基于滴水算法的印刷体维吾尔文切分方法

朱 兰,袁保社,余 伟
(新疆大学 信息科学与工程学院,新疆 乌鲁木齐 830046)

摘 要:为进一步提高印刷体维吾尔文字符的切分准确性,尤其针对每一行内的两个连体段之间(包括两个独写字母之间以及连体段与独写字母之间的情况)存在重叠现象时的切分,现有的方法未能有效地处理该情况,所以针对此问题提出了一种基于改进的滴水算法的切分方法。该方法首先判断两个连体段之间的关系,若存在空白间隙,则选择空白间隙的左右端作为切分点;若存在重叠现象,则选择基线空白间隙的中点作为切分点,然后根据滴落规则对两个连体段进行切分。实验结果表明,该方法能有效提高切分准确率。

关键词:印刷体维吾尔文;切分;连体段;重叠;基线;滴水算法
中图分类号:TP391.43 **文献标识码**:A **文章编号**:1673-629X(2015)07-0107-04
doi:10.3969/j.issn.1673-629X.2015.07.023

Segmentation Method of Printed Uyghur Based on Drop Fall Algorithm

ZHU Lan, YUAN Bao-she, YU Wei
(College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

Abstract: In order to further enhance the segmentation accuracy of printed Uyghur characters, especially for the segmentation where there is an overlap between two conjoined sections (including the situation between two independent characters as well as between the conjoined section and the independent character) within each line, the existing methods have not dealt with the situation effectively. Therefore, a segmentation method based on improved drop fall algorithm for this problem is proposed. The method first determines the relationship between the two conjoined sections. If there is a gap, then select the left and right ends of the gap as the segmentation points. If there is an overlap, then select the midpoint of the baseline of the gap as the segmentation point, and then segment the two conjoined sections according to the dripping rules. The experimental results show that this method can effectively improve the segmentation accuracy.

Key words: printed Uyghur; segmentation; conjoined section; overlap; baseline; drop fall algorithm

0 引 言

随着新疆信息化建设的飞速发展,少数民族语言文字信息技术也成为了其不可或缺的一个方面,其中以维吾尔文为典型代表,占据了很大比例。基于印刷体及屏幕图片上的维吾尔文字识别一直是新疆少数民族语言文字识别的一个重要研究方向。新疆大学哈力木拉提教授与清华大学合作在印刷体维吾尔文字符识别技术上取得了重要成果[1],其后的研究主要在进一步提高识别率上。根据工信部 2009 年度电子信息产业发展基金项目中印刷体维吾尔文字符识别系统的研制需求,本课题组开展了相关技术的研究开发工作,并取得了一些初步成果[2-4]。

从已经开发的印刷体维吾尔文识别软件测试效果来看,影响识别率的关键问题就是字符切分[5]。目前对印刷体维吾尔文字符的切分使用较多的是基于投影的切分方法[6]。该方法首先采用水平投影法对整篇文档图片进行行切分,然后再对每一行文字做垂直投影切分,根据投影中空白间隙阈值[7]的大小先对阈值较大的部分进行切分(即词间切分),从而将单词切分出来,再用同样的方法对阈值较小的部分进行切分,也就是将单词内的连体段和独写字母切分出来,最后再对连体段进行细切分,把粘连的字母切分出来。

通过对已有维吾尔文切分方法的研究发现,目前主要是针对连体段内字母的切分提出了很多新方法,而对于每一行内的两个连体段之间(包括两个独写字母之间以及连体段与独写字母之间的情况)的切分,

收稿日期:2014-08-28 **修回日期**:2014-11-28 **网络出版时间**:2015-06-23
基金项目:工信部 2009 年度电子信息产业发展基金项目(工信部财[2009]453)
作者简介:朱 兰(1988-),女,硕士研究生,研究方向为中文信息处理;袁保社,教授,研究方向为中文信息处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150623.1031.025.html

目前仍普遍采用基于垂直投影的切分方法。该方法忽略了当行内两个连体段之间存在重叠现象(在基线处实际未相连,但垂直投影相连)时的切分问题,这将影响整个切分过程的准确性。鉴于此情况,文中在采用传统水平投影法进行行切分之后,采用改进的滴水算法对行内两个连体段进行切分。

1 滴水算法

滴水算法首先由 G Congedo[8] 于 1995 年提出,其基本思想是模拟了水滴因受重力作用而从高处向下滴落,在这个过程中水滴会沿着字符轮廓向下滴落或在水平位置左右滚动,当水滴陷入字符轮廓凹陷处时则直接竖直向下穿透笔划并继续滴落,最后根据水滴的滚动轨迹来确定切分路径。该方法由于其执行效果较好,目前主要运用在粘连字符的切分上,如:手写数字串、印刷体公式及少数民族文字等方面都取得了一定成果[9-11]。

滴水算法的决定因素主要有切分起始点的选择、水滴滴落规则以及方向的确定。根据这些因素的不同,可将滴水算法分为四种不同的分割方法,即左侧向下、右侧向下、左侧向上和右侧向上[12]。

采用左侧向下的分割方法对粘连字符进行切分,其过程为:对一幅图像自上而下、由左到右进行扫描,找到第一个满足像素扫描条件为(…1 * 0…01…)的白像素点(*),将其作为切分起始点(即切分起始点就是位于左右两侧黑像素点之间的第一个白像素点),该点的下一移动位置根据其下方、右下、左下、右方、左方的白像素点情况决定,且按照正下、右下、左下、向右、向左的次序来选择下一切分点的位置[13]。

特别地,当水滴位于字符轮廓凹陷处时会左右来回滚动,此时水滴会在笔划内的某一极值点处直接进行垂直渗漏。

滴水算法执行效果固然很好,但同时也存在缺点:首先切分起始位置的选择很关键,复杂度很高,起始点的位置直接影响到滴水算法的执行结果,若不能准确合理地选择起始位置,则可能导致切分错误;其次,垂直渗漏过程会对字符的笔画造成损害,使笔画不完整,即切分出来的字符或不完整或出现多余笔画。所以文中考虑借用滴水算法的思想,将其运用在维吾尔文不粘连的字符切分上,即将这一算法运用在行内两个连体段之间的切分上,这就需要针对维吾尔文的书写特点,在传统滴水算法的基础上做出改进。

2 滴水算法的改进及其应用

文中基于传统滴水算法,分别对切分起始点的选择、滴落规则以及方向做出了改进,并将其运用在行内两个连体段之间的切分上。该方法首先结合每列基线域像素值和该列整个像素值的情况找到切分点,然后以这些点作为滴水算法的起始切分点,按照改进的滴落规则分区域进行滴落并形成最终需要的切分路径。

2.1 切分起始位置的选择

维吾尔文的特点是沿着基线书写,每一行中的两个连体段之间在基线处都有较为明显的空白间隙,所以切分起始位置就选在基线的空白间隙上。具体选择过程如下:

设第 j 列的像素值为 $V(j)$,第 j 列基线域的像素值为 $VB(j)$($V(j)$ 和 $VB(j)$ 的具体定义将在后文中给出),对一幅输入图像从左往右逐列进行扫描,则:

(1)当有连续 c($c \geqslant 2$)列的 $V(j)=0$ 时,取这 c 列中的第一列和最后一列所对应的基线位置作为切分起始位置;

(2)当有连续 c($c \geqslant 2$)列的 $(VB(j)=0) \cap (V(j) \neq 0)$ 时,取这 c 列所对应的基线的中点作为切分起始位置;

(3)当有且仅有一列的 $VB(j)=0$ 时,该列所对应的基线位置即为切分起始位置。

2.2 滴落规则

由于文中将滴水算法用在不粘连字符的切分上,所以最根本的原则是不能破坏字母原有的各个笔画。鉴于此,文中在传统滴落规则的基础上进行了如下改进(见图 1)。

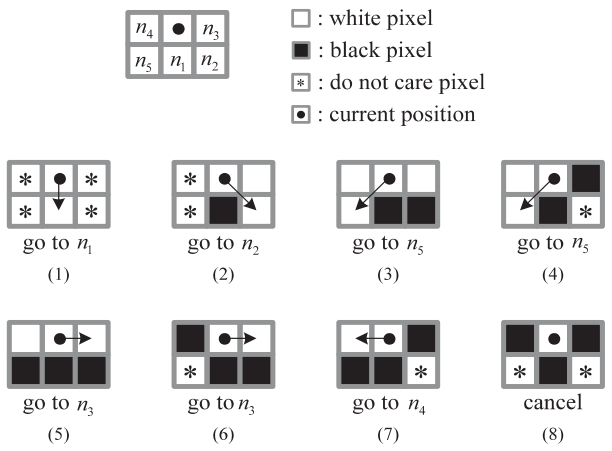


图 1 改进的滴落规则

按 $n_1 \rightarrow n_2 \rightarrow n_3 \rightarrow n_4 \rightarrow n_5$ 的顺序对图 1 规则进行 0-1 编码,若这些位置为白像素则为 0,黑像素则为 1,得到 0-1 编码 ζ 如下(其中,* 代表任意位任意数,ζ 代表当前点其周围五个位置的 0-1 情况):

(1)If ζ match "(0)(.*)",则遵循图 1(1)规则,水滴向下;

(2)If ζ match "(100)(.*)",则遵循图 1(2)规则,水滴向右下;

(3)If ζ match "11000",则遵循图 1(3)规则,水滴

向左下;

(4)If ζ match "(1)(.*)(100)",则遵循图 1(4)规则,水滴向左下;

(5)If ζ match "11001",则遵循图 1(5)规则,水滴向右;

(6)If ζ match "(1101)(.*)",则遵循图 1(6)规则,水滴向右;

(7)If ζ match "(1)(.*)(101)",则遵循图 1(7)规则,水滴向左;

(8)If ζ match "(1)(.*)(11)(.*)",则遵循图 1(8)规则,水滴终止,并将整条切分路径还原。

情况一:当行内两个连体段之间存在空白间隙时,当前水滴的下一滴落位置只需遵循图 1(1)这一条规则进行,即直接沿着当前位置的正下方 n_1 位置滴落。

情况二:当行内两个连体段之间存在重叠现象时,当前水滴的下一滴落位置需要根据图 1 中这八条规则进行判断,即由它下方的三个像素点和它左右两个像素点共五个像素点的情况共同决定[14]。

特别地,当切分起始点刚好在一些带有“弯钩”笔画的特殊字母上时,如:ﻭ ﻯ ﺱ ﻝ等,即这些字母的基线处为白像素点,但基线下方或上方为黑像素点时,那么当水滴按照规则滴落后会位于字符轮廓的凹陷处,即如果遇到图 1 中 6→7 或 6→5…5→7 或 8 这三种情形时,水滴会在左右两个方向进行来回滚动或停留在某一局部极小点处,此时采取的策略是将此水滴的滚落轨迹全部删除,即表明在该位置不进行切分。

2.3　分区域滴落

从前面已选定的切分起始点开始,按照上述的改进滴落规则,先对基线下方区域采用向下的滴落方式,再对基线上方区域采用向上的滴落方式(即将字符图像作垂直镜像后向下滴落),然后将这两部分合并起来形成完整的水滴滚动轨迹,也就得到了最终所需的切分路径。当两个连体段之间仅有唯一一列空白间隙时,切分路径就是这一空白列;当两个连体段之间的空白间隙大于一列时,空白间隙的最左列即为第一个连体段的右边界,空白间隙的最右列即为第二个连体段的左边界;当两个连体段之间存在重叠情况时,切分路径是位于两连体段中间的一条不规则曲线。

2.4　具体实现步骤

Step1:采用传统的水平投影法进行行切分,在切分过程中记录下每一行文字的上下边界,分别记为 a 和 b,并计算出行高 $h = b - a$。

Step2:设第 j 列的像素值 $V(j) = \sum_{i=a+1}^{b} p(i,j)$(其中 $p(i,j)$ 为行切分过程中已定义过的图像中第 i 行、第 j 列的像素值,$p(i,j)=0$ 表示白像素,$p(i,j)=1$ 表示黑像素),由于文字行中连接字母的基线部分的高度相同,因此从左往右逐列对文字行进行扫描,记录文字行中每一列的像素值 $V(j)$,统计出具有相同 $V(j)$ 的列的数目,具有最多相同 $V(j)$ 的列所对应的 $V(j)$ 值就是基线的高度,记为 h_b。

Step3:计算出基线高度 h_b 后根据公式(1)计算得到基线位置[15](其中 B_{top} 和 B_{btm} 分别表示基线的上边界和下边界),那么文字行的基线域就是高度为 h_b 的水平投影值最大的带状区域。

$$\begin{cases} B_{top} = \arg_i \max\left(\sum_{k=0}^{h_b-1} H_{i+k}\right), i = 0,1,\ \cdots, h - h_b \\ B_{btm} = B_{top} + h_b - 1 \end{cases} \tag{1}$$

Step4:设第 j 列基线域的像素值 $VB(j) = \sum_{i=B_{top}}^{B_{btm}} p(i,j)$,从左往右对文字行中的每一列进行扫描从而确定各切分起始点。

(1) $VB(j)$ 从 $VB(x_1)$ 到 $VB(x_2)$ 连续为 0 且 $V(j)$ 从 $V(x_3)$ 到 $V(x_4)$ 也连续为 0:

①当 $x_1 \leqslant x_3 < x_4 \leqslant x_2$ 时,则切分起始点为 $(x_3, B_{top} + \lfloor \frac{h_b}{2} \rfloor)$ 和 $(x_4, B_{top} + \lfloor \frac{h_b}{2} \rfloor)$;

②当 $x_1 \leqslant x_3 = x_4 \leqslant x_2$ 时,则切分起始点为 $(x_3, B_{top} + \lfloor \frac{h_b}{2} \rfloor)$。

(2) $VB(j)$ 从 $VB(x_1)$ 到 $VB(x_2)$ 连续为 0 且 $V(j)$ 从 $V(x_1)$ 到 $V(x_2)$ 连续不为 0(其中 $x_1 \leqslant x_2$)时,则切分起始点为 $(x_1 + \lfloor \frac{x_2 - x_1}{2} \rfloor, B_{top} + \lfloor \frac{h_b}{2} \rfloor)$。

Step5:从选定好的切分起始点开始,遵循改进的滴落规则,先对基线下方区域进行滴落,再将图像作垂直镜像后对另一半区域进行滴落,这两部分的水滴滚动轨迹合并起来就是所需的完整的切分路径。

2.5　与传统方法的比较

文中从切分准确性、对重叠现象的适应性和切分速度这三个方面对垂直投影法和改进的滴水算法的性能进行了如下对比:

(1)垂直投影法:静态的垂直投影空白间隙无法体现重叠情况,导致切分准确性较低;直接根据投影空白间隙进行垂直切分,其切分速度较快但对重叠现象的适应性较差。

(2)改进的滴水算法:可以动态地查找切分路径,所以对重叠现象的适应性较好,因而提高了切分准确性;但该方法在具体执行时要判断当前水滴周围五个像素的情况后选择下一滴落位置,所以计算量相对较大,切分速度较慢。

3 实验结果与分析

首先扫描得到一幅原始图像后对其进行去噪、二值化、倾斜校正等预处理,然后采用水平投影法进行行切分,得到如图 2 所示的图片。

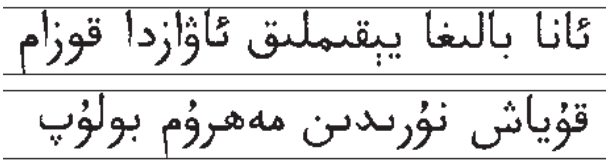


图 2 行切分效果图

实验 1:在行切分的基础上采用传统的垂直投影法(根据垂直投影空白间隙切分)对每一行内的两个连体段之间进行切分。首先对图 2 中的每一行分别进行垂直投影,得到相对应的投影图,然后取投影的每一空白间隙的左右两端作为切分位置进行切分,得到如图 3 所示的切分效果图。

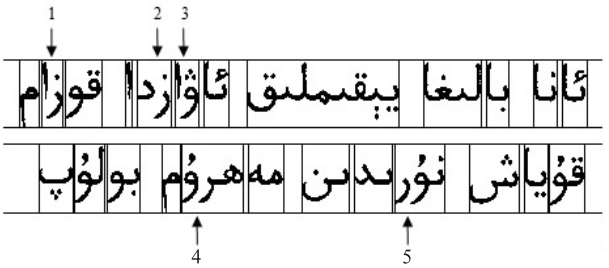


图 3 基于垂直投影法的切分效果图

实验 2:在行切分的基础上采用改进的滴水算法对每一行内的两个连体段之间进行切分,得到如图 4 所示的切分效果图。

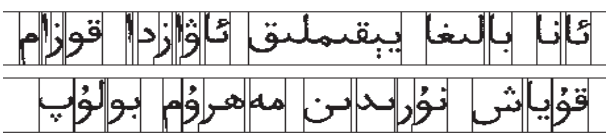


图 4 基于改进的滴水算法的切分效果图

通过对图 3 和图 4 的比较可以发现:图 3 中的 1,2,3,4,5 这五个位置都存在字符重叠现象,利用垂直投影法则容易出现漏切分现象,这必然会影响到后面连体段内字母的切分,从而影响整个切分过程的准确性;而文中提出的改进的滴水算法则可以很好地弥补垂直投影法带来的漏切分情况,从而提高切分正确率。

实验 3:选取 10 幅扫描得到的短文图像进行切分正确率的测试,包含独写字母、连体段和标点共计 1 793个,若将它们全部正确切分开,则应该切分的数目为 1 729。本实验对垂直投影法和改进的滴水算法的切分正确率进行了比较,实验数据如表 1 所示。可见,针对连体段之间的切分,采用传统的垂直投影法的切分正确率低于提出的改进的滴水算法。

由表 1 可知,文中提出的方法也出现了漏切情况,主要有以下几个原因:

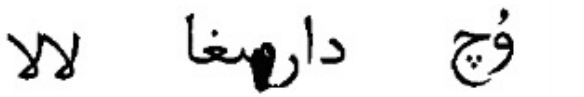
表 1 切分正确率对比

方法	应切数	实切数	漏切数	切分正确率/%
垂直投影法	1 729	1 491	238	86.23
改进的滴水算法	1 729	1 686	43	97.51

(1)二值化过程中选择不同阈值处理得到的图像有差异,原本的独写字母粘连在一起形成了连体段,如图 5(a)所示;

(2)图像受到各类噪声的干扰,甚至有的字符信息被人为损坏,如图 5(b)所示;

(3)当水滴滚落到凹陷处时按照规则删除水滴滚动轨迹,不进行切分,但当出现图 5(c)情况时则导致漏切。



(a)独写字母粘连 (b)字符信息损坏 (c)重叠但规则不适用

图 5 易出现漏切的情况

4 结束语

文中在借用传统滴水算法思想的基础上对其做了改进,并将其运用在行内两个连体段之间字符的切分,当行内两个连体段之间存在重叠现象时,该方法能有效地分割且执行效果良好,具有较好的适用性,也为后面连体段内字母的细切分提供了更准确的基础。同时该方法也存在一定不足,如:对图像质量要求较高、滴落规则对一些特殊情况不适用、判断水滴下一位置过程中的计算量和开销都较大等,在今后的研究中将针对这些问题做出改进。

参考文献:

[1] 哈力木拉提,丁晓青. 多字体印刷维吾尔文的切分[J]. 中文信息学报,1997,11(3):35-40.

[2] 李 晓,袁保社,陈 卿,等. 基于像素积分投影的印刷体维文字母切分方法[J]. 计算机技术与发展,2012,22(4):41-44.

[3] 陈 卿,袁保社,李 晓,等. 基于模板匹配的印刷维吾尔文字符识别研究[J]. 计算机技术与发展,2012,22(4):119-122.

[4] 万金娥. 印刷体维吾尔文字识别系统关键技术研究与实现[D]. 新疆:新疆大学,2013.

[5] Sankur B, Sezgin M. Image thresholding techniques: a survey over categories[J]. Pattern Recognition, 2001, 34(2): 1573-1583.

[6] Fujisawa H, Nakano Y, Kurino K. Segmentation methods for character recognition: from segmentation to document structure

(下转第 115 页)

检测出 DDoS 攻击。再对流量中存在 DDoS 攻击流量和突发流量的 50 段流量进行检测,检测情况如表 3 所示。

表 2　加入 DDoS 流量攻击检测性能表

检测方法	检测异常	实际攻击	正确检测	错误检测	检测率/%	误报率/%
传统匹配方法	21	20	17	4	85	19
文中方法	20	20	18	2	90	10

表 3　加入突发流量和 DDoS 攻击性能对比表

检测方法	检测异常	实际攻击	正确检测	错误检测	检测率/%	误报率/%
传统匹配方法	28	20	17	11	85	39.3
文中方法	21	20	18	3	88	11.3

由表 3 可以看出,当网络中存在突发流量和 DDoS 攻击时,相对于传统的匹配方法,文中提出的复合式 DDoS 攻击检测方法能更好、更有效地检测出 DDoS 攻击,误报率更低。

5　结束语

文中提出的复合式 DDoS 攻击检测方法是对网络流量数据进行离线操作分析,如果把该方法应用到实际的网络环境中,在检测速率、检测效率方面有待进一步实验测试和提高;另外,还可把基于重尾特性的检测方法应用到网络流量的其他属性,例如包的持续时间、流速、包的大小等,对 DDoS 攻击进行更加有效的检测。等技术成熟还可应用到对其他网络攻击的检测中去,这将有更大的实际意义。

参考文献:

[1] 徐　图,何大可. 网络流单边连接密度的时间序列分析[J]. 四川大学学报:工程科学版,2007,39(3):136-140.

[2] 程　光,龚　俭,丁　伟. 基于抽样测量的高速网络实时异常检测模型[J]. 软件学报,2003,14(3):594-599.

[3] 程　光,龚　俭,丁　伟. 网络测量及行为学研究综述[J]. 计算机工程与应用,2004,40(27):1-8.

[4] 许晓东,杨海亮,朱士瑞. 基于重尾特性的 SYN 洪流检测方法[J]. 计算机工程,2008,34(22):179-181.

[5] 巩永旺. 基于扫描流量统计的本地网蠕虫检测方法[J]. 计算机技术与发展,2011,21(7):145-148.

[6] 任勋益,王汝传,王海艳. 基于自相似检测 DDoS 攻击的小波分析方法[J]. 通信学报,2006,27(5):6-11.

[7] 孙知信,李清东. 基于源目的 IP 地址对数据库的防范 DDoS 攻击策略[J]. 软件学报,2007,18(10):2613-2623.

[8] 孙知信,唐益慰,程　媛. 基于改进 CUSUM 算法的路由器异常流量检测[J]. 软件学报,2005,16(12):2117-2123.

[9] 母军臣,甘志华,许宏云. 基于动态包过滤的 RoQ 攻击防御策略[J]. 电脑知识与技术,2007,1(6):1532-1533.

[10] 谢　逸,余顺争. 新网络环境下应用层 DDoS 攻击的剖析与防御[J]. 电信科学,2007,23(1):89-93.

[11] Yang G,Gerla M,Sanadidi M Y. Defense against low rate TCP attacks:dynamic detection and protection[C]//Proceedings of network04. New York,NY,USA:ACM,2004:189-198.

[12] Shevetkar A,Anantharam K,Ansari N. Low rate TCP denial-of-service attack detection at edge routers[J]. IEEE Communications Letters,2005,9(4):363-365.

[13] Guirguis M,Bestavros A,Matta I. Exploiting the transients of adaptation for RoQ attacks on internet resources[C]//Proc of the 12th IEEE international conference on network protocols. [s. l.]:IEEE,2004:184-195.

[14] Leland W E,Taqqu M S. On the self-similar nature of Ethernet traffic[J]. IEEE/ACM Trans on Networking,1994,2(1):1-15.

(上接第 110 页)

analysis[J]. Proceedings of the IEEE,1992,80(7):1079-1092.

[7] Gasser A, Hazem R. An automatic text reader using neural networks[C]//Proceedings of the Canadian conference on electrical and computer engineering. [s. l.]:[s. n.],1993:92-95.

[8] Congedo G,Dimauro G,Impedovo S,et al. Segmentation of numeric strings[C]//Proceedings of the third international conference on document analysis and recognition. [s. l.]:[s. n.],1995:1038-1041.

[9] 张　闯,蔺志青,肖　波,等. 适用于银行票据手写数字串切分的滴水算法[J]. 北京邮电大学学报,2006,29(1):13-16.

[10] 李小园,田　刚,封　超. 印刷公式中粘连字符的切分[J]. 科学技术与工程,2011,11(3):628-632.

[11] 刘　赛,王江晴,张振绘. 一种用于脱机手写体女书字符切分的方法[J]. 计算机应用研究,2011,28(3):1188-1190.

[12] 何耘娴. 印刷体文档图像的中文字符识别[D]. 秦皇岛:燕山大学,2011.

[13] 马　瑞. 非限制手写字符分割中相关技术与算法的研究[D]. 南京:南京理工大学,2007.

[14] 李兴国,高　炜. 基于滴水算法的验证码中粘连字符分割方法[J]. 计算机工程与应用,2014,50(1):163-166.

[15] 靳简明,丁晓青,彭良瑞,等. 印刷维吾尔文本切割[J]. 中文信息学报,2005,18(5):76-83.

基于滴水算法的印刷体维吾尔文切分方法

作者： 朱兰， 袁保社， 余伟， ZHU Lan， YUAN Bao-she， YU Wei
作者单位： 新疆大学 信息科学与工程学院,新疆 乌鲁木齐,830046
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(7)