文章编号: 2095-2163(2018)02-0126-03 中图分类号: TP391.41 文献标志码: A

用于手写汉字识别的文本分割方法

雷 鑫, 李俊阳, 宋 宇, 赛琳伟

(河海大学 常州校区数理部, 江苏 常州 213022)

摘 要: 本文提出了一种手写汉字文本的分割方法,填补了汉字识别领域在文本行分割方面的空白。本方法首先对预处理后的文本图片进行池化处理,然后运用并查集算法得到每行为一个连通区域,最后调整每行上下的孤立区域的归属,最终把多行文本图片分割为单行,为后期的汉字列分割做准备。此方法虽然用行分割,但也为汉字的列分割提供了新的思路。

关键词: 手写汉字识别; 池化; 文本分割; 并查集

Text segmentation method applied for handwritten Chinese characters recognition

LEI Xin, LI Junyang, SONG Yu, SAI Linwei

(Department of Mathematics and Physics, Hohai University Changzhou Campus, Changzhou Jiangsu 213022, China)

Abstract: In this paper, a text segmentation method for handwritten Chinese characters is developed, which fills in the blank of text line segmentation in the field of Chinese character recognition. In this method, the pretreatment of the text image is pooled. Then, using the method of searching and collecting, coarse segmentation of the pictures after the pooling step is processed, and the result of the coarse segmentation is segmented by the calculation of the connected set. Finally, the multi-line text picture is segmented into a single line to prepare for the later Chinese character segmentation. Although this method is used for row segmentation, it also provides a new way for column segmentation of Chinese character text.

Key words: handwritten Chinese character recognition; pool; text segmentation; the method of searching and collecting

引言

汉字识别技术经过长期的发展已经日趋成熟,无论是联机汉字识别还是难度更高的脱机手写汉字识别[1],其识别成功率均有较大的提升,并在相关的领域得到了一定的推广应用。作为汉字识别中的关键组成部分,汉字分割技术的进展也将制约着汉字识别率的研究提升。能够完整无误地分割出整个汉字对汉字识别来说尤为重要,这也是目前汉字识别技术攻关中的研究处理重点。

区别于字母、数字,汉字的结构复杂,形式多样,不同人的书写习惯和选用字体也形色多样,书写起来多具有很大的随意性,因此汉字分割较其它字符的分割也更显难度。近年来,学界已基于汉字识别技术研发提出了一系列的汉字分割方法。这些方法主要包括: 基于汉字结构的切分方法[2]、基于识别的切分方法[3]、基于词的整体切分方法以及基于统计的切分方法[4]等。如上的汉字分割方法虽然对特定的汉字图片获得了可观的成功率[5],但却都各自存在着一定缺陷。例如基于结构的切分方法中汉字笔画的提取十分复杂,基于识别和词整体切分方法又会产生效率与识别率双重走低的问题,而基于统计的切分方法则只适用于非黏连的汉字。综上论述可见,汉字分割依然是汉字识别技术中亟待解决的研究难题。

通常的汉字分割方法只是围绕少量汉字或者单行汉字而展开研究,对汉字文本分割的研究迄今为止仍属罕见。众所皆知,应用到汉字识别的地方多数为大篇幅的文章和段落,因此只有先对文本实现行分割后才能再次转入汉字分割,行分割是汉字分割的基础。不同于印刷体干净整齐的排布,手写汉字往往由于人为原因而扭曲变形,每一行汉字并不能做到严格水平布置,从而产生倾斜和扭曲,甚至出现相邻行的黏连现象,这都是不可避免的。如果用投影法或者连通域法[6]对文本进行分割则无法处理相隔较近的行,对严重扭曲和黏连的文本也难以达到满意的效果。

因此本文综合了多种算法后,设计提出了一种

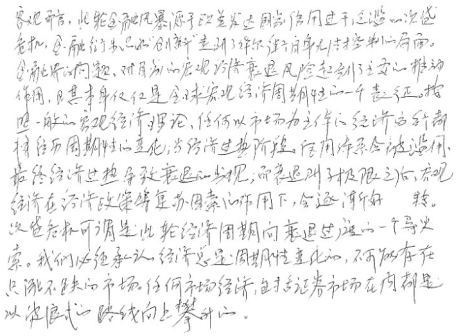
作者简介: 雷 鑫(1997-),男,本科生,主要研究方向: 机器视觉; 李俊阳(1996-),男,本科生,主要研究方向: 机器视觉; 宋 宇(1997-),男,本科生,主要研究方向: 机器视觉; 赛琳伟(1984-),男,博士,讲师,主要研究方向: 机器视觉、自然语言处理。

收稿日期: 2018-03-06

新的分割方法。不同于以往传统的汉字分割方法，该方法先用池化操作将汉字文本变换为大片的连通区域；再用并查集算法找出每个连通分支（往往是一行），这样就初步分离出每一行；而后确定相邻两行之间的小块连通区域的归属，对行引入微调处理，最终实现了文本的行分割。

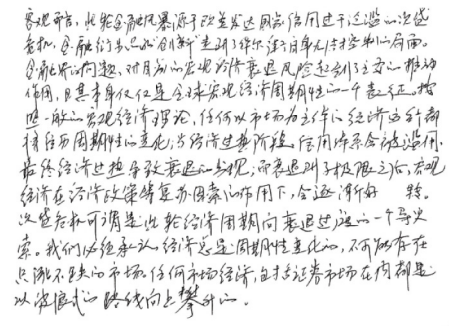
1　预处理

为了方便后期的文本分割和识别研究过程，即要对汉字图像进行必要的预处理。本研究中对初期录入的文本图片，如图 1（a）所示，进行了去噪声、二值化等操作，得到了图 1（b）。经过处理后发现，图片质量得到了明显的提升。



（a）初始图片

（a）Initial picture



（b）预处理后图片

（b）The picture after pretreatment

图 1　文本图片

Fig. 1　Text picture

2　池化处理

如果对图 1（b）直接进行分割显然难度较大，文本中每行汉字并不是严格的水平排列，且相邻两行的汉字存在黏连的情况。因此研究采取了池化的方法对预处理后的文本图片做更深层次的调整与加工。虽然一行可能有严重的倾斜，但一个小片段倾斜度往往很小，可以近似为水平。假设一个汉字的平均宽度是 64，就可选取 257* 3 大小的长条形卷积核对整个图片进行扫描，得出的结果如图 2 所示。从图 2 中可以清楚地看到，一行汉字经过池化后形成了一个连通的黑色区域，而相邻两行的黏连情况也被消除了，这样就可以根据这个结果对每一行进行提取，从而实现行分割的目的。

图 2　池化后的结果

Fig. 2　The result of pooling

3　粗分割

3.1　并查集算法

并查集算法[7] 可以高效地计算一个集合的连通分支。利用并查集中的合并算法，将图 2 中处于一个连通域的黑色像素点并为一个集合，这样就可以把不同的行区分出来。其中，每一个合并的集合就是一个列表，这个列表里包含了整个连通域的所有像素，而背景（本示例图片为白色）也合并为一个集合。这样就得到了许多不同的集合，这些集合包括每一行汉字、背景以及池化产生的噪点。这些噪点是由于汉字的某一部分在池化后脱离了本行而形成了一个单独的连通域，如果简单地删除会影响汉字的识别过程，因此在后续的操作中则需要将这部分噪点合并到其所在的行，从而将一行汉字完整地分割出来。

3.2　初步分割出行

并查集中的搜索算法可以查找某一个像素点所在的集合，根据这个特性可以从上向下对图 2 中的像素点进行扫描，从而找出图中所有的集合（除了背景集合），再将这些集合整理并从上到下顺次排序（包括噪点）。通过观察图 2 可以发现那些噪点无论是长度、还是宽度都远小于每一行，经过一定条件的筛选就可以暂时剔除这些噪点，这样研究就初步分割出了每一行（不包括噪点），并按照原有位置从上到下得到了有序排列，具体如图 3 所示，分割边界用线条显示。

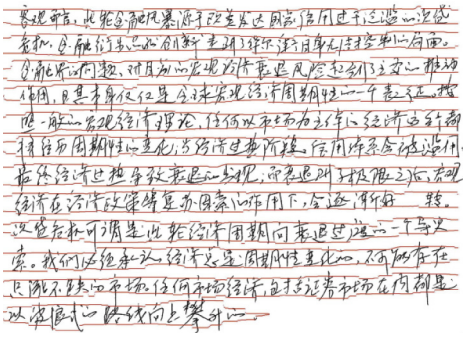


图 3 初步分割

Fig. 3 Preliminary segmentation

4 细分割

如图 3 所示，初步分割并不能将一行完整地分离出来，有些汉字上部或者下部被截断，而这些截断的汉字部分在池化时被抹去或者形成了单独的噪点，在细分割的时候需要将这些部分还原到其所在的行中。在图 3 的基础上对每一行边界向外(上边界向上，下边界向下) 计算连通集，如果存在连通的部分则相应地调整分割边界，从而将截断的部分包含于该行内。第一行的上边界和最后一行的下边界易于调整，其余的情况将大致分为 3 类，每一类的功能阐释可见如下。

(1) 修改上一行的下边界。从每一行(除去最后一行) 的下边界处向下进行连通集查找，直至下一行的上边界处结束，如果存在连通的部分则修改本行的下边界。若上下两行有黏连情况，即上下两行某部分连通在一起，则根据日常人们的书写习惯将这部分归于上一行。

(2) 修改下一行的上边界。从每一行(除去第一行) 的上边界向上进行连通集查找，直至上一行的下边界处结束，如果存在连通部分则修改本行的上边界。

(3) 确定两行中间区域的归属。当两行中间的噪点既不与上一行连接、也不与下一行连接时，通过确定阈值来判断这些噪点与上、下两行的位置关系，从而找出离噪点更近的行来设置合并。经过修改后的分割边界将每一行的汉字都完整地包含进去，得到的行分割结果如图 4 所示。例如第一个字“客”，该字下面的“口”在之前分割中被断开，经过这一步后，调整了下边界为口的下方。类似地，“客”字上面的一点之前也在上边界之外，现在重新修改了上边界为点的上边。这样，“客”字完全落在第一行的红色区域之内。

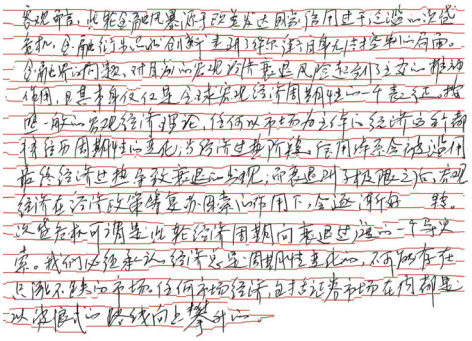


图 4 细分割结果

Fig. 4 Fine segmentation result

5 结束语

手写汉字文本相对于印刷体存在着扭曲、变形、黏连等现象。针对于这一状况，本文提出了一种文本分割的方法，采用并查集算法和计算连通集的方法对池化后的文本图片进行处理，从而实现行分割的目的，有效应对了文本扭曲和黏连的情况，对大量手写汉字样本进行实验均达到了理想的分割效果。该方法也可用于一行汉字的列分割。

参考文献

[1] 金连文. 手写体汉字识别的研究 [D] . 广州: 华南理工大学，1996.

[2] 熊鹏. 汉字笔迹的笔划提取 [D] . 武汉: 华中科技大学，2008.

[3] 邵洁，成瑜. 关于手写汉字切分方法的思考 [J] . 计算机技术与发展，2006，16(6) : 184-186，190.

[4] 赵继印，郑蕊蕊，吴宝春，等. 脱机手写体汉字识别综述 [J] . 电子学报，2010，38(2) : 405-415.

[5] 马瑞，杨静宁. 一种有效的手写汉字多步分割方法 [J] . 中国图象图形学报，2007，12(11) : 2062-2067.

[6] 陈艳，孙羽菲，张玉志. 基于连通域的汉字切分技术研究 [J] . 计算机应用研究，2005(6) : 246-248.

[7] 曾志雄. 并查集的树型存储表示及优化实现 [J] . 现代计算机，2001(7) : 61-63.

[8] 杨峰，张黎，王立克，等. 二值图像中基于连通集的滤波算法 [J] . 山东师范大学学报(自然科学版) ，2006，21(2) : 27-29.