基于像素积分投影的印刷体维文字母切分方法

李 晓[1],袁保社[1],陈 卿[1],任宏宇[2],张建华[3]
(1.新疆大学 信息科学与工程学院,新疆 乌鲁木齐 830046;
2.94537部队气象台,河南 许昌 461101;
3.新疆公众信息产业股份有限公司,新疆 乌鲁木齐 830046)

摘 要:维吾尔文字属于左向连写文字,字母之间的连笔与变形使得切分字母很困难,印刷体维吾尔文字母的准确切分是识别的关键。文中试验了一种基于像素积分投影的印刷体维吾尔文字母切分方法,包括使用行水平投影切出文字行与文字基线,通过垂直投影切出单词及单词中不粘连的字母,结合水平投影与垂直投影数据,外加相邻投影谷距、字母宽度与基线像素值等信息,设置了细化的连体段字母切分规则。实验结果表明,该方法能够较为准确的将印刷体维吾尔文字母切分开,为OCR系统的准确识别提供了基础。

关键词:维吾尔文;印刷体;切分;像素投影积分;光学字符识别
中图分类号:TP391.43 **文献标识码**:A **文章编号**:1673-629X(2012)04-0041-04

A Segmentation Method of Printed Uyghur Character Based on Projection Histogram of Pixels

LI Xiao[1], YUAN Bao-she[1], CHEN Qing[1], REN Hong-yu[2], ZHANG Jian-hua[3]
(1. College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China;
2. Meteorological Observatory, 94537 Troops, Xuchang 461101, China;
3. Public Information Industry Co., Ltd. of Xinjiang, Urumqi 830046, China)

Abstract: The Uyghur language is a kind of the right to left and concatenate writing language, the connection and the deformation between the characters make the segmentation difficult, the accuracy of the segmentation of the characters is crucial to the recognition. Experiment a Uyghur characters segmentation method which based on the projection histogram of the pixels, including the use of the horizontal projection to separate the line of the text, separate the word and the character without adhesion in the word through the vertical projection, detailed a set of segmentation rules of the conjoined characters according to the information of the horizontal and the vertical projection, in addition the distance of the adjacent valleys of the projection, the width of the characters and the value of the pixel on the baseline, etc. Experimental results show that the method can segment the printed Uyghur text into characters effectively which provide a basis for the recognition of the OCR system.

Key words: Uyghur; printed text; segmentation; pixels projection histogram; optical character recognition (OCR)

0 引 言

光学字符识别(OCR)是涉及到图像和文字处理技术、自然语言处理、模糊数学、组合数学、信息论、人工智能、模式识别等学科的一种新技术,是人工智能领域和模式识别的一个重要的研究方向[1-3]。由于光学字符识别技术可以在很大程度上提高计算机的输入效率,所以广泛应用于办公自动化、计算机翻译、新闻出版等领域中。目前汉字、英文的识别系统已经达到国际较先进的水平[4],但是维吾尔文印刷体识别的识别率却不尽如人意。

我国维吾尔文文字识别系统还处于初级阶段,目前还没有一套完整的印刷体维吾尔文识别系统。随着维吾尔文信息化进程的发展,印刷体维吾尔文识别系统的需求将非常迫切。为了将来的研究和学习,很多维吾尔文的书籍资料需要用计算机来保存。如果这么多文字都用人工录入的方式输入计算机,效率的低下程度可想而知。所以,为了更好地、快速地对维吾尔文印刷体文字进行处理,印刷体维吾尔文文字识别软件就显得非常必要了。

收稿日期:2011-09-15;**修回日期**:2011-12-20
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:李 晓(1986-),男,湖北襄阳人,硕士研究生,研究方向为中文信息处理;袁保社,教授,研究方向为多语种系统平台与嵌入式技术。

印刷体维吾尔文文字切分就是将整篇维吾尔文文档切割成字母,供系统进行识别,其中准确的将印刷体维吾尔文字母切分开是提高OCR系统识别率的关键。对于维吾尔文字母切分,在《多字体印刷维吾尔文字符识别系统的研究与开发》[5]中,作者采用像素积分投影方法做了相关论述,文中也采用像素积分投影方法进行行和词的切分,但是对字母的切分做了一些改进,达到了比较好的效果。

1 印刷体维文切分方法

1.1 维吾尔文文字特征描述

维吾尔文文字特点如下[6,7]:

(1)文字来源:维吾尔文属于阿尔泰语系突厥语族西匈语支,维吾尔文字母借用了阿拉伯文和部分波斯文字母。

(2)文字构成特点:维吾尔文字母书写笔画简单,很多字母的主体笔画相同,增加附加部分构成不同字母。维吾尔文字母一共有32个,其中包括8个元音字母和24个辅音字母。在32个维文字母中有20个字符包含附加笔画,附加笔画包括不同数目和位置的点以及等图基本形状。附加笔画部分与字母主体部分上、下不粘连。

(3)文字书写特点:维吾尔文字是左向连写文字,行向和其他文字一样为从上往下。每个维吾尔文字母最多有四种不同的书写形式,32个字母共有120多种书写形式,字母在文本中出现的书写形式根据此字母在单词中独立与否和在连体段中的前、中、后的位置不同而不同。维吾尔文单词由一个字母或多个字母组成前后相连的连体段或字母和多个连体段混合组成。在印刷体维吾尔文连体字母段中,字母是沿着一条水平线相连的,这条水平线通常被称为基线。由于附加笔画在基线的上方或下方分布,所以可以将基线作为区分点笔画上下位置的参考线,如图1所示:

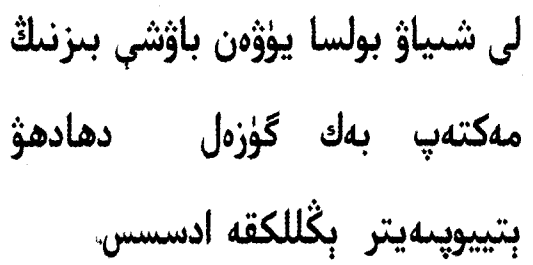

图1 维吾尔文字图片

要想对单个维文字母进行识别,就需要把维文字母从整个图像中分离出来,而要想分离出单个字母,就需要先把单词从文本中分离出来,这就是切分要完成的任务。由图1可以看到,对文字图像进行预处理之后得到的是一个整体的图片文档,其中行与行之间、单词与单词之间都有一定的空白间隙,文字部分的像素值和空白间隙部分的像素值不同,可以对文字部分和空白间隙部分进行像素积分投影,根据积分投影值不同来确定切分位置。

1.2 像素积分投影方法

由于维吾尔文印刷体识别是在Windows平台上,对印刷在纸上的多种维吾尔文字体进行扫描,以二值文件作为输入,所以在进行切分之前,首先要对图像进行去噪、二值化和图像倾斜校正处理,这样就得到了一幅没有干扰点,所有像素只有两个索引值(0表示黑,255表示白),文字像素排列没有倾斜的文字图片。

1.2.1 像素积分投影原理和目的

对图像上的像素进行探测,若索引值为0,则表示这一像素点是文字部分;若索引值为255,则表示这一像素点是空白部分。由于维吾尔文字母是沿着基线水平相连的,所以对图像上的每一行像素进行探测,记录下每行黑色像素点的数目,即对图片上的像素做积分投影,根据投影值来判断哪些行是文字行,哪些是文字间的空隙行,以确定切分位置。

在文字图片上,有文字行的理论的黑色像素点数目大于零,并且有连续的n行都为有文字的行(即有连续n行的理论黑色像素点数目大于零),在这连续的n行中,黑色像素点数目最多的一行为基线行。文字之间的空隙行的理论黑色像素点数目等于零,并且有连续的m行都为空隙行(即有连续的m行的理论黑色像素点数目等于零)。文字行和空隙行临界的地方,就是要确定作为行切分位置的行。其中,n、m值根据字体的大小有所不同。

词切分和字母切分原理与行切分基本相同,不同之处在于词切分和字母切分要在切出的每一文字行内部进行垂直投影。

1.2.2 维吾尔文的行、单词和字母切分具体实现

1)行切分。

经过扫描仪扫描的印刷体图像文本是一个整体,要想完成识别,必须将每个字符图像块切分出来,形成单个字符像素阵列,但是要想切分出单个单词,就必须先进行行切分。图像文本的行与行之间通常都有一定的空白,就对这些空白间隙做水平投影积分进行切分,将整篇文档的每一行切分开[8]。

设文字图像中的第i行、第j列的像素值为$g(i,j)$。

$$g(i,j)=\begin{cases}0 & \text{背景}\\ 1 & \text{文字上}\end{cases} \tag{1}$$

则第i行水平方向上的积分投影为$\sum_{j=1}^{L_i} g(i,j)$,其

中 L_1 为一行像素的个数。

由于图像在预处理的过程中被二值化,图像有文字部分的像素值和空白部分的像素值一定不相等,如果对图像进行水平积分投影,则图像文字行部分和空白间隙部分的积分投影值会有较大的不同。利用这一性质来确定文本图像的行切分位置。由于图像像素的存储是从下往上的,所以将最下面的一行设为第一行,按照自下而上的顺序进行行切分。

首先求出各行的积分投影:

$$F(i)=\sum_{j=i}^{L} g(i,j) \tag{2}$$

(1) 确定文本行下界[9]。

对预处理过的尚未切分的文本按照像素从下往上的顺序进行逐行搜索:

① 有连续 n 行满足:

$$(F(i)>p)\cap(F(i+1)>p)\cap\cdots\cap(F(i+n-1)>p) \tag{3}$$

② 从第 i 行到第 $i+n-1$ 中至少有一行满足:$F(k)>q$,其中 $i\leqslant k\leqslant i+n-1$。

取第一个满足以上两个条件的像素行 i 作为文本行的下界。

(2) 确定文本行上界[9]。

对预处理过的尚未切分的文本按照像素从下到上的顺序进行逐行搜索:

①有连续 m 行满足 $(F(i)<r)\cap(F(i+1)<r)\cap\cdots\cap(F(i+m-1)<r)$ (4)

② 从第 i 行到第 $i+m-1$ 中至少有一行满足:$F(k)<t$,其中 $i\leqslant k\leqslant i+m-1$。

取第一个满足以上两个条件的像素行 i 作为文本行的上界。在确定文本上下界过程中的参数 p、q、m、n、r、t 均为根据预处理去噪效果和实验情况得到的常数,其中 $n=5$,$m=3$,$r=2$,$p=2$,$q=2$,$t=2$,此数据作为参考,具体数据还需根据实验情况而定。

2)单词切分。

由图1可以看出,维吾尔文每个单词之间有明显的空隙(类似于英文),可以利用这些空隙对每一行的各个单词进行切分。虽然,单词内部也有小的空隙,但是其宽度和词与词之间的空隙宽度相比要小得多,所以切分的时候也比较容易将它们区分开,排除误切。对单词的切分是在每一行内进行的,一行切分完之后,再对下一行进行切分,直到整篇文档切分完为止。

单词的切分采用垂直投影的方法。在行切分过程中记录下每一行文字的上下边界分别为 a、b,在 a 行与 b 行之间进行垂直投影(对 a 行与 b 行之间的每一列进行投影),投影积分为

$$H(j)=\sum_{j=a+1}^{b} g(i,j) \tag{5}$$

(1)单词左边界确定。

对行切分已经切割出来的文本行按照像素从左向右的顺序进行逐列搜索:

有连续的 n 列满足:

$$(H(j)>p)\cap(H(j+1)>p)\cap\cdots\cap(H(j+n-1)>p) \tag{6}$$

取第一个满足上述条件的像素列 j 作为本行文本的一个单词的左边界列。

(2) 单词右边界确定。

对行切分已经切割出来的文本行按照像素从左向右的顺序进行逐列搜索:

有连续的 m 列满足:

$$(H(j)<r)\cap(H(j+1)<r)\cap\cdots\cap(H(j+m-1)<r) \tag{7}$$

取第一个满足上述条件的像素列 j 作为本行文本的一个单词的右边界列。

(3) 排除单词内部空隙干扰

为了排除单词内部的列空隙的干扰,词切分过程中记录下前一个单词的尾列和下一个单词的首列。

假如前一个单词的尾列为 c,后一个单词的首列为 d,则应满足:

$$(b-a)/(d-c)<e \tag{8}$$

在以上三个过程中,a、b、c、d 为实验过程中记录下的行号列号,p、r、e 为根据去噪效果和实验经验得到的常数,其中 $p=3$,$r=3$,$e=4$,此数据作为参考,具体数据还需根据实验情况而定。

3)字母切分。

在维吾尔文中,单词之内字母之间是互相连接的,和阿拉伯文很相似,而且维吾尔字母很多也来自于阿拉伯文字母。阿拉伯文字母的识别方法已经做了很多研究,这些方法对维吾尔文 OCR 系统也适用,总结有以下五种[10~12]:

(1)假设输入的是字母。由于单个字母在一篇文档里面很少以单词出现,所以这个方法不太切合实际。

(2)将输入的单词切分成比字母更小的部分。对于一个单词,在所有连接位置都做切分,这样切出来的部分甚至可能比一个字母还要小。在这种方法下,通常是先识别切出来的部分,然后将它们连接组合成字母。这种方法在联机识别和细化过的脱机识别中用的比较多。这种方法用于印刷体识别的时候会将有些字母(例如图:)切割成三部分,最终这种方法就变成了对字母、字母片段、连体段的识别。

(3)将单词切分成字母。这种方法,试着将字母准确地切分,然后再对字母进行识别。在这种方法中,切割部分对整个识别过程来说就显得至关重要。这种切割方法类似于第二种方法,但是需要加上一些判别

条件,尽量避免将一个字母切割成多个部分。有方法对细化后的单词进行切分,跟踪单词的基线,检测基线行上像素位置的上升或下降。有些字母有连续的尖端(例如图),有些单个的尖端有一些附属的点笔画在尖端的上面或下面,系统可以检测这些点来判断什么时候切分。一种有效的方法是沿着基线寻找连接点。用的比较多的还是垂直投影法,这种方法认为投影值低于某一个阈值的列为切割位置。由于在一些字母内部会有这样的列,研究人员采用不同的方法去避免误切分,例如,有些研究人员采用启发式的规则根据一个特定宽度来确定是否进行切分;有些研究人员用一种多样化投影,在这种方法中,投影值不是通过对一列所有像素进行投影,而是计算了垂直方向上的两个极值像素的距离,所切分的列就应该在极值距离小于特定阈值的列。有些对字母的切分根据字母的距离,通过训练的方法来对距离进行估计,但是通过训练数据来估计字母宽度是比较困难的,所以有些研究人员通过水平投影估计字母的高度,然后假定一个高度和字母之间的平均宽度的内在联系。有些研究人员根据单词的轮廓来进行字母切分,将单词的上轮廓上的曲率由正变为负的地方作为切分点。

(4)识别单词,将切割作为辅助部分。此方法从单词的最右边开始检查若干列,将这若干列作为一个字母识别,如果识别失败,则加一行继续识别,直到识别出一个字母。一旦这个字母被识别出来,就将它从单词中移除,用同样的方法继续识别,这种方法提高了识别速度。但是,无论在单词的哪一个部分,尤其是在单词开始的时候,如果识别失败,后面的部分就无法继续识别。对于这个问题,有人提出了如果在识别过程中失败,就从另一端(从左向右)开始用同样的方法识别。

(5)直接将输入的单词从单词的内部缝隙进行切割,将单词切割成连体段,或直接将单词作为一个整体识别而不做切割。这种方法通常将连体段或整个单词作为一个整体,用全局匹配技术将输入的单词或连体段和数据库中的单词或连体段进行全局匹配来识别。但是这种方法需要建立一个庞大的单词或连体段数据库,工作量很大。

由于印刷体文字有比较稳定的文字宽度,而且所有文字都沿着基线行。所以,识别方法选用第三种——将单词直接切分成字母。切分方法以垂直投影为主,对于一些特殊情况,用特定方法进行处理。

字母切分一般要采用予切框的概念,通过垂直投影结合基线位置得到一个可能的切点,两个切点之间构成一个予切框,通过考察本予切框内的投影特性,结合相邻字符的投影峰距与已知的字母宽度信息调整予切框,准确的框定一个完整的字母。

具体实现算法如下:

1. 对输入的单词进行垂直积分投影为 $H(j)=\sum_{i=0}^{S} g(i,j)$,先对单词内部缝隙进行切分:

① 若满足$(H(j)<u)\cap(H(j+1)<u)\cap(H(j-1)>v)$,则$j$列为连体段的右边界

② 若满足$(H(j)<u)\cap(H(j-1)<u)\cap(H(j+1)>v)$,则$j$列为连体段的左边界

③ 若满足$(H(j)<u)\cap(H(j-1)>v)\cap(H(j+1)>v)$,则$j$列为两个连体段之间的唯一一列空隙。

以上这几种情况,都是单词内部缝隙出现的几种状态,其所确定的列,是必定要在此处切分的列,列构成一个集合U。这些列还可以作为下一步连体段切分的位置参考。

其中,u,v可以分别取2,4作为参考。

2. 对单词内部的连体段进行切分:

① 若有连续的超过w列的像素积分满足$0<H(j)<7$,且这些列都满足$g(\text{baseline},j)=0$,且文字像素轮廓没有起伏,即$g(\text{baseline}+2,j)=1$,则这些列的中间位置为一个切分位置。这些切分位置构成一个集合I。

② 若在I中有元素i和集合U中的元素u之差的绝对值小于最窄字母的宽度z,即$|u-i|<z$,则去掉集合I中的i列,此处是一个误切分。

其中w、z可以选取3、8作为参考,baseline为一行文字的基线行。

2 实验结果

图2为一幅原始的未做切分维吾尔文文档图片:

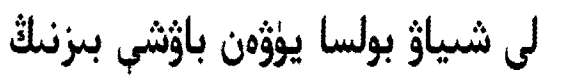

图2 未做切分的维吾尔文文档图片

维吾尔文字行切分、词切分效果,水平和垂直线段表示切分的位置标记,如图3所示:

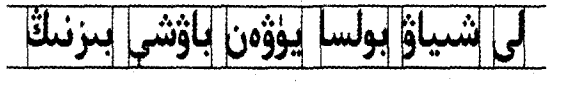

图3 对单词切分后的维吾尔文文档图片

在行、词切分基础上的字母切分效果如图4所示:

图4 在行、词切分的基础上的字母切分效果图

(下转第49页)

识别质量,为后续的模式挖掘工作打下了良好的基础。

4 结束语

Web 日志挖掘是当前研究的热点,数据预处理是 Web 日志挖掘首先要解决的问题,在数据预处理的各个步骤中,会话识别起到至关重要的作用,会话识别的质量直接决定着数据预处理质量的好坏,从而影响着 Web 日志挖掘的最终结果。文中重点研究了预处理阶段的会话识别算法,提出了基于访问时间的一次会话识别和利用断开和合并操作的二次会话识别的新会话识别方法,弥补了传统会话识别算法中真实性较低的不足,提高了数据预处理的质量。同时,文中算法也存在着不足,基于时间的一次会话所采用的访问时间是先验值,会与实际试验数据存在不匹配的地方,而且算法的效率还有待于提高,需要进一步深入的研究。

参考文献:

[1] 韩家炜,孟小峰. Web 挖掘研究[J]. 计算机研究与发展,2001,38(4):405-414.

[2] 钱小军,平玲娣,潘雪增. Web 文本挖掘技术研究及其实现[D]. 杭州:浙江大学,2002.

[3] Hseush W, Pu C. A Practical Technique for Asynchronous Transaction Processing[C]//Proceedings of ICDCS. [s. l.]: [s. n.],1995:110-117.

[4] 王听忠,王　辉,武新梅,等. 基于协同推荐的 Web 日志预处理过程[J]. 微计算机信息,2006(22):150-151.

[5] Lin Haibin, Keselj V. Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests[J]. Data & Knowledge Engineering,2007,61(2):304-330.

[6] 赵　伟,何丕廉,陈　峡,等. Web 日志挖掘中的数据预处理技术研究[J]. 计算机应用,2003,23(5):62-66.

[7] 朱　岩,杨永田,张玉清,等. 基于层次结构的信息安全评估模型研究[J]. 计算机工程与应用,2004,40(6):40-43.

[8] Wang Xidong, Ouyang Yiming, Hu Xuegang, et al. Discovery of User Frequent Access Patterns on Web Usage Mining[C]//The 8th International Conference on Computer Supported Cooperative Work in Design. [s. l.]:[s. n.],2004:765-769.

[9] 欧阳一鸣,汪曦东,郭　骏. Web 使用挖掘数据预处理中的会话构造[J]. 计算机工程与应用,2005,41(25):148-151.

[10] 张海强,胡学龙. 一种基于引用日志文件的启发式会话识别算法[J]. 扬州大学学报,2007(3):57-61.

[11] Facca F M, Lanzi P L. Mining interesting knowledge from Weblogs: a survey[J]. Data and Knowledge Engineering, 2005,53(3):225-241.

[12] 戴智丽,王鑫昱. 一种基于动态时间阈值的会话识别方法[J]. 计算机应用与软件,2010,27(2):244-246.

[13] Spiliopoulou M, Mobasher B, Berendt B, et al. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis[J]. INFORMS Journal on Computer,2003, 15(2):171-175.

(上接第 44 页)

3 结束语

上述实验结果表明,基于像素积分投影的印刷体维吾尔文切分方法,对于图 2 中的维文字体有比较好的切分效果,其他印刷体维吾尔文字体也可以采用这种方法。但是这种方法采用的是像素积分投影,由于不同字体的行列像素厚度有一定的差异,所以对于其他维吾尔文字体就需要根据字体像素厚度情况设定不同的阈值。

参考文献:

[1] 尹　芳,王卫兵,陈德运. 印刷体英文文档识别系统的设计与实现[J]. 哈尔滨理工大学学报,2008,13(6):9-12.

[2] 罗剑锋. 字符图像的分割与识别[D]. 北京:北京理工大学,2003.

[3] 吴晓峰. 基于单词全局特性的印刷体英文单词识别系统研究[D]. 广州:中山大学,2005.

[4] 欧　珠,普次仁. 印刷体藏文文字识别技术研究[D]. 西藏:西藏大学,2009.

[5] 哈力木拉提,阿孜古丽. 多字体印刷维吾尔文字符识别系统的研究与开发[J]. 计算机学报,2004,27(11):1480-1484.

[6] 王　华,丁晓青. 多字体多字号印刷体维吾尔文字符识别[J]. 清华大学学报,2004,44(7):946-949.

[7] 袁保社,吾守尔·斯拉木. 一种手写维吾尔文字母识别算法[J]. 计算机工程,2010,36(2):198-188.

[8] 董国君. 印刷体俄文文字识别研究[D]. 乌鲁木齐:新疆大学,2009.

[9] 求是科技,苏彦华. Visual C++数字图像识别技术典型案例[M]. 北京:人民邮电出版社,2004.

[10] Albadr B. A Segmentation-Free Approach to Text Recognition with Application to Arabic Text[D]. Washington: University of Washington,1995.

[11] An K H. Concurrent pattern recognition and optical character recognition[D]. USA: University of North Texas,1991.

[12] Majid M. Altuwaijri. A Parallel Recognition System for Arabic Cursive Words with Neural Learning Capabilities[D]. USA: The Graduate Faculty of The University of Southwestern Louisiana,1995.