基于折线切分路径的粘连搭接字符切分算法的研究

刘阳兴
(南开大学 微电子学研究所, 天津 300457)

摘 要: 针对粘连和搭接字符切分算法的不足,提出一种基于折线切分路径的字符切分算法。该算法利用投影法将粘连搭接字符与非粘连搭接字符分离开,而后结合粘连搭接字符独有的外形特征,通过引入惩罚权重的路径搜索算法快速而准确地得到粘连搭接字符间的折线切分路径;为了避免一些字符在以上的切分过程中被误切碎,利用识别反馈信息对一些字符子图像进行合并。实验结果表明,该算法对印刷体日英混排字符切分有很强的适应性,取得了较理想的切分效果。

关键词: 字符切分; 字符识别; 粘连搭接字符; 折线切分路径
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2011)10-3998-03
doi:10.3969/j.issn.1001-3695.2011.10.110

Non-linear partitioning path based approach for touching and kerned character segmentation

LIU Yang-xing
(*Institute of Microelectronics*, *Nankai University*, *Tianjin* 300457, *China*)

Abstract: Segmentation of touching and kerned characters has been the most difficult problem in character segmentation. This paper presented a novel approach based on exploiting non-linear partitioning paths to segment touching and kerned characters. Firstly, employed character projection to isolate touching and kerned characters with other characters. Then in order to find the correct non-linear segmentation path of touching and kerned characters, used a heuristic method seeking minimal-penalty curved cut to determine candidate paths from all possible segmentation paths to remove redundant paths and reduce the computational cost. Some characters might be segmented into several regions in above process. So evoked a merging procedure to combine some neighboring regions that belong to a single character. Experimental results demonstrate that our algorithm is robust in segmenting touching and kerned characters with respect to different orientation and language.

Key words: character segmentation; character recognition; touching and kerned characters; curve-based partitioning path

字符切分是指从图像文本中将每一个单字(包括英文字母、数字、标点等)分离出来。字符的识别和切分是密不可分的,只有把每个字符完整无误地切分出来,识别器才能正确地识别。所以字符切分是 OCR(optical character recognition)系统中极为重要的一环[1]。

统计表明,OCR 系统中的许多错误来源于字符的粘连搭接所引起的误切分[2,3]。字符的粘连搭接是指两个或两个以上的字符存在有粘连部分或两个独立的字符在空间上存在重叠的部分,如图 1 所示。由于文本图像质量(扫描样张不好或扫描分辨率低等情况)的不尽人意,使得粘连搭接字符大量存在于字符图像中。而字符的粘连或搭接都容易造成字符的误切分,把一个独立的字符切成几部分后或几个独立的字符因误切分而合在一起后送往识别器,必然造成识别结果的不正确[4],因此粘连搭接字符切分的正确性直接影响着 OCR 系统的整体性能。

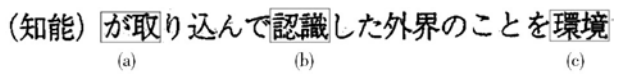

图 1 混有粘连搭接字符的字符行

要完成粘连搭接字符的正确切分,必须首先确定粘连搭接字符在文本图像中的准确位置[5]。但由于粘连搭接字符存在于文本字符集中,具有一定的隐蔽性,实用经典投影方法必然找不到正确的切分点[1];并且有时一些粘连搭接字符合在一起也能得到很好的识别结果从而造成字符的误切分。即使在准确得到粘连搭接字符的准确位置后,由于粘连搭接字符存在有公共点或空间投影存在重叠,并不存在将粘连搭接字符正确分割开的直线切分路径。这些都给粘连搭接字符的切分带来了困难。

本文针对粘连搭接字符的切分难点,提出了一种基于折线切分路径的切分算法。为了确保粘连搭接字符检测的准确性,该方法先利用简单的切分方法分割字符图像,对识别有问题的图像块再进行分析,判断其是否属于粘连搭接字符,并通过搜索粘连搭接字符间的最优折线切分路径将字符分割开。

与其他方法相比,本方法采用投影法快速分离粘连搭接字符与非粘连搭接字符,大大提高了切分速度;在切分粘连搭接字符时,采用了引入惩罚权重的路径搜索算法快速搜索折线切分路径,使粘连搭接字符的切分更加准确有效。同时在字符切分过程中充分利用识别反馈信息,进一步提高了切分的精确度。

1 已有的字符切分算法

1.1 投影法

字符投影是应用到字符切分的一种常用方法,它计算简单

收稿日期: 2011-02-23; **修回日期**: 2011-03-14
作者简介:刘阳兴(1978-),江西泰和人,副教授,博士,主要研究方向为图像处理、模式识别、VLSI-CAD(liuyangxing@nankai.edu.cn).

并且切分速度快。垂直投影是通过计算水平方向每一列上所有黑点数的总和而得到。在单行字符的垂直投影中,如果字符是完全分开的,字间的投影将为零值[6],如图 2 所示。通过计算字符行(列)的投影值,可以把许多质量较好的字符快速切分开。

图 2　字符行投影

1.2　曲线切分路径法

Wang 和 Jean[7] 提出了寻找粘连字符曲线切分路径的方法。根据粘连搭接字符的外形特征,如果始终沿着一个方向搜索切分路径会把一个完整的字符误切成几部分,因此要完整无误地把粘连搭接字符切分开,需找出曲线切分路径。该方法对粘连搭接字符的切分非常有效,但如果应用到整个字符行中时,则需花费大量的字符切分时间。

1.3　识别反馈法

基于识别的方法是利用识别反馈对字符切分结果进行判别。该方法利用识别器对待识别图像识别后,识别器会返回一个衡量识别图像是否为一个独立字符的可信度值。根据这个返回值,可以判断送往识别器识别的图像是否为一个独立的字符。只有当返回值大于一定的数值(也就是阈值,记为 T)时,才认为识别结果是正确的,从而判断出送往识别器识别的图像为一个独立字符。字符切分与识别的紧密结合可以提高字符切分算法的鲁棒性[8],但以识别反馈为基础的切分比较耗时,实际应用较少[9]。

2　字符分割算法

在印刷体字符样张中存在许多质量较好的字符,通过普通投影法就可以将它们正确切分开。对于粘连搭接字符,投影方法则难以找到正确的切分点[10],并且可能造成一些字符的误切分。但是如果把样张中所有字符当成粘连搭接字符进行切分,必然造成时间上的大量浪费;如果不考虑样张中粘连搭接字符的存在,又将使粘连搭接字符得不到正确切分,从而严重影响切分的正确率。

为了保证字符切分的速度和正确率,本文提出的方法主要分为三步:a)利用字符投影快速把粘连搭接字符和非粘连搭接字符分离开;b)利用含惩罚权重的评价函数来搜索优化粘连搭接字符的折线切分路径,准确地把粘连搭接字符切分开;c)利用识别反馈法将一些误切分字符进行合并。图 3 给出了本算法的流程。

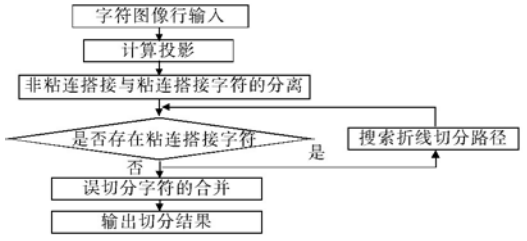

图 3　算法流程

2.1　非粘连搭接字符与粘连搭接字符的分离

通过计算字符行的投影,可以把字符行分成一段段的字符图像。为了判断出哪些字符图像段为粘连搭接字符段,分别对字符段的宽度和识别结果进行如下的判断:

a)设每一段字符图像的宽度为 $w(i)$,字符行一个独立字符的平均宽度为 w_a。当 $w(i)$ 比 w_a 大出许多时,则可以判定该字符段为粘连搭接字符段。实验结果表明,$1.5\times w_a$ 是较好的判定界限,即当 $w(i)\geqslant 1.5\times w_a$ 时,可以判定该字符段为粘连搭接字符段。

b)识别所有宽度小于 $1.5\times w_a$ 的字符段后,可以判定那些识别结果不正确的字符段也为粘连搭接字符段。

以图 4 给出的字符行为例,通过字符的宽(高)度信息和识别器的返回值,可以判断图 5 中(a)~(c)包含的并不是一个独立的字符,(b)(c)粘连搭接字符显然没有被正确切分开。由于字符"が"获取的粘连,造成投影后"が"字符与(a)的误切分。这就要求在第二步的切分过程中,不仅要把粘连搭接字符正确切分开,还要对切分结果中不正确的部分进行合并。

图 4　含有粘连搭接字符的字符行及其投影

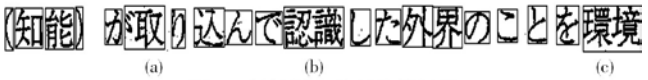

图 5　经过第一步切分的结果

2.2　粘连搭接字符曲线切分路径的启发式搜索

在寻找切分路径过程中,每一处粘连搭接字符的备选切分路径数目巨大,而每一条路径都必须计算出相应的消耗值,这使得计算量非常大。

为了减少计算量而提高切分速度,在切分路径搜索过程中提出了一种启发式搜索方法。该方法是在路径搜索过程中对每一个搜索的位置进行评价,得到最好的位置,再从这个位置进行搜索直到目标,从而达到省略大量无谓的搜索路径、提高切分效率的目的。在启发式搜索中,对位置的估价非常重要,关键是选择合适的评价函数。如果所用的评价函数掌握得过严而对确实有希望的某些路径也拒绝的话,则结果可能导致切分错误;另一方面如果所用的评价函数掌握得过宽,而对许多错误的路径都接受的话,结果将会出现计算量的大大增加。通过对粘连搭接字符切分路径特点的观察,发现选用对切分路径穿越黑点的次数以及切分路径的起始点和终止点在字符分布方向上的距离进行判断,计算简单并且可以快速排除大量的错误切分路径。

设 (x_1,y_1) 和 (x_2,y_2) 分别为切分路径的起始点和终止点,W_a 为样张中字符的平均宽度,W_s 为样张中字符的平均笔画宽度,$B_{\text{num}}(i)$ 为第 i 条路经穿越的黑点数。

$$C_1(x,y) = W_s - B_{\text{num}}(i)$$

$$C_2(x,y) = \gamma\times W_a - |x_2 - x_1|$$

其中 γ 为一常数。

$$f(i) = C_1(x,y)\ C_2(x,y)$$

$f(i)$ 即为所设的评价函数,$C_1(x,y)$ 对切分路径穿越过的黑点数进行了评价,$C_2(x,y)$ 对切分路径的起始点和终止点在字符分布方向上的距离进行了评价。通过计算 $f(i)$ 值,当 $f(i)<0$ 时,可知切分路径穿越的黑点数超过了字符笔画的宽度或是切分路径的起始点和终止点的距离大于一个独立字符的宽度,从而可以判断出这条路径是错误的,无须继续计算这条路径的消耗值。在计算出所有 $f(i)>0$ 对应路径的消耗值后,也就可以得到消耗值最小的切分路径。

利用上面提出的切分路径的启发式搜索方法,即可很快找出图 4 中粘连搭接字符的正确切分路径,如图 6 所示。

图 6　图 4 中粘连搭接字符的折线切分路径

2.3　误切分字符的合并

经过以上的切分过程,粘连搭接字符被切分成不存在粘连搭接的几部分。把这几部分的识别结果与阈值 T 相比较,可以判断出有些部分并不是一个独立的字符,它们只是一个独立字符的一部分。通过以下的判别规则,可以把一个独立字符的几部分正确合并。

首先把识别不正确的一部分与相邻的一部分合并后进行识别,设识别返回值为 v_r;然后对每一部分进行识别,识别后的返回值分别是 v_1、v_2。

1)定义 $v_{emp}=\alpha\times v_1+\beta\times v_2$。其中 $\alpha+\beta=1$,α、β 为根据字符质量预先设定好的常数。

2)当 $v_r>\max(v_{emp},T)$ 时可知,两部分合并后成为一个独立字符,因此应该将两部分合并;当 $v_r\leqslant\max(v_{emp},T)$ 时可知,两部分合并后并不是一个独立字符,因此不合并两部分。

图 7 即为经过第三步切分后的切分结果,粘连搭接字符都已被正确切分开。

图 7　最终的切分结果

3　实验结果

笔者扫描了含大量日英横竖混排字符的 213 页日文报纸及杂志作为测试样本。图 8 ~ 10 形象地描述了本文提出的算法对粘连搭接字符的提取,进而对粘连搭接字符切分的全部过程。

图 8　原始图像(含粘连搭接字符的字符行)

图 9　粘连搭接字符的提取

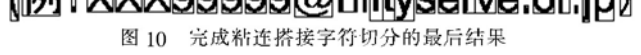

图 10　完成粘连搭接字符切分的最后结果

即使在测试样本中有多个(两个以上)字符连续粘连或搭接在一起时(图 11 方框中的字符),仍可以通过多次递归使用该算法将每个独立的字符切分出来,图 12 为该算法的切分结果。实验结果进一步表明该算法具有很强的自适应性。

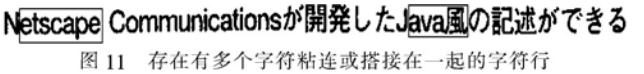

图 11　存在有多个字符粘连或搭接在一起的字符行

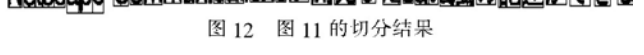

图 12　图 11 的切分结果

图 13 和 14 给出了更完整的字符图像切分结果。213 页测试样本的切分统计结果见表 1。从样本的测试结果可以看出,该方法的切分正确率可达 97.5% 以上,为粘连搭接字符的切分提供了有价值的解决方案。

表 1　字符切分结果统计

比较项	杂志	报纸
样本数/页	135	78
切分正确率/%	98.1	97.6

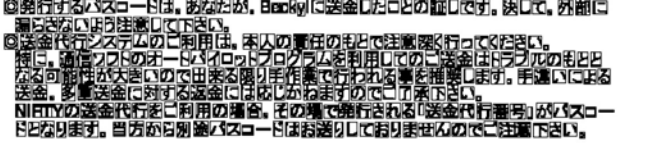

图 13　横向排版字符图像切分结果

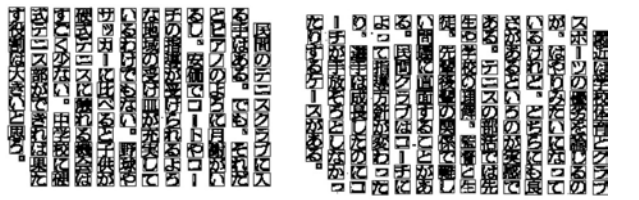

图 14　竖向排版字符图像切分结果

4　结束语

较高的字符切分正确率是保证 OCR 系统高性能的前提条件。针对字符切分中的难点问题之一,本文提出了一种基于折线切分路径的粘连搭接字符切分算法。由于粘连搭接字符存在有公共点或空间投影存在重叠,独立字符易被误切分为几部分或独立字符的部分像素点与其他字符易被误合并在一起。因此,通过引入惩罚权重的评价函数对切分路径进行搜索优化,得到粘连搭结字符间的最优折线切分路径,从而将粘连搭接字符切分开。同时在切分过程中,借助于识别反馈信息,避免了将一些呈左右(ハリル)或上下(ニラテ)分布的日文字符切碎,从而进一步提高了切分正确率。

参考文献:

[1] NAGY G. Twenty years of document image analysis in PAMI[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000,22(1):38-62.

[2] 李佐,王妹华,蔡士杰. 一种基于前端预测识别的粘连字符分割方法[J]. 计算机研究与发展,2001,38(11):1337-1344.

[3] LU Yi. Machine printed character segmentation:an overview[J]. Pattern Recognition,1995,28(1):67-80.

[4] 陈臻刚,丁晓青,刘长松,等. 文档识别中误切分字符拒识问题的研究[J]. 计算机工程与应用,2002,38(17):69-72.

[5] NOMURA A,MICHISHITA K,UCHIDA S,*et al*. Detection and segmentation of touching characters in mathematical expressions[C]// Proc of the 7th International Conference on Document Analysis and Recognition. Washington DC: IEEE Computer Society, 2003: 126-130.

[6] LU Yi,HAIST B,HARMON L,*et al*. An accurate and efficient system for segmenting machine-printed text[C]//Proc of the 5th Advanced Technology Conference. Washington DC:IEEE Press,1992:93-105.

[7] WANG J,JEAN J. Segmentation of merged characters by neural networks and shortest path[J]. Pattern Recognition,1994,27(5): 649-658.

[8] TSUJIMOTO S,ASADA H. Resolving ambiguity in segmenting touching characters[C]//Proc of the 1st International Conference on Document Ananlysis and Recognition. 1991:701-709.

[9] 张振绘,刘赛. 女书文字切分算法的设计与实现[J]. 中国科技信息,2010(12):119-120.

[10] 吕岳,施鹏飞,张克华. 基于汉字结构特征的自由格式手写体汉字切分[J]. 电子学报,2000,28(5):102-104.