基于汉字笔画特征的文本图像倒置判断算法

王景中,朱其猛
(北方工业大学 信息工程学院,北京 100144)

摘 要:针对目前对文本图像倒置判断过分依赖文本标点的局限以及判断准确率不理想的问题,提出了一种新的中文文本图像倒置判断算法。算法运用投影法,对汉字进行定位,充分利用汉字笔画连续属性以及动态搜寻路径寻找撇笔迹,最后根据撇笔画的轮廓与走向特征运用特定的策略与算法判定出文本的方向。此法不仅很好地解决了上述问题,同时对扭曲的文本图像的倒置判断也有良好的效果。实验结果也验证了此法的可行性与有效性。通过实验结果与现有倒置判断算法相比,此法更具普遍适用性,在效率和准确率上也得到了较大的提高。

关键词:文本定位;字符切分;笔画特征;撇笔迹;倒置判断
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2014)05-0129-05
doi:10.3969/j.issn.1673-629X.2014.05.031

A Judgment Algorithm for Inverted Chinese Text Image Based on Characteristics of Stroke

WANG Jing-zhong, ZHU Qi-meng
(Department of Information Engineering, North China University of Technology, Beijing 100144, China)

Abstract: Aiming at the problems of the judgment of inverted text image excessively relied on the punctuation and the text accuracy of judgment was not ideal, a new inversion judgment algorithm for Chinese text images is proposed. This algorithm makes full use of the outline of stroke and the characteristics of Chinese characters, on the basis of the left strokes founded by the software, using a particular algorithm to determine the direction of the text. The algorithm solves the problems above properly, at the same time, it has good effect for inversion judgment of distorted text image. The experimental results also verify the feasibility and effectiveness of this method. Compared with the existing inverted judgment algorithm, this method not only has more universal significance, efficiency and accuracy has also been largely increased.

Key words: text localization; character segmentation; character of Chinese strokes; skimming handwriting; inverted judgment

0 引 言

智能阅读器是针对盲人和视障人群开发设计的应用产品。产品的主要功能就是对印刷文本资料进行拍照或者扫描,然后对图像文本进行分析和基本处理[1],以获取文字信息并转换为语音输出。在 OCR 过程中,文本图像的方向对字符的识别结果起到决定性作用。然而由于盲人无法对文本的方向进行正确的放置,致使文本图像方向出现倾斜和倒置的情况,其中图像倾斜可以利用倾斜校正算法得以解决,目前已有很多版本的倾斜算法出台。但倒置(180 度)判断目前依然是个难题。

对于中文文本图像倒置判断,目前可以采用的方法包括基于 OCR 识别结果、基于图像特征[2]及基于文本标点符号[3]等方法。基于 OCR 识别结果方法由于要经过两次 OCR 识别处理,因此效率很不理想,并不能满足时效性的要求。基于图像特征方法主要是对图像进行投影运算,通过对投影数据进行归类分析或者利用字符行与正方向数据模板的相似度来确定图像的方向,由于图像中含有噪声或背景,此种方法对图像方向的判断准确率较低。文献[3]基于文本标点符号的方法根据文本排版中标点与文本行的相对位置属性来判断文本的方向,此法在一定程度上提高了文本图像倒置判断的效率与准确率,但对文本扭曲致使文本行

收稿日期:2013-07-15 **修回日期**:2013-10-24 **网络出版时间**:2014-02-11
基金项目:"十一五"国家科技支撑平台重点项目(2009BA171B02);国家自然科学基金资助项目(61371142)
作者简介:王景中(1962-),男,教授,研究方向为图像处理技术与应用、系统辨识与信息安全等;朱其猛(1989-),男,硕士生,研究方向为图像处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140211.1613.030.html

中字和标点相对错位的图像,此法的准确率就很不理想,同时,此法完全依靠标点,因而对标点符号少的文本图像,就失去了用武之地,因此此法的适用范围有限,不具有普遍性。

在研究了文献[4-10]中几种笔画提取的方法后,文中根据汉字笔画连续属性及撇笔画走向特征,提出一种新的中文文本图像倒置判断算法,此法消除了中文文本倒置判断过分依赖标点的限制,同时也提高了对部分扭曲的文本图像倒置判断的准确率。

1 相关理论

为了说明问题的方便,首先给出一些定义和相关理论。

1.1 邻域集

图像点阵中与当前像素点 $P(x,y)$ 相邻的像素点构成的集合,称为像素点 P 的邻域集,记为 Partner (p),Partner$(p)=\{P_0,P_1,P_2,P_3,P_4,P_5,P_6,P_7\}$,如图 1 所示。

P_5	P_6	P_7
P_2	P_3	P_4
P	P_0	P_1

图 1 邻域集

其中当前点为 $P_{(i,j)}$,则 $P_0x=i+1,P_0y=j;P_1x=i+2,P_1y=j;P_2x=i,P_2y=j+1;P_3x=i+1,P_3y=j+1;P_4x=i+2,P_4y=j+1;P_5x=i,P_5y=j+2;P_6x=i+1,P_6y=j+2;P_7x=i+2,P_7y=j+2$。

1.2 字符切分

涉及到字符分割和文本分割的文章有很多[3,11-14],文中选用的是文献[3]中的字符分割方法,并做了适当的改进。

1.2.1 行切割

行切割的一般方法是:对二值化图像从上到下逐行扫描并同时计算每扫描行的前景像素数目,以获取图像的水平投影,根据水平投影值确定文字行的位置,利用文字行间空白间隙造成的水平投影空白间隙,以确定文本行的上沿和下沿坐标,即可将各行文字分割开来,同时得到文本行的位置与行高等信息。

算法在项目中放在 OCR 版面分析之后,不需考虑文本图像中包含图片或非文本图案的噪声,但是由于文字信息中可能会包含横线或下划线等条形图案,而且此类符号的高度通常小于文本行高度的一半,即肯定低于平均行高的一半,因此可设定平均行高为阈值,对已标记行进行筛选,保证取到有效的文字行坐标。

1.2.2 字切割

字切割就是从左往右搜索一行文字单字的左右界,切分出单字和标点符号。因此可以用同样的方法,对已标记的文本行逐行进行垂直投影,利用文本行内字与字之间的空隙对各个字符加以区分,并进行标记,即可得到每个字符的精确坐标信息,同时得到每个字符的宽度、高度等信息。

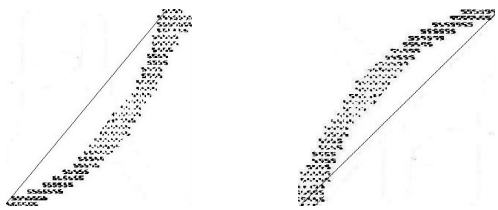
由于水平投影时已经得到每一行的高度,为了便于统计分析,认为在同一行内的文字具有同样的高度,但是,每个字符的宽度则通过垂直投影由其自身来决定。

为了提高效率,对汉字的选取也设置了阈值,就是字符的宽高比。根据统计,汉字的宽高比一般在 0.6 到 1.2 之间。鉴于此,实验中设置了双阈值:0.5 与 1.5。对于宽高比不在此范围内的字符,不做笔画的寻找。

2 算法基本原理

2.1 汉字撇笔画走向分析

汉字的主要基本笔画有四种:横、竖、撇、捺,每种笔画都有其特定的走向特征。根据统计,以上四种基本笔画出现的概率分别为横 65%,竖 50%,撇 35%,捺 24%。也就是说,平均每不到三个字中肯定会出现一个理想的撇笔画。文中在充分研究汉字笔画走向特征的基础上,发现汉字撇笔画(撇点除外)一个共同特征:从图 2 可以看出,当把撇笔画的始点和终点连线的话,笔画的其他像素点几乎位于连线的一侧(捺笔画也有相同的特性)。



(a) 正撇 (b) 倒撇

图 2 撇笔画

2.2 算法流程

鉴于上述对汉字笔画特征的分析,提出算法的基本思路:首先对文本图像进行预处理,包括倾斜校正、二值化和去噪等;然后根据投影算法确定出文本行位置及各行中字符的坐标信息。利用汉字撇笔画的走向特性,选出汉字中撇笔画,再利用一定的筛选条件舍去一些不具有代表性的撇笔画;最后根据撇笔画的走向规律,判别出图像的方向,进而进行后续的处理。算法的具体流程图如图 3 所示。

由于算法是根据笔画进行判断,所以只有提取到正确的目标笔画才能保证算法的正确率。如果对所有字符进行标记必然影响算法的性能,同时对算法的正确率没有任何帮助,所以,可以根据需要只选择其中某

几行文本进行文字定位,从中选取一定有效个数的笔画即可,这样算法可以在保证正确率的情况下同时具有较高的实时性。

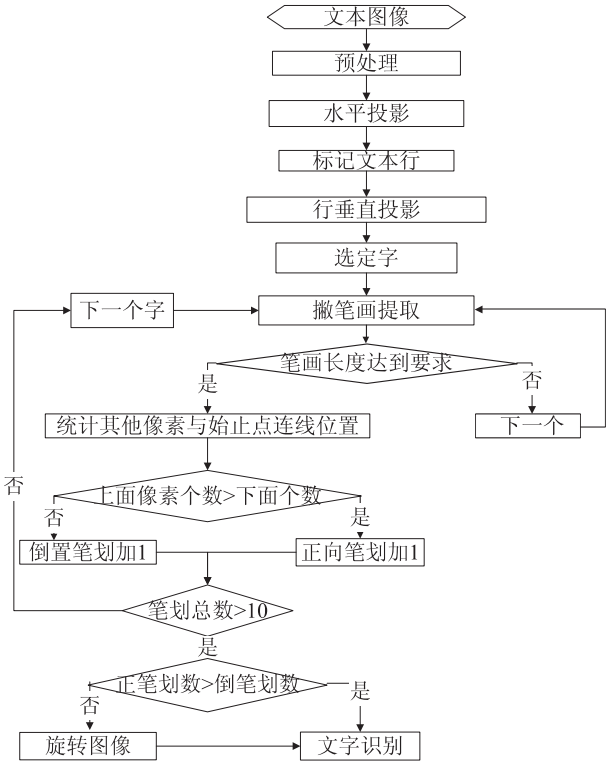


图 3　算法流程图

3　核心算法描述

3.1　撇笔画动态寻找算法

通过以上分析,可以把笔画的寻找限定在一个字符的范围内。能否找到最符合实际的目标笔画是整个算法的关键。考虑到笔画宽度,文中利用动态递归的方法对撇笔画进行跟踪。

设当前像素点为 $P_{(i,j)}$(二值化后像素值)

1) 若 $P_{(i,j)}=255$,转到下一像素点;

2) 若 $P_{(i,j)}=0$,记录起始 P 点坐标信息,优先遍历 P_3,即 $P_{(i+1,j+1)}$;

3) 若 $P_{(i+1,j+1)}=0$,i,j 均加 1,转到步骤 3;

若 $P_{(i+1,j+1)}=255$,遍历 $P_{(i,j+1)}$;

4) 若 $P_{(i,j+1)}=0$,则:

若 $P_{(i+1,j+2)}=0$,则 j 加 1,并转到步骤 4;

若 $P_{(i+1,j+2)}=255$,则转到步骤 3;

5) 若 $P_{(i,j+1)}=255$,遍历 $P_{(i+1,j)}$;

6) 若 $P_{(i+1,j)}=0$,则:

若 $P_{(i+2,j+1)}=0$,则 i 加 1,并转到步骤 5;

若 $P_{(i+2,j+1)}=255$,转到步骤 3;

7) 若 $P_{(i+1,j)}=255$,则结束。

核心算法伪代码描述:

```
int t=j,d=i;
if (lpImage[j * linebyte+i]==0) {
// 当前为黑像素点;
loop://代码跳转点 1
while(lpImage[(t+1) * linebyte+d+1]==0){
d++;t++;//寻找 45 度方向黑像素
}
if(lpImage[(t+1) * linebyte+d+1]! =0&&lpImage[(t+1)
* linebyte+d]==0){
while(lpImage[(t+1) * linebyte+d]==0&&mark1==
TRUE){ // 竖直方向查找要满足条件
if (lpImage[(t+2) * linebyte+d+1]==0){
t++;
}else{
mark1=FALSE;//此次竖直查找结束;
}
}
mark1 = TRUE;//标记复位,便于循环查找
if (lpImage[(t+1) * linebyte+d+1]==0){
goto loop;// 竖直查找后回到 45 度方向
}
}else if(lpImage[(t+1) * linebyte+d+1]! =0&&lpImage[(t
+1) * linebyte+d]! =0){//水平方向查找
while(lpImage[t * linebyte+d+1]==0&&mark2==TRUE)
{水平查找要满足条件
if (lpImage[(t+1) * linebyte+d+2]==0){
d++;
}else{
mark2 = FALSE;
}
}
mark2 = TRUE;
if (lpImage[(t+1) * linebyte+d+1]==0){
goto loop; //水平查找后回到 45 度方向
}
}
}//撇笔画查找结束
```

3.2　笔画筛选策略

上述遍历算法可以找到目标撇笔画,然而,由于笔画粘连的负面作用,使笔画提取的准确率受到了限制。必须对此进行处理,才能保证算法的强有效性。假设提取到笔画的始点与终点分别为 $P_{(i,j)}$、$P'_{(d,t)}$,对笔画矩形范围内进行像素的统计,根据像素所占整个面积的比例 ratio 对笔画加以过滤。此外,为了保证提取到的笔画的有效性,对选取笔画的长度也进行了限制。利用笔画长与宽的和与整个字的宽与高的和的比值作为筛选依据,设置一个经验值 length 为标准去除过短的撇笔画。

3.3　图像方向判断

从图 4 中可以看出,撇笔画的大多像素点都处于

笔画起始点连线的一侧,文中根据上述笔迹搜索中遍历像素点与连线的相对位置判别出汉字笔画的方向,进而判断文本方向。

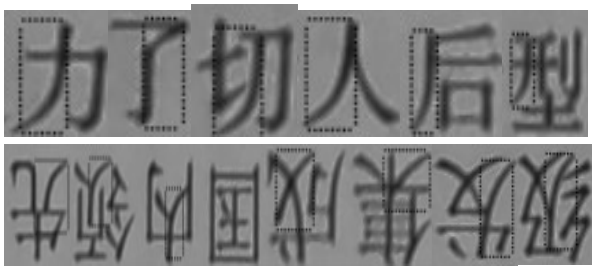


图 4 笔画提取实验效果图

算法描述如下:

1)根据上面笔迹跟踪算法定位到撇笔画的起止点,根据两点坐标信息建立直线方程;

2)统计在笔迹跟踪过程中被遍历点与直线的相对位置,若点在直线上方,则 upper 加 1,在下方则 under 加 1;

3)比较 upper 与 under 的数值大小,确定提取到笔画的方向;

4)统计一定数量的笔画,根据大部分已统计笔画的方向确定文本的方向。

算法伪代码描述:

```
k=(double(t-j))/(d-i);   //直线斜率
t=j;d=i;
if (lpImage[j*linebyte+i]==0){
loop1://代码跳转点 2
while(lpImage[(t+1)*linebyte+d+1]==0)
{ //统计像素点与连线的相对位置
d++;     t++;
f=k*(d-i)+j;
if (t<f){
under++;
}else if (t>f) {
upper++;
}
}
//统计笔画搜索过程中遍历的像素点
if(lpImage[(t+1)*linebyte+d+1]!=0&&lpImage[(t+1)
*linebyte+d]==0){
while(lpImage[(t+1)*linebyte+d]==0&&mark1==
TRUE){ // 竖直方向查找要满足条件
if (lpImage[(t+2)*linebyte+d+1]==0){
t++;
}else{
mark1=FALSE;//此次竖直查找结束;
}
}
mark1 = TRUE;//标记复位,便于循环查找
if (lpImage[(t+1)*linebyte+d+1]==0){
goto loop1;// 竖直查找后回到 45 度方向
}
}else if(lpImage[(t+1)*linebyte+d+1]!=0&&lpImage[(t
+1)*linebyte+d]!=0){
//水平方向查找
while(lpImage[t*linebyte+d+1]==0&&mark2==TRUE)
//{水平查找要满足条件
if (lpImage[(t+1)*linebyte+d+2]==0){
d++;
}else{
mark2 = FALSE;
}
}
mark2 = TRUE;
if (lpImage[(t+1)*linebyte+d+1]==0){
goto loop1; //水平查找后回到 45 度方向
}
}//笔画像素相对位置统计结束;
if (under>upper) {
izhengxiang++;//笔画正向
}else{
ifanxiang++;//笔画反向
}
……
if(izhengxiang>ifanxiang){
return 1;//图像正向
}else{
return 0;//图像倒置
}//倒置判断结束
```

特别要说明的是,为了减少一些因素(如因图像质量或部分笔画单一连接而导致笔画提取不准确)的影响,如图(4)中“了”和“先”。因此为了更好地提高算法的准确性,在算法中判断倒置,并不依据单个笔画,而是选取一定数量的笔画来判断文本图像是否倒置,通过实验验证,选取 10 个笔画就完全可以达到准确判断的要求。

4 实验结果与分析

文中算法在 Visual Studio 2005 环境下采用 C++语言实现,测试环境为:Inter(R) Core(TM)2 Duo CPU;内存 4 G;操作系统 Windows XP 的 PC 机。为了验证算法的有效性,根据实际应用需要,选取小五号、五号、四号共 3 种字号的图书和杂志样本进行验证性测试,测试内容针对 OCR 版面分析后截取的文本区域,测试结果如表 1 所示(表中的平均时间为准确判断出文本图像倒置的运行时间)。

备注:此数据是对同一样张的两种算法的测试数据。

表 1 文中算法与文献[3]方法实验结果对比

字号		小五(杂志)	五号	四号
样本数		30	30	20
文中方法	判断正确数	29	30	19
	平均时间/s	0.03	0.04	0.033
	正确率/%		97.5	
文献[3]方法	判断正确数	28	28	18
	平均时间/s	0.16	0.17	0.16
	正确率/%		92.5	

通过以上数据和结果可以看出,文中提出的基于汉字笔画的文本图像倒置检测算法能够对不同的文本图像实现快速高效的方向判断,以便根据图像方向进行后续智能处理。表 1 是针对正常样张的测试数据,与文献[3]中的算法相比,此方法的判断准确率和时间效率都得到了一定的提高,增进了整个项目的双重效率。表 2 是两种算法对一定扭曲的文本图像做倒置判断的对比,实验数据证明文中算法有相对较好的稳定性,同时,此算法对标点少的文本图像同样适用,因此文中方法有更好的鲁棒性。

表 2　两种算法对轻微扭曲文本图像的判断对比

字号		小五(杂志)	五号	四号
样本数		10	10	10
文中方法	判断正确数	8	8	9
	正确率/%	83.3		
文献[3]方法	判断正确数	6	5	7
	正确率/%	60		

5　结束语

文中从研究汉字笔画走向的特征出发,根据撇笔画的走向规律,提出了一种新的文本图像倒置判断算法:在动态跟踪撇笔画的基础上,利用线性函数,判断出文本图像的方向。实验结果也表明了该方法的科学性与有效性,切实解决了项目中面临的问题。如图 4 所示,文中方法在笔划单一连接处会发生笔画寻找错误,进而可能影响准确率的进一步提高。因此下一步的工作着重于消除笔画单一连接对准确提取撇笔画的影响,以期得到更好的判断准确率。

参考文献:

[1] 杨淑莹. VC++图像处理程序设计[M]. 第 2 版. 北京:清华大学出版社;北京交通大学出版社,2005.

[2] Liu Chunmei. Degradedcharacter recognition by image quality evaluation[C]//Proc of 2010 international conference on pattern recognition. [s. l.]:IEEE,2010:1908-1911.

[3] 曾凡锋,张国锋,陈　侃. 中文文本图像倒置检测算法[J]. 计算机工程与设计,2012,33(9):3512-3516.

[4] 刘峡壁,贾云得. 汉字笔段形成规律及其提取方法[J]. 计算机学报,2004,27(3):389-395.

[5] 孙星明,杨茂江,刘国华,等. 完全基于结构知识的汉字笔画抽取方法[J]. 计算机研究与发展,2000,37(5):543-550.

[6] 李正华,胡奇光. 汉字笔画提取的算法与实现[J]. 计算机应用与软件,2004,21(7):96-97.

[7] 张世辉. 一种新的基于距离的汉字笔画抽取方法[J]. 计算机工程,2003,29(14):37-38.

[8] 史　伟,傅　彦,陈安龙,等. 一种动态的汉字笔段提取方法[J]. 计算机应用研究,2008,25(7):1998-2000.

[9] 李建华,王　宏,闫文芝,等. 一种新的汉字细化和笔画提取方法[J]. 仪器仪表学报,2008,29(4):226-229.

[10] 何浩智. 字符识别中笔段及特征提取方法的研究[D]. 长沙:湖南大学,2007.

[11] 黄长专,王　彪,杨　忠,等. 图像分割方法研究[J]. 计算机技术与发展,2009,19(6):76-79.

[12] Grau V, Mewes A U J, Alcaniz M. Improved watershed transform for medical image segmentation using prior information [J]. IEEE Transaction on Medical Imaging,2004,23(4):447-458.

[13] Sun Jun, Wang Yan, Wu Xiaohong, et al. A new image segmentation algorithm and its application in lettuce object segmentation[J]. TELKOMNIKA,2012,10(3):557-563.

[14] Pezeshk A, Tutwiler R L. Text segmentation and reorientation from scanned color topographic maps [C]//Proc of 10th IASTED international conference on signal image process. [s. l.]:[s. n.],2008:94-97.

(上接第 128 页)

译. C 语言版. 北京:清华大学出版社,2004.

[10] Hill M D, Marty M R. Amdahl's law in the multicore era[J]. Computer,2008,41(7):33-38.

[11] 威尔金森,艾　伦. 并行程序设计[M]. 陆鑫达,译. 第 2 版. 北京:机械工业出版社,2005.

[12] Williams S, Oliker L, Vuducc R, et al. Optimization of sparse matrix-vector multiplication on emerging multicore platforms [C]//Proc of the 2007 ACM/IEEE conference on supercomputing. New York, NY, USA: ACM,2007.

[13] Amazon Company. Amazon elastic compute cloud getting started guide[M]. Seattle: Amazon,2012.

[14] Juve G, Deelman E, Vahi K, et al. Scientific workflow applications on Amazon EC2 [C]//Proc of 5th IEEE international conference on e-science workshops. Oxford: [s. n.],2009:59-66.

基于汉字笔画特征的文本图像倒置判断算法

作者： 王景中， 朱其猛， WANG Jing-zhong， ZHU Qi-meng
作者单位： 北方工业大学 信息工程学院,北京,100144
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(5)