。因此,本文结合字符笔画的几何特征和空间特征,设计了基于多重规则和路径评价的中英文混合文本分割算法。

2.1 基于多重规则的中英文手写文本切分

对于在线手写文本而言,文本的切分就是笔画序列的正确分割和整合。本文结合水平相对位置、垂直重叠率、面积重叠率对笔画进行整合,相关定义如下:

定义1 垂直重叠率是相邻两个笔画在垂直方向重叠的比率。

$$p_o = \frac{l_o}{l_a + l_b - l_o} \times 100\% \quad (7)$$

根据定义1进行笔画整合的示意图如图2(a)所示,其中: l_o 为两笔画重叠长度; l_a 为笔画 a_1 的长度; l_b 为笔画 b_1 的长度。

定义2 面积重叠率是相邻两个笔画或笔画组合片段的最小外包矩形面积的重叠部分与两块面积中较小者的比值,其计算公式如式(8)所示:

$$a_o = \frac{S_o}{\min(S_c, S_d)} \times 100\% \quad (8)$$

根据定义2进行笔画整合的示意图如图2(b)所示,其中: S_o 为重叠面积; S_c 为笔画 c 的最小外包矩形的面积; S_d 为笔画组合片段 d 的最小外包矩形的面积。

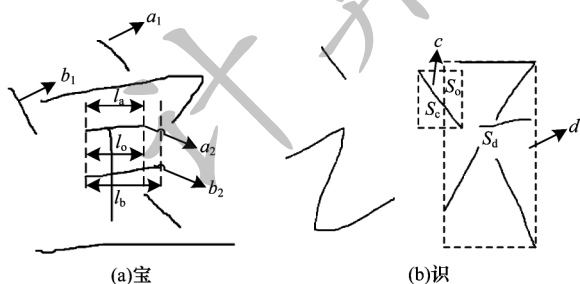


图2 笔画整合示意图

Fig.2 Schematic diagram of stroke integration

切分算法使用相邻两笔策略,假设2个相邻笔画 a 和 b , a 书写在前, b 书写在后,手写文本的笔画序列使用如下规则进行整合:

规则1 水平相对位置规则。若笔画 b 的最右端在笔画 a 最右端的左侧,则认为2个笔画属于同一字符片段,进行笔画整合,如图2(a)中 a_1 和 b_1 。

规则2 垂直重叠率规则。若笔画 a 和笔画 b 的垂直重叠率超过阈值(本文取50%),则认为2个笔画属于同一字符片段,进行笔画整合,如图2(a) a_2 和 b_2 所示。

根据上述两个规则,笔画序列中的某些笔画已完成了整合,称为笔画组合片段;若2个相邻笔画或笔画组合片段 c 和 d 不满足规则1、2,如图2(b)所示,则需使用如下规则进一步整合:

规则3 面积重叠率规则。若 c 和 d 的面积重叠率超过阈值(本文取40%),认为2个笔画或笔画组合片段属于同一字符片段,进行笔画整合。

笔画整合完成之后,若笔画组合片段的宽度值超过阈值(本文取笔画片段高度的1.8倍),则认为该笔画存在连笔情况,应进行切分。

根据大量统计和相关文献^[12]的研究可知,中文字符中的大部分连笔笔画均具有一个明显的特征,即存在一个较长的、方向稳定的笔画,且笔画的书写方向为从左下方到右上方。不仅中文连笔字符具有这个特征,而且英文也具有同样特征。另外,英文还有一种连笔情况,即字符笔画的书写方向为从左上方到右上方。连笔笔画还具有相同的位置特征,即连笔笔画的位置位于整个笔画的中间部位。依据这2个特征就可以找到字符连笔笔画并进行切分。

本文使用八方向特征来处理字符连笔的切分。八方向特征是特征提取中常用的方法^[13],它是四方向特征(水平、垂直、斜上、斜下)的细化,能够较好地提取8个方向的笔画,八方向分解图如图3所示,字符连笔情况多出现在 $D7$ 、 $D8$ 的方向特征图中。

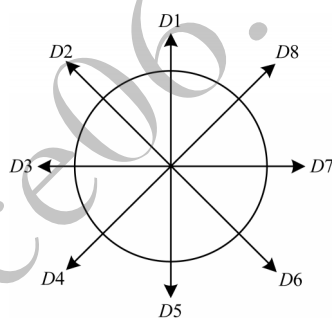


图3 八方向分解

Fig.3 8-direction decomposition

八方向特征图是通过计算字符点序列中每个点的方向生成。给定某一字符中的某个坐标点 p_k ,前一点为 p_{k-1} ,后一点为 p_{k+1} ,它的方向向量计算公式如下:

$$V_k = \begin{cases} \overrightarrow{p_k \times p_{k+1}}, & p_k \text{ 是起始点} \\ \overrightarrow{p_{k-1} \times p_{k+1}}, & p_k \text{ 是非终结点} \\ \overrightarrow{p_{k-1} \times p_k}, & p_k \text{ 是终止点} \end{cases} \quad (9)$$

得到方向向量 V_k 后,将其投影到8个方向上并进行向量分解,得到八方向特征图。

针对字符连笔书写情况,本文设计了一种检测连笔笔画并切分的方法,具体步骤如下:

步骤1 连笔检测。计算字符笔画或笔画组合片段的宽度值,如果宽度值大于阈值,那么认为该笔画或笔画组合片段存在连笔情况,筛选出该笔画或笔画组合片段,如图4所示,其中词组“中国”是一笔写出来的。

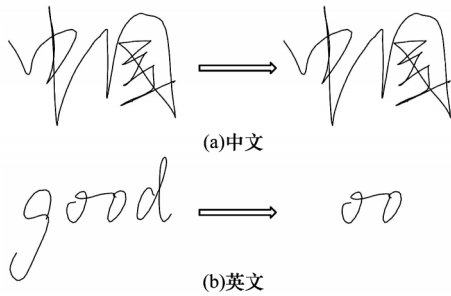


图4 笔画筛选示意图

Fig.4 Diagrammatic sketch of stroke filter

步骤2 根据筛选出的连笔笔画生成对应的八方向特征图,并根据连笔方向情况选择D7、D8方向图,如图5所示。

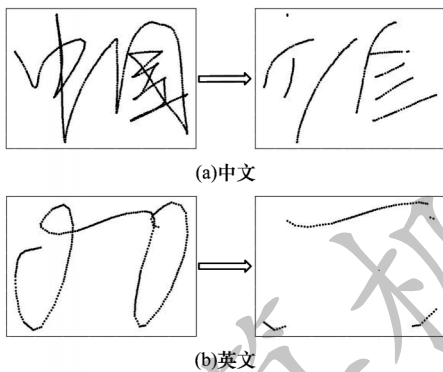


图5 特征方向图搜索

Fig.5 Search for feature direction diagrams

步骤3 搜索连笔笔画。在特征方向图中,搜索到的 $\left[\frac{1}{5}W, \frac{4}{5}W\right]$ 范围内较长的连续点序列即为连笔笔画, W 为特征方向图的宽度。

步骤4 连笔笔画切分。在现有研究中,切分点大多采用连笔笔画的中点,且在切分过程中并不会删除连笔区域的冗余点坐标数据,即只做切分,不做其他处理。但是,冗余的点坐标数据对字符识别准确率有一定影响。因此,本文定位2个切分点,并删除连笔部分的冗余笔迹,即2个切分点中间的点坐标数据做删除处理。切分点的位置定在笔画的除连笔部分剩余其他部分的最小外包矩形与字符笔迹的交点上,如图6所示,图中圆圈为确定的2个切分点。

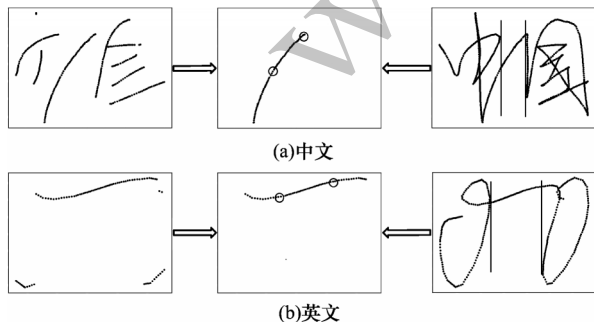


图6 确定切分点示意图

Fig.6 Schematic diagram of determining the point of division

经过以上步骤,得到字符连笔切分的效果如图7所示。

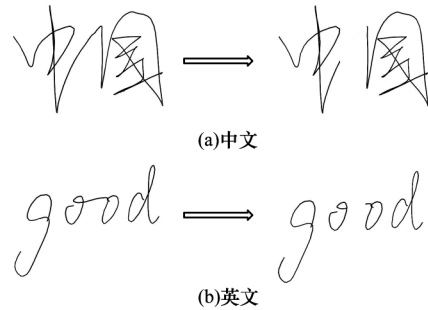


图7 连笔切分效果图

Fig.7 Effect drawing of continuous pen segmentation

预处理完成的手写文本笔画序列通过水平相对位置、垂直重叠率、面积重叠率3个规则进行整合,之后进行连笔检测并切分,最终得到切分完成的中英文字符片段序列,切分算法如算法1所示。

算法1 基于多重规则的中英文手写切分算法

输入 笔画列表 $S=(S_0, S_1, \dots, S_n)$

输出 切分完成的字符片段列表 S'

1. 遍历笔画列表 S , 得到每个笔画的水平、垂直方向最大值 $S_{xmin}, S_{xmax}, S_{ymin}, S_{ymax}$;
2. 如果 $S_{i+1}.xmax \leq S_i.xmax$, 则笔画合并;
3. 如果 $p_0 \geq 0.5$, 笔画合并;
4. 计算每个笔画的最小外包矩形面积, 以及相邻笔画间的重叠面积;
5. 如果 $a_0 \geq 0.4$, 笔画合并;
6. 计算每个笔画或笔画组合片段的宽度;
7. 如果宽度值 \geq 高度值的1.8倍, 生成该笔画或笔画组合片段的八方向坐标点序列特征图;
8. 根据连笔方向选择方向图, 搜索连笔特征点序列;
9. 求笔画的除连笔部分剩余其他部分的最小外包矩形与点序列的交点, 作为切分点;
10. 删除2个切分点中间点序列数据, 并进行笔画切分。

2.2 基于几何特征和识别置信度的字符片段分类算法

由于两种语言类别数相差较大、字符结构不同、相关度不高,混合识别不能达到较好的效果。因此,通过基于多重规则的切分算法得到的字符片段需要进行中英文分离,把分离后的中、英文字符片段序列进行合并,之后分别送入单语言模型进行识别。中英文混合字符片段的分离通过基于笔画个数、宽高比、中心偏离距离、平滑度等几何特征和字符片段识别置信度相结合的分类算法来完成。

如图8所示,本文提取的字符片段几何特征包括字符片段的宽度、高度、宽高比、笔画个数、字符间距、中心偏移距离、平滑度。具体定义如下:

h : 字符片段的高度值。

w : 字符片段的宽度值。

h_w : 字符片段的宽高比值。

n : 字符片段的笔画个数。

d : 字符片段间的距离。

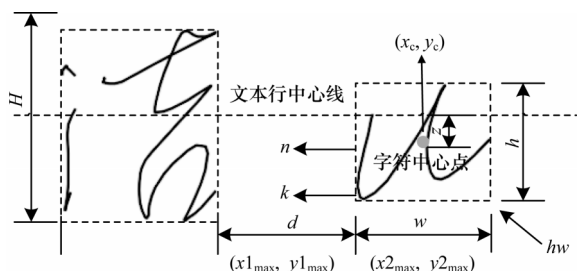


图8 几何特征提取

Fig.8 Geometric feature extraction

定义4 文本行高度估计值 H 。对所有笔画按高度值升序排序;如果输入笔画数小于阈值 β ,则 H 取所有笔画中的高度值最大的笔画高度。如果输入笔画数大于阈值 β ,则 H 取笔画序列中高度值较大的1/2笔画的平均值。设置阈值 β (本文为10)是为了防止输入笔画过少,导致 H 估算偏差较大。

定义5 中心偏移距离 z 。字符片段中心点与文本行中心线的距离,字符片段中心点在文本行中心线下方为负值,在文本行中心线上方为正值。

$$z = x_c - \frac{H}{2} \quad (10)$$

定义6 字符片段的笔迹平滑度 k ,反映了书写笔迹的弯曲程度。每个笔画上随机选择5个点,计算每个点的局部曲率值。假设笔画 L 由坐标点 $P[(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)]$ 构成,则对应的方程为 $y=f(x)$,笔画 L 在点 $M(x, y)$ 处切线的斜率为 $y'=\tan\alpha$,则

$$y'' = \sec^2\alpha \times \frac{d\alpha}{dx} \quad (11)$$

由式(11)推导得:

$$d\alpha = \frac{y''}{1+(y')^2} dx \quad (12)$$

又 $ds = \sqrt{1+y'^2} dx$,其中 ds 称为曲线弧长的微分,则局部曲率为:

$$K = \frac{|y''|}{(1+y'^2)^{3/2}} \quad (13)$$

则笔迹平滑度为:

$$k = \frac{\sum_{i=1}^5 K_i}{5} \quad (14)$$

定义7 识别置信度是为了估计字符识别结果的准确性。本文识别置信度为卷积神经网络输出的Softmax概率值。

根据以上特征,本文设计了基于几何特征的粗分类器和基于识别置信度的细分类器。

将以下4个特征作为粗分类器的主要依据:

1)中文字符片段的笔画个数明显多于英文字符片段;

2)中文字符片段的高度高于英文字符;

3)英文字符片段笔迹的平滑度高于中文字符;

4)英文字符中心点位于文本行中心线下方。

粗分类器能够将大部分字符片段正确分类,而无法分类的字符片段将进入细分类器。细分类器包含了基于CNN的在线手写英文识别模型和在线手写汉字识别模型。进入细分类器后,每个字符片段将会得到2个模型对应的识别置信度,若手写汉字识别模型的识别置信度较大,则归为中文片段,否则归为英文片段。具体的字符片段分类算法如算法2所示。

算法2 基于几何特征和识别置信度的分类算法

输入 字符片段列表 $S=(S_0, S_1, \dots, S_n)$

输出 英文字符片段 E ,中文字符片段 C

- 1.初始化英文字符片段列表 E 和中文字符片段列表 C ;
- 2.遍历字符片段列表,进入几何特征分类器;
- 3.如果 $n \geq 5$ 或者 $h \geq$ 行高以及 $w \geq h$,加入片段列表 C ;
- 4.如果 $h_w < 1, z < 0, d \geq$ 行高,加入片段列表 E ;
- 5.定义笔画的曲线函数,求得字符片段平均曲率 k ;
- 6.如果 $k \geq 0.6$,加入片段列表 E ,如果 $k \leq 0.2$,加入片段列表 C ;

- 7.遍历字符片段列表,进入细分类器;
- 8.计算字符片段的中、英文识别置信度 c_0, c_1 ;
- 9.如果 $c_0 \geq c_1$,则加入片段列表 E ,否则加入片段列表 C 。

针对算法2中分类器的限定条件作如下说明:

在一般情况下,英文字符的笔画最多为3个,若 $n \geq 5$,可以认为该字符片段为中文;若 $h_w < 1, z < 0, d \geq$ 行高,即字符片段的宽度小于高度、字符片段的中心点位于文本行中心线的下方且字符片段间的距离相对较大,可以认为该字符片段为英文;经过对大量英文字符的平均曲率进行计算统计,发现 k 的最小值约为0.4,若 $k \geq 0.6$,可以认为该字符片段为英文,若 $k \leq 0.2$,则认为该字符片段为中文。

2.3 结合自然语言模型和动态规划算法的路径评价

通过上述文本切分和字符片段分类的算法,得到了字符串基本切分片段,由于中文字符笔画数多、结构复杂,且大部分字符不能一笔完成,因此字符片段中存在欠合并的现象。所以,本文结合自然语言模型和动态规划的路径评价算法搜索最优的字符合并路径。基于字符片段识别框架,首先将一个字符串切分为基本片段,接着将一个或者多个基本片段合并为候选字符,生成候选识别网络,如图9所示。候选字符首先被基于CNN的在线手写中、英文字符识别模型进行识别并得到识别置信度;然后结合自然语言模型,通过路径评价算法得到路径评分;最后,使用路径搜索算法选出评分最优的合并路径,得到合并完成的待识别字符序列。

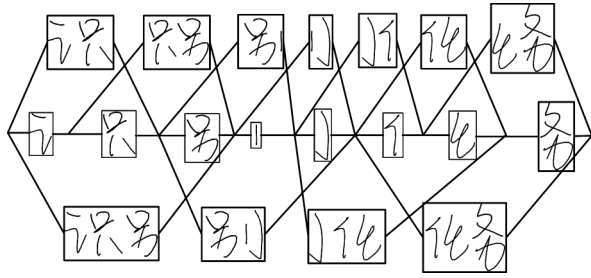


图9 部分候选识别网络

Fig.9 Part of the candidate identification network

对于自然语言概率模型而言,假设文本行 S 的识别结果为 $R=(R_1, R_2, \dots, R_n)$,以 $P(S)$ 代表该识别结果的概率,则概率评估函数为:

$$P(R) = P(R_1, R_2, \dots, R_n) \quad (15)$$

根据链式法则,概率评估函数可转化为:

$$P(R) = P(R_1)P(R_2|R_1) \cdots P(R_n|R_1, R_2, \dots, R_{n-1}) \quad (16)$$

由于输入法对识别时间要求较高,考虑到计算量以及语料库的大小,本文使用N-gram模型的二元语言模型来计算式(16)的概率,因此:

$$P(R) = P(R_1)P(R_2|R_1) \cdots P(R_n|R_{n-1}) \quad (17)$$

其中:每个字符出现的概率只取决于前一个字符。

本文训练的自然语言概率模型所使用的数据库为搜狗实验室公开发布的搜狐新闻数据(SogouCS)以及全网新闻数据(SogouCA)。在不考虑其他模型的情况下,自然语言模型概率最大的字符组合即为最佳的识别路径。如图10所示,为字符片段通过计算自然语言模型概率得到的最优识别路径。

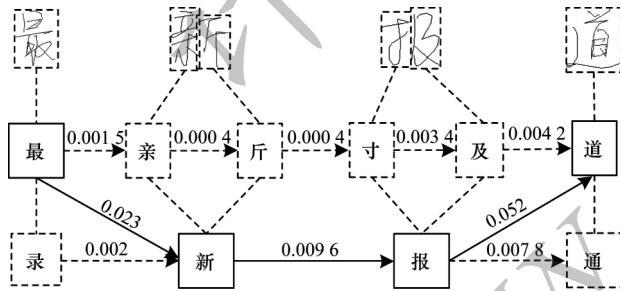


图10 二元语言模型路径

Fig.10 Binary language model path

对字符片段组合加以规则约束,可以减少候选片段组合的数量,进而提高路径搜索效率。本文定义规则如下:

- 1) 候选字符合并个数不超过3个;
- 2) 候选字符合并后的宽度不超过高度的2倍;
- 3) 待合并的2个候选字符的水平距离不超过候选字符宽度的1.5倍。

基于规则的组合策略,对候选字符片段进行组合,一次组合称为路径 s 。组合后的片段分别提取特征得到 $X=(x_1, x_2, \dots, x_n)$,如果假设字符串识别结果为 $R=(r_1, r_2, \dots, r_n)$,那么该识别结果的后验概率^[14]为:

$$P(R|X) = \sum_s P(s|X)P(R|X^s) \quad (18)$$

其中: $P(s|X)$ 代表在获取特征 X 的情况下组合路径 s 的后验概率, $P(R|X^s)$ 代表在获取组合路径 s 的情况下识别结果 R 的后验概率。

考虑到字符片段组合后包含大量的路径以及能够避免大量的计算,最优结果可以近似计算为:

$$R^* = \arg \max_{sR} P(s|X)P(R|X^s) \quad (19)$$

其中: $P(s|X)$ 以判断该字符是否有效切分来表示当前路径的概率。由于本文使用的文本行数据库没有切分点数据,以及加入了基于规则的组合策略,因此本文没有使用该分类器的概率值。

因为本文识别技术主要应用于输入法,没有考虑符号、数字等其他字符,所以没有使用几何模型,仅使用了单字符识别概率值和自然语言模型。 $P(R|X^s)$ 可以表示为:

$$P(R|X^s) = \frac{1}{p} \prod_{i=1}^n p(r_i|x_i)p(R) \quad (20)$$

其中: p 为常数; $p(r_i|x_i)$ 为字符分类的结果; $p(R)$ 为自然语言模型的结果。

考虑到不同分类器的权重问题以及克服路径长度的影响,本文使用了修正的片段宽度加权方法,通过公式两边取对数,并在每一项前加入权值来解决权重问题;通过归一化字符片段宽度以及语言模型对整个长度做归一化来克服路径长度的影响。计算公式如下:

$$f(X^s, R) = \sum_{i=1}^n \left(\frac{w_i}{w} \log_a p_i^0 + \frac{\lambda_1}{n} \times \log_a p_i^1 \right) \quad (21)$$

其中: w_i 代表第 i 个路径中片段的宽度; $\log_a p_i^0$ 代表单字符分类器概率结果的对数值; $\log_a p_i^1$ 代表自然语言模型的概率结果的对数值; λ_1 为自然语言模型参数。

通过路径评价算法得到本次组合的评分,接下来,要从所有组合路径中选择一条评分最高的路径。虽然采用了基于规则的组合策略对字符片段组合加以约束,但仍有大量的组合方式。若对全部的组合方式进行计算,文本识别性能将会变得极为低效。所以,快速有效的路径搜索算法对提高文本识别的性能至关重要。路径评价函数是计算所有候选字符得分的加和值,取最大加和值的字符路径为最优路径,因此可以使用动态规划算法进行路径搜索,在搜索的中间节点中保留一条最优路径,从而使路径搜索快速且有效。路径搜索的算法如算法3所示。

算法3 路径搜索算法

输入 字符片段序列

输出 最优片段组合方式

1. 遍历字符片段序列,通过规则组合为候选字符;
2. 遍历候选字符;
3. 计算候选字符识别置信度得分;
4. 将候选字符进行语言模型匹配,计算语言模型得分;
5. 通过路径评价函数得到最终路径得分;
6. 保留得分最高的候选字符作为当前组合的候选结果;
7. 从最后一个候选结果开始回溯,直到起始位置。

3 基于CNN的在线手写字符识别

在文字识别领域,CNN模型取得了巨大的成功^[15-17]。本文把前述分割得到的中、英文字符序列分别送入CNN模型并进行训练识别。

3.1 CNN模型

单字符的识别采用了经典的CNN模型LeNet-5^[18-19],并在其基础上进行改进:

1)输入输出层:输入尺寸修改为本文输入尺寸,后续各层的尺寸相应改变,在输出层添加Softmax激活函数,从而加速模型收敛,缓解Sigmoid函数发生梯度消失的问题。

2)卷积层、池化层:当分类数越大时,模型所需要的特征信息也相对增多,于是增加模型的层数和特征图数量;按照两层卷积层、一层池化层的组合排列,添加了6层卷积层和2层池化层,特征图的数量从50到400逐层增加。卷积层采用 3×3 大小的滤波器,池化层采用 2×2 的滤波器。

3)全连接层:本文采用2个全连接层,每层有1 024个单元。由于训练样本有限,模型参数过多、模型层次过深,导致训练时易发生过拟合现象。为避免该现象的发生,本文加入了dropout算法。

基于以上改进,本文设计并实现了14层CNN模型,模型包括8层卷积层、4层池化层、2层全连接层,如图11所示。

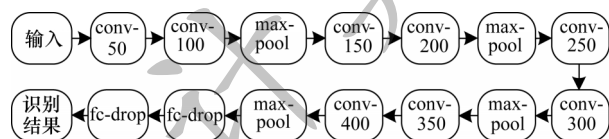


图11 CNN模型结构

Fig.11 Structure of CNN model

3.2 在线手写英文字符识别

由于英文字符类别数较少,因此本文将提取的单字符特征图作为网络模型的输入。

对手写字符进行线性插值、平滑、归一化等预处理后,通过计算该字符的最小外包矩形得到字符边界,将其平均分为 $12\times 12=144$ 块,使该字符的所有点坐标落入小方块中,统计每个小方块中字符点坐标的个数,若个数大于0,则该方块的特征值为1,否则为0;得到 12×12 的特征图,特征图提取过程如图12所示。最终,把得到的特征图作为CNN的输入。

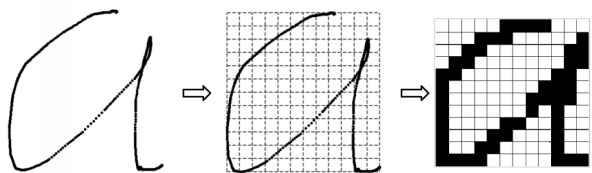


图12 特征图提取

Fig.12 Feature map extraction

模型训练的数据集为哈尔滨工业大学收集的HIT-OR3C^[20]中的Letter子集以及华南理工大学收集的SCUT-OUCH2009^[21]英文字母子集。

3.3 在线手写中文汉字识别

文中用于在线手写中文汉字识别的流程大致分为3个步骤:预处理,特征提取,CNN训练识别。

首先,对字符进行预处理。主要有长宽比映射关系归一化、平滑、线性插值、加入虚拟笔画等,加入虚拟笔画有助于字形的区分(这里的虚拟笔画是指上一笔结束点和下一笔起始点之间的连线,也就是当书写完成当前笔画后准备书写下一笔画时,笔尖脱离纸面在空中划出的轨迹),如图13所示。

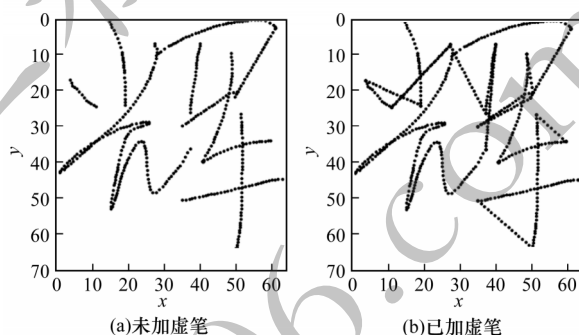


图13 虚拟笔画

Fig.13 Virtual stroke

然后,将预处理后得到的字符点坐标序列进行方向分解,生成 $D1\sim D8$ 这8个方向的特征,即点坐标的八方向特征图提取。

虽然CNN在数据处理时,不需要显式构造特征,但原图输入最具有代表性,且将对最终的结果产生积极的影响。因此,本文把8方向特征图加上原图构成9通道特征图(由9张 32×32 像素的图组成)作为CNN的输入,如图14所示。



图14 9通道特征图

Fig.14 9 channel characteristic diagram

模型训练的数据集为中科院收集的CASIA-OLHWDB 1.0^[22]、CASIA-OLHWDB 1.1以及HIT-OR3C的中文子集。

4 实验与结果分析

本文所提在线中英文混合手写文本识别方法通过预处理、文本切分、字符片段分类、字符片段合并以及单字符识别,最终得到文本识别结果,识别流程如图15所示。

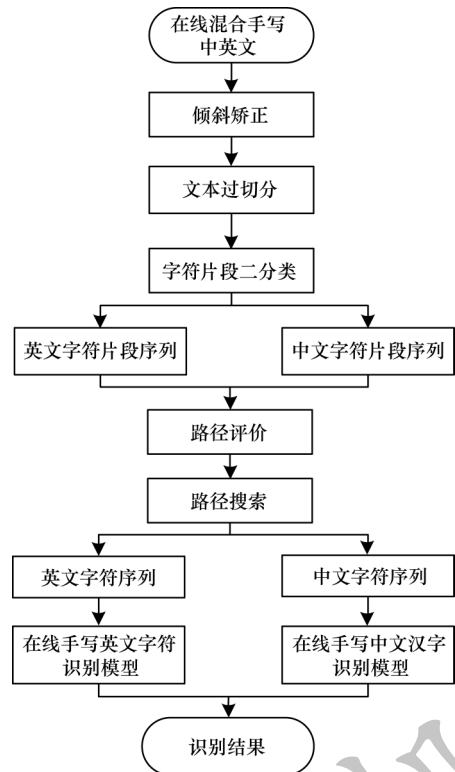


图15 本文方法识别流程

Fig.15 Identification procedure of the method in this paper

4.1 实验数据

选用公开的在线手写中文文本数据集 CASIA-OLHWDB2.0-2.2^[23] 以及本文采集的在线混合手写中英文文本行数据集 OH-C_E_TextDB, 并将常用中文字词和英文单词随机重组为文本样本, 共计 3 000 条, 30 名采集人员(大学生 10 名, 研究生 10 名, 教师 10 名)进行手写数据采集, 每人随机采集 100 条。部分文本样本如表 1 所示。

表 1 部分样本数据

Table 1 Partial sample data

序号	中英文混合文本行样本
1	菠 津 简 洁 umbrella coral
2	束 income 深邃 alone 逢
3	bedroom easily 终 简洁 垄
4	漂亮 娜 办 laughter 高耸
5	assistance 季 消瘦 舒心 衫
6	agriculture surface 衍 合身 春联
⋮	⋮
3 000	whisper 优 欣喜 优秀 boycott

4.2 结果分析

本文通过切分正确率 R_c 和切分有效率 R_v 来验证过切分算法的性能, 计算公式如下:

$$\begin{cases} R_c = \frac{M_c}{M_t} \\ R_v = \frac{M_t}{M_z} \end{cases} \quad (22)$$

其中: M_c 表示真实切分点与正确切分点的匹配个数, 即正确切分个数; M_t 表示真实切分点总数; M_z 表示所有切分点的个数。 R_c 的值越大说明命中正确切分点的数量越多, R_v 的值越大说明字符出现过切分的情况更少。

表 2 和表 3 分别给出了本文切分算法及其他切分算法在 CASIA-OLHWDB 2.0-2.2 数据集、OH-C_E_TextDB 数据集上的切分性能测试结果。

表 2 不同方法在 CASIA-OLHWDB 2.0-2.2 数据集下的切分对比实验结果

Table 2 Experimental results of segmentation comparison of different methods under CASIA-OLHWDB 2.0-2.2 data set

切分方法	切分正确率/%	切分有效率/%	切分时间/(行·ms ⁻¹)
文献[24]方法	97.82	69.38	5.42
文献[25]方法	96.31	66.93	5.03
本文方法	98.95	70.26	4.47

表 3 不同方法在 OH-C_E_TextDB 数据集下的切分对比实验结果

Table 3 Experimental results of segmentation comparison of different methods under OH-C_E_TextDB data set

切分方法	切分正确率/%	切分有效率/%	切分时间/(行·ms ⁻¹)
文献[24]方法	93.53	67.12	0.54
文献[25]方法	94.25	68.24	0.41
本文方法	99.13	71.15	0.32

通过表 2 和表 3 的对比实验结果可以发现, 本文切分算法相比其他切分算法的切分正确率、切分有效率均有所提高, 并且减少了切分耗时。相比表 2, 本文切分算法在表 3 的切分正确率、切分有效率有所提升, 而其他 2 种切分算法均有所下降。究其原因, 发现 OH-C_E_TextDB 数据集中有大量的英文连笔和中文连笔数据, 而其他 2 种算法对字符连笔情况处理效果较差, 尤其是英文连笔的切分。图 16 给出了 3 种切分方法在实际数据中的对比图。通过实验结果可知, 本文切分算法不仅对在线手写中文文本行切分有效, 而且对包含字符连笔的在线混合手写中英文文本行切分有较好的切分效果。



图 16 不同切分方法在实际数据中的对比

Fig.16 Comparison of different segmentation methods in actual data

为证明本文方法的有效性, 采用字符串编辑距离的思想, 具体用了 3 个评判标准: 文本行识别率 (Row Rate, RR), 文本正确率 (Correct Rate, CR), 文

本精确率(Accurate Rate, AR),计算公式如下:

$$\begin{cases} R_{RR} = \frac{T_r}{T_z} \\ C_{CR} = \frac{N_i - D_e - S_e}{N_i} \\ A_{AR} = \frac{N_i - D_e - S_e - I_e}{N_i} \end{cases} \quad (23)$$

其中: T_r 代表识别完全正确的文本行数; T_z 代表识别的总文本行数; N_i 代表每行真实文本个数; D_e 代表真实字符与识别结果对比的删除错误数目; S_e 代表真实字符与识别结果对比的替换错误数目; I_e 代表真实字符与识别结果对比的插入错误数目。

在 OH-C_E_TextDB 数据集上的实验结果表明,本文方法对在线混合手写中英文文本的识别正确率、文本识别精确率以及文本行识别率分别可达 93.67%、92.25%、91.53%,验证了本文在线中英文混合手写文本识别方法的有效性。

把本文识别方法应用到在线输入系统中,该系统利用动态维护候选字符序列的思想,进行实时切分识别。对系统进行实时性分析发现,每当新笔画输入时,系统动态更新笔画序列并进行切分、分类、合并以及识别,当抬笔时间超过 1 s 时,系统判定字符输入结束并立即输出识别结果。系统识别效果如图 17 所示。图 18 展示了输入“online 手写中 English 混合识别”的具体识别过程。由图 18 可知,字符连笔可以被正确分割并识别;在书写中文字符“识”的过程中,先写“讠”,系统更容易认为是英文字符“i”,而当把另一部分“只”书写完成后,正确识别为“识”。

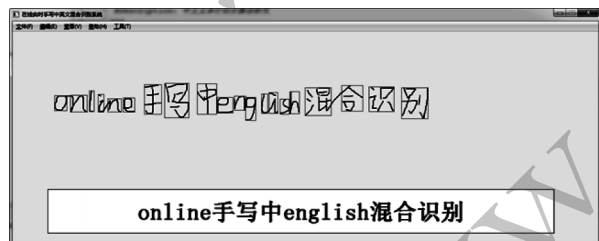


图 17 在线中英文手写识别效果

Fig.17 Online Chinese and English handwriting recognition effect

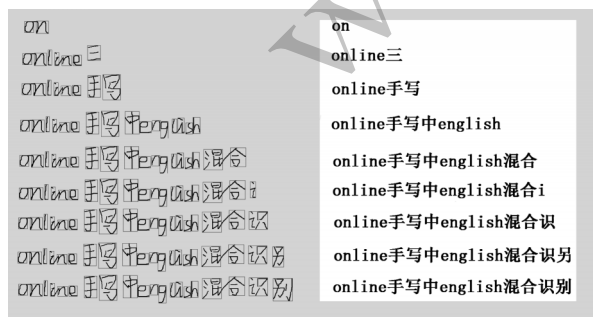


图 18 文本识别过程

Fig.18 Text recognition process

5 结束语

针对多数在线输入法不支持中英文混合手写文本识别的问题,本文提出一种在线中英文混合手写文本识别的新方法。通过切分文本得到字符片段,并使用分类算法对字符片段进行分类。此外,结合自然语言模型和动态规划算法将字符片段合并为字符序列,并通过在线手写识别模型得到中英文混合手写文本识别结果。实验结果表明,相比其他切分算法,本文算法对在线手写中文文本行及包含字符连笔的在线混合手写中英文文本行均能较好地进行切分,在线中英文混合手写文本识别正确率达 93.67%。但本文研究的文本识别方法没有考虑标点符号、数字等特殊字符,下一步将通过研究中文、英文、数字、符号 4 类别的识别方法,完善本文模型。

参考文献

- [1] 王恺,王庆人. 中英文混合文章识别问题[J]. 软件学报, 2005, 16(5): 149-161.
WANG K, WANG Q R. Identification of mixed articles in chinese and english [J]. Journal of Software, 2005, 16(5): 149-161. (in Chinese)
- [2] 任荣梓,高航. 基于反馈合并的中英文混排版面 OCR 技术研究[J]. 计算机技术与发展, 2017, 27(3): 39-43.
REN R Z, GAO H. Research on OCR technology of Chinese and English mixed layout based on feedback merging [J]. Computer Technology and Development, 2017, 27(3): 39-43. (in Chinese)
- [3] WANG D H, LIU C L, ZHOU X D. An approach for real-time recognition of online Chinese handwritten sentences [J]. Pattern Recognition, 2012, 45(10): 3661-3675.
- [4] ZHOU X D, WANG D H, TIAN F, et al. Handwritten Chinese/Japanese text recognition using Semi-Markov conditional random fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(10): 2413-2426.
- [5] CHEN K, TIAN L, DING H, et al. A compact CNN-DBLSTM based character model for online handwritten Chinese text recognition[C]//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Washington D. C., USA: IEEE Press, 2017: 1068-1073.
- [6] KEYSERS D, DESELAERS T, ROWLEY H A, et al. Multi-language online handwriting recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1180-1194.
- [7] CARBUNE V, GONNET P, DESELAERS T, et al. Fast multi-language LSTM-based online handwriting recognition [J]. International Journal on Document Analysis and Recognition, 2020, 23(6): 89-102.
- [8] KANG L, RIBA P, VILLEGAS M, et al. Candidate fusion: integrating language modelling into a sequence-to-sequence handwritten word recognition architecture [J]. Pattern Recognition, 2020, 112(6): 107-113.
- [9] SIMAYI W, IBRAHIM M, ZHANG X Y, et al. A benchmark for unconstrained online handwritten Uyghur word recognition[J]. Document Analysis and Recognition, 2020, 23(3): 205-218.
- [10] YAAGROUP K M M, MUSTAFA M E. Online arabic handwriting characters recognition using deep learning

- [EB/OL]. [2020-12-10]. https://www.researchgate.net/publication/330076104_IJARCCE.
- [11] GHOSH S, CHATTERJEE A, SEN S, et al. CTRL-CapTuRedLight: a novel feature descriptor for online Assamese numeral recognition[J]. *Multimedia Tools and Applications*, 2021, 80(11):30033-30056.
- [12] YAO Z, DING X, LIU C. On-line handwritten Chinese word recognition based on lexicon[C]//*Proceedings of the 18th International Conference on Pattern Recognition*. Washington D. C., USA: IEEE Press, 2006:320-323.
- [13] BAI Z, HUO Q. A study on the use of 8-directional features for online handwritten chinese character recognition[C]//*Proceedings of the 8th International Conference on Document Analysis and Recognition*. Washington D. C., USA: IEEE Press, 2006:262-266.
- [14] WANG Q F. Handwritten Chinese text recognition by integrating multiple contexts[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(8):1469-1481.
- [15] 施恩, 李骞, 顾大权等. 基于局部特征的卷积神经网络模型[J]. *计算机工程*, 2018, 44(2):282-286.
SHI E, LI Q, GU D Q. Convolutional neural network model based on local feature[J]. *Computer Engineering*, 2018, 44(2):282-286. (in Chinese)
- [16] GUPTA D, BAG S. CNN-based multilingual handwritten numeral recognition: a fusion-free approach[J]. *Expert Systems With Applications*, 2021, 165:113784-113792.
- [17] LAN W, WU L, JIA X Y, et al. Research on application of an improved deep convolutional neural network in handwritten character recognition[J]. *Journal of Physics. Conference Series*, 2020, 1629(1):12002-12013.
- [18] YEKTA S C, ERDEM K. Automatic CNN-based arabic numeral spotting and handwritten digit recognition by using deep transfer learning in ottoman population registers[J]. *Applied Sciences*, 2020, 10(16):5430-5441.
- [19] LECUN Y, JACKEL L D, BOTTOU L, et al. Comparison of learning algorithms for handwritten digit recognition [C]//*Proceedings of International Conference on Artificial Neural Networks*. Washington D. C., USA: IEEE Press, 1995, 53-60.
- [20] ZHOU S, CHEN H, et al. HIT-OR3C: an opening recognition corpus for Chinese characters[C]//*Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*. Boston MA, US:[s. n.], 2010:223-230.
- [21] JIN L, GAO Y, LIU G, et al. SCUT-COUCH2009—a comprehensive online unconstrained Chinese handwriting database and benchmark evaluation [J]. *International Journal on Document Analysis and Recognition*, 2011, 14(1):53-64.
- [22] WANG D H, LIU C L, YU J L, et al. CASIA-OLHWDB1: a database of online handwritten Chinese characters[C]//*Proceedings of the 10th International Conference on Document Analysis and Recognition*. New York, USA: ACM Press, 2009:1206-1210.
- [23] WU C, FAN W, HE Y, et al. 1st place in ICDAR 2013 chinese handwriting recognition competition [C]//*Proceedings of IapR International Conference on Document Analysis and Recognition*. Washington D. C., USA: IEEE Press, 2013:1464-1470.
- [24] 陈肇欣. 联机手写汉字文本行识别算法研究[D]. 广州: 华南理工大学, 2013.
CHEN Z X. Research on on-line handwritten Chinese character text line recognition algorithm [D]. Guangzhou: South China University of Technology, 2013. (in Chinese)
- [25] 邱理权. 基于多种上下文信息的联机手写中文文本识别方法及系统实现[D]. 广州: 华南理工大学, 2017.
QU L Q. Multi-contexts based online handwritten Chinese text recognition methods and system implementation [D]. Guangzhou: South China University of Technology, 2017. (in Chinese)

编辑 赖玉玲