中英文混合文章识别问题*

王 恺+, 王庆人

(南开大学 机器智能研究所,天津 300071)

Research on Chinese/English Mixed Document Recognition

WANG Kai+, WANG Qing-Ren

(Institute of Machine Intelligence, Nankai University, Tianjin 300071, China)

+ Corresponding author: E-mail: wangkai@expervision.com.cn

Received 2004-03-03; Accepted 2004-05-08

Wang K, Wang QR. Research on Chinese/English mixed document recognition. *Journal of Software*, 2005,16(5):786−798. DOI: 10.1360/jos160786

Abstract: Currently, OCR (optical character recognition) classifiers are generally designed for one character set (or language). On the other hand, multilingual document increases drastically due to the globalization. Therefore, designing a document processing system with multilingual capability is very important. A general scheme is presented in this paper: two OCR techniques, a system, and a language classification. For embodying the scheme, a Chinese/English mixed document processing system is implemented. Three key problems are considered: the control of the system flow, the classification of Chinese/English regions, and the segmentation of English characters. Compared with old systems presented in other papers, the module of the classification of Chinese/English regions is added in the system, and a novel approach based on the equidistance is applied to the module. To verify the effectiveness of the system, another system is implemented according to the methods presented in other papers. Experiment shows, the new system is more effective than the old system. The recognition rate increases from 98.48% to 99.13% on magazine samples and from 98.68% to 99.25% on book samples, respectively.

Key words: systems design; language discrimination; character segmentation; multilingual OCR (optical character recognition) system; document image processing

摘 要: 当前,已经有大量为单一字符集(或语种)而设计的 OCR(optical character recognition)分类器.同时,随着全球一体化,多语文档的出现越来越普遍.因此,设计多语文档处理系统势在必行.提出了一般性的解决方案:两项 OCR 技术、一个系统和语言判断.为了使研究工作具体化,实现了一个中英文混合文章处理系统.其中主要涉及了 3 个关键问题:系统流程控制、汉英语言区域分离和英文字符切分.与以往的系统相比,该系统增加了汉英语言区域分离模块,并将基于等间距性的新方法应用于该模块.为了验证本系统的有效性,综合以往的方法实现了另一个系统.实验结果表明,该系统的性能明显优于另一个系统,在杂志样和书籍样上的识别率分别从 98.48%

* Supported by the National Natural Science Foundation of China under Grant No.TY10026002-04-04-01 (国家自然科学基金天元基金)

作者简介: 王恺(1979－),男,天津人,博士生,主要研究领域为文档图像处理,人工智能;王庆人(1944－),男,教授,博士生导师,主要研究领域为文档图像处理,文字识别,机器人学,计算机博弈,软件开发技术.

和 98.68%提高到 99.13%和 99.25%.

关键词: 系统设计;语言判别;字符切分;多语光学字符识别系统;文档图像处理
中图法分类号: TP391 文献标识码: A

光学字符识别(optical character recognition,简称 OCR)技术的最初研究始于欧洲,德国人 Taushek 早在 1929 年就获得一项有关 OCR 的专利.为了把大量纸张文档上的文字信息电子化,并进一步利用计算机处理信息,欧美国家自 20 世纪 50 年代起就开始计算机西文 OCR 技术研究.在信息化时代,中、日、韩面临文字信息处理的关键问题是东方语言文字的识别,为此,日本在 20 世纪 70 年代投入大量资金和人力,韩国于 20 世纪 90 年代也开展了大量的 OCR 研究工作,中国的汉字识别研究始于 20 世纪 70 年代末 80 年代初,.

英文或西文 OCR 商业软件在 20 世纪 90 年代前期日渐成熟,中文 OCR 商业软件也在稍后趋于成熟.在美国联邦政府能源部的资助下,美国内华达大学拉斯维加斯分校(UNLV)信息科学研究所(ISRI)于 20 世纪 90 年代前期对各种商业系统中的 OCR 核心技术进行评测,这也是国际上该领域最严肃的评测.我国学术界稍后也对中文 OCR 核心技术进行过多次评测.以下是曾经公布的主要评测结果:

(1) UNLV 英文评测:由南开大学机器智能研究所研究的西文 OCR 核心技术(简称南开 OCR,它后来以美国 ExperVision 公司的名义推向全球市场)在 1992 年~1994 年的英文 OCR 核心技术评测中,获得性能评测全面世界第一[1−3].UNLV 评测内容包括识别率、怀疑正确标注率以及版面理解水平三大项,20 小项.其中版面理解已经超出字符识别范围,属于文档理解(document understanding)的内容了.这里顺便说明一下:在现代排版技术日趋完善的情况下,理解含有复杂段落、图形和表格的文档版面,其实是极为困难的.

(2) UNLV 中文评测:由北京信息工程学院研究的中文 OCR 核心技术(简称北信 OCR)在唯一的一次 UNLV 中文 OCR 性能评测中获得最佳成绩[4].这次评测只有识别率一项.

(3) 我国中文 OCR 评测:我国另外两项优秀的中文 OCR 核心技术分别是清华大学电子工程系丁晓青教授带领研究的中文 OCR 技术(简称清华 OCR),以及中国科学院自动化研究所刘昌平研究员带领研究的中文 OCR 技术(后来并入北京汉王公司研究院,本文简称汉王 OCR).清华 OCR、汉王 OCR 和北信 OCR 是 20 世纪 90 年代最好的 3 项中文 OCR 技术,在我国学术界组织的各次评测中都曾分别获得最佳成绩.

以上 4 项 OCR 核心技术分别适合英文或中文,但哪一项技术都无法在中、英两种语言文字表现出全面最优识别性能.或者一般地说,没有一项 OCR 核心技术能够同时圆满识别西文和东方语言两类文字,其原因是很明显的:

- 字符远近粘连:中文汉字、日文汉字以及韩文字符等东方文字有可能由两个或多个部件组成,西文字符则(除 i,j 以及顶部带变音符的字符之外)都是连通体;另一方面,东方文字的相邻字符是不粘连的,但西文相邻字符的粘连则很普遍,有时粘连还十分严重.因此,中文的字符切分困难不大,而且主要是解决如何将同一字符的分离部件合并的问题;其解决方案也很简单、很局部化.西文字符切分则是十分困难的课题,任何解决方案只有充分利用全识别过程的信息,才有可能获得满意的效果.
- 字体字符数量:东方文字字体很少,但字符数目往往在几千数量级.西文字符的数目则不过百个上下;但其字体数量很大,在几百到几千范围.因此,西文 OCR 则兼顾字符识别和字体判断,利用字体信息来提高字符识别率;东方文字 OCR 的关键在于字符识别,其受字体的干扰远没有西文那样严重.
- 形状拓扑差别:东方文字一般结构复杂、笔画繁多,但各种字体之间的差别却不是很大;西文字符则结构简单、笔划稀少,但字体变化很大,有的字体,其单个字符连“1”人都很难识别(例如 Old English 字体,人们只能靠文章上下文去猜其中的字符).毫无疑问,这两类语言文字的核心识别技术必然有很大差异.

出于以上原因,学界难以研究兼顾东、西文字识别的核心技术,难以两全其美.为适应当今信息化和全球一体化时代中英文混排文件自动处理需求,我们必须设法将最优良的西文 OCR 技术和最优良的东方文字 OCR 技术合成在一个系统之中,或者将最好的英文 OCR 技术和最好的中文 OCR 技术结合起来,构造最优良的中英文混排文件 OCR 系统.这就是本文所致力的研究课题.以下是我们提出的一般性解决方案:

(1) 两项技术.选用一项优良的西文 OCR 技术和一项优良的东方文字 OCR 技术;

(2) 一个系统.构造一个包括这两项核心技术的 OCR 系统,该系统在相对较高的层次上设置语言判断模块,以决定调用哪一个 OCR 技术来处理当前的局部文字环境;

(3) 语言判断.为语言判断模块研究一些新的技术,并改造系统结构、系统其他模块来支持该语言判断模块,使系统的双语识别性能最优化,满足识别和理解西文、东方文字混排文档的要求.

本文将这项研究工作具体化,将实验对象设定为中英文混排文章.

在中英文混合文章识别方面,已经进行了大量的研究工作.文献[5]利用投影的方法将文本行切分为字符块,并根据高度、宽度和相邻块间距进行汉英判别.实验表明,95%以上的英文和数字被正确地分离出来并送到英文 OCR 中进行识别,其他英文、数字则需要根据结构识别以及中文 OCR 拒识从文本行中分离出来,中文 OCR 采用多分类器技术以获得更好的鲁棒性.文献[6]利用投影的方法将文本行切分为字符块,并根据宽高比来决定哪些字符块可能需要合并或切分,从而生成一系列切分点,最后根据识别结果选择最佳切分路径,它使用 Gabor 特征进行中英文识别,并以笔划和外围特征对识别结果进行验证.文献[7]是文献[6]的进一步完善,它为了弥补文献[6]中不能处理粘连英文的不足,将一些易粘连字符对作为整体进行识别,并使用基于 MCE 的反例训练以拒绝非法字符图像.文献[8]利用字符拓扑结构进行语言判别,并根据识别可信度以及语义分析对判别结果进行进一步的验证.文献[9]首先对名片图像进行倾斜矫正,并抽取出文本行,然后根据连通体高宽特性以及连通体间的关系进行语言分类,并分别送到相应的分类器中进行识别,最后根据识别可信度、语义分析以及名片内在规则对判别结果进行验证.

总体上看来,文献[5–9]在汉英切分(指中文的字符切分和英文的字符切分)和汉英判别(指语言属性的判断)方面考虑较少.只是运用简单的算法进行切分,根本无法有效解决在英文中普遍出现的粘连情况.利用字符的局部拓扑结构特性进行汉英判别,这种方法无法达到很高的正确率,因此,它们把大部分的汉英判别工作寄托于识别器的性能上,但是识别前必然要进行切分,而切分要根据不同的语言属性运用不同的切分算法,如果语言属性的判别又要依赖于识别,那么就会产生矛盾,尽管用迭代的方法可以缓解这一矛盾,但是最根本的解决办法还是要从语言属性的判别入手.

本文阐述如何利用一项中文 OCR 技术和一项英文 OCR 技术构造中英文混排 OCR 系统,并使系统在双语环境下的性能极为接近或达到两项技术分别在单语环境下所达到的性能.这项研究的 3 个关键问题是:

- 系统流程控制

按照可计算性理论,模式识别问题就其本质是不可计算的.因此,一个高性能的系统不可能只依靠一个简洁算法,它必须具备复杂的控制结构、允许大量具有互补性能的算法发挥作用,相互弥补不足来保证系统的整体性能.如何协调各种算法使其发挥出最优性能是系统构造最为关键的问题.

- 汉英语言区域分离

即将文本行划分成多个区域,每个区域中的字符具有相同的语言属性(对于中英文混排文本行来说,或者是中文区域,或者是英文区域),并且相邻区域具有不同的语言属性.具有不同结构特性的语言需要采用不同的切分方法,因此,在解决多种不同结构语言混排问题时,将具有不同语言属性的区域相分离是切分前的必要步骤.对于中英文混排 OCR 系统来说,汉英语言区域的分离是最根本的操作.

- 英文字符切分

纯中文区域的汉字切分是非常简单的,英文字符则因为严重粘连而难于切分.因此,英文字符切分是中英文混排 OCR 系统中必须着重考虑的问题.

本文第 1 节~第 3 节依次就以上 3 个问题进行分析并提出解决方案.第 4 节将描述我们所做的数据规模较大的实验,并从各个方面分析实验数据,验证本文所提解决方案的有效性和普遍性.

1 系统流程控制

文献[10]中提到,生理学和心理学研究表明,一切生物系统都是反馈系统.因此,作为对人类视觉功能的模拟,OCR 系统应该也是一个反馈系统.中英文混排 OCR 系统流程如图 1 所示,分为前处理、后处理和双语 OCR 三大部分,下面分别进行介绍.

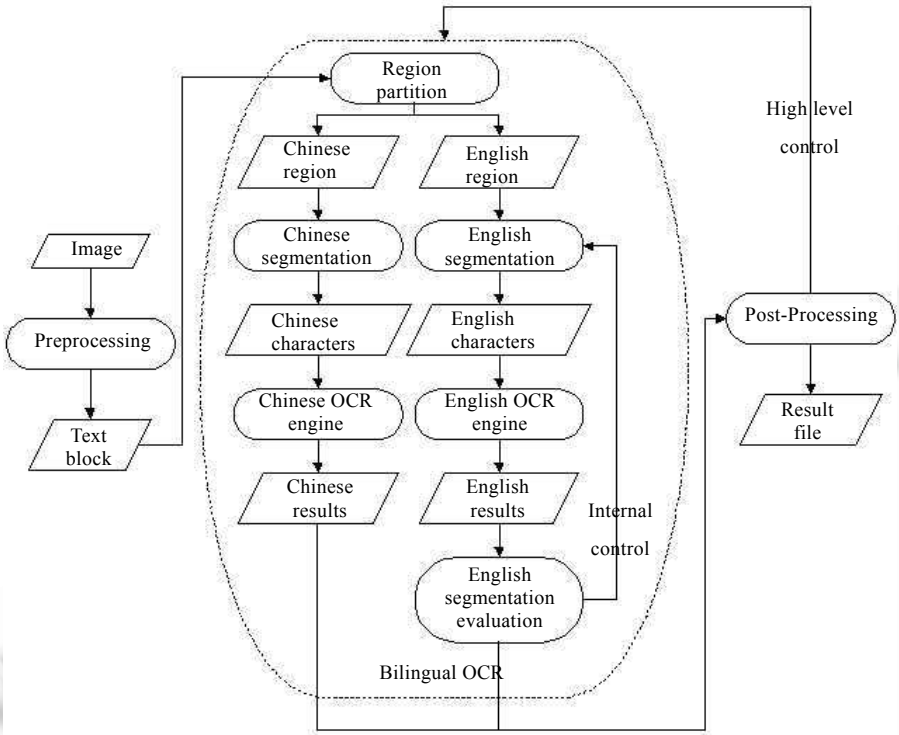


Fig.1 The flow of Chinese/English mixed OCR system
图 1 中英文混排 OCR 系统流程

(1) 前处理

包括图像输入、预处理、版面分析、段行切分和粗切分 5 个步骤,其中,前 4 个步骤的功能与普通 OCR 系统一致:

① 图像输入.将图像从扫描仪或文件系统导入内存中,开始各步处理.

② 预处理.对图像做二值化、噪音滤除、页面自动定向和倾斜校正等操作.

③ 版面分析.将文档图像划分为文本区域、图像区域和表格区域.

④ 段行切分.文本区域划分为若干文本段落和文本行.

⑤ 粗切分.为后面的双语 OCR 处理所做的特殊准备工作.

粗切分作为双语 OCR 的初始划分,利用投影操作或连通体操作(或者是二者的结合)将文本行划分为若干文本短行,使得每个文本短行:

或为某一个汉字,

或为某一汉字的一个部分,

或为某一个英文字符,

或为几个连续的英文字符.

注意,我们这里既排除了多个汉字的情况,也排除了汉字和英文字符混合的情况.这是否能够做到呢?只要以下两条假设成立,文本短行所满足的上述条件是可以保障的:

(a) 假设两个相邻汉字之间是不粘连的;

(b) 假设汉字与其附近的英文字符是不粘连的.

可以看得出,上述假设(a)和假设(b)在具有现实意义的印刷品中都是成立的.本文介绍系统的基础算法,不讨论那些不满足假设(a)或假设(b)的极端情况.出现那些极端情况的概率极小,商业系统可能根据客户需求和特殊条件给予解决,但都不在研究论文中讨论.

(2) 后处理

包括性能评价、版面恢复和结果输出 3 个步骤.

① 性能评价.根据词典拼写式检查和识别可信度评价识别结果.

② 版面恢复.依据原始排版和文字识别结果,恢复原始文件版面.

③ 结果输出.按所恢复版面输出文件,并允许用户选择各种文件格式.

高层控制

性能评估是用于决定高层控制的:根据性能评价结果来决定是否接受双语 OCR 的识别结果,如果不接受,就调整识别系统参数,重新调用双语 OCR,有时可能在某些局部区域多次反复,以提高整个系统的性能.有时某个区域所获识别性能评价无法提升,系统可能按照另一种语言属性重新进行切分-识别操作,纠正前处理的错误之后再调用双语 OCR.

(3) 双语 OCR

双语 OCR 是系统的核心部分,如图 1 所示,包括以下几个步骤:

① 汉英语言区域分离.通过归并同类文本短行将文本行划分为中文区域和英文区域,以根据不同的语言区域应用不同的字符切分算法.

② 中/英文字符切分.分别针对中/英文区域应用中/英文字符切分算法,将单个字符图像从文本行中分离出来.

③ 中/英文 OCR 引擎.OCR 引擎可以定义为实施一项具体 OCR 核心技术的软件.中文 OCR 引擎采用由北京信息工程学院研究的中文 OCR 核心技术,英文 OCR 引擎采用由南开大学机器智能研究所研究的西文 OCR 核心技术.

④ 英文切分评价.根据词典拼写式检查和识别可信度评价英文切分结果,对于评价不高的切分反复迭代英文字符切分-识别-评价,逐步改善识别结果.这实际上是英文 OCR 引擎的一种内部控制.

2 汉英语言区域分离

由于汉字保持“单摆浮隔”(即每个汉字所占宽度相等、相邻字符保持间隔)式排版,中文文本行具有明显的全局特性——字符中心的等间距性(如图 2 所示).这一特性一般不受字体和风格影响,即当同一文本行中出现不同字体、不同风格的汉字时,字符中心等间距性仍然维持.不同字号的汉字在同一文本行中出现的概率很小,本文不对这种情况进行分析.

Fig.2 The equidistance between centers of Chinese characters

图 2 汉字字符中心的等间距性

含有英文的文字行,则因英文字母、数字和标点符号间距不同而破坏等间距性(如图 3 和图 4 所示).因此,我们需用某种方法,例如线性最小二乘法来拟合出这种全局特性;然后,再根据这种全局特性完成区域划分操作;最后,利用字符本身的局部特性进行区域语言的判别.

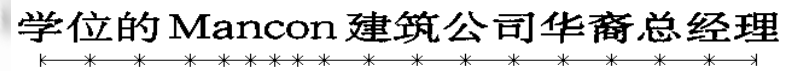


Fig.3 The equidistance is destroyed by English characters

图 3 英文字母破坏等间距性

Fig.4 The equidistance is destroyed by punctuations

图 4 标点符号破坏等间距性

2.1 全局特性抽取

上面分析指出,中文文本行具有字符中心等间距性,这一全局特性是可以用线性最小二乘法计算出来的.如果文本行因其中夹有英文字符、数字、标点符号而破坏等间距性,这种局部破坏也可以在实施线性最小二乘法时发现.

线性最小二乘法可以根据一系列数据点拟合出一条直线,使得总体方差达到最小.我们以 $y=ax+b$(a 和 b 未知)表示直线方程,假设 n 个数据点的坐标为(x_i,y_i)(i=1,2,...,n),则根据线性最小二乘法可得:

$$\begin{cases}\sum_{i=1}^{n}2x_i(ax_i+b-y_i)=0\\ \sum_{i=1}^{n}2(ax_i+b-y_i)=0\end{cases} \tag{1}$$

我们建立一个平面坐标系,取字符中心点在行中的水平位置作为数据点的 x 坐标,取字符中心点与前一字符中心点的水平距离作为数据点的 y 坐标.由于我们的目的是重新拟合出汉字中心的等间距性,因此,直线方程的形式可以简化为 $y=b$(即 a=0),代入式(1)得:

$$b=\left(\sum_{i=1}^{n}y_i\right)/n \tag{2}$$

在采集数据点时,为了尽量减少非汉字的干扰,我们使用以下两个规则对数据点进行过滤:

(a) 如果数据点在局部上不满足字符中心的等间距性(即当前字符中心距与前一字符中心距和后一字符中心距都不近似相等),则滤除该数据点;

(b) 滤除由宽度和高度都比较小的标点符号(如“.”,“、”等)产生的数据点.

图 5~图 7 分别是根据图 2~图 4 得到的字符中心分布.根据上述过滤规则:(a) 图 3 中英文“Mancon”各字符中心距变化较大,只有“*n*”与“*c*”和“*c*”与“*o*”之间的中心距近似相等,因此,由其他英文字符产生的数据点被滤除;(b) 图 4 中的“.”和“、”宽度和高度都比较小,因此,由它们产生的数据点被滤除.

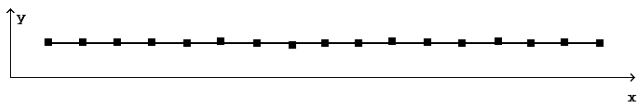


Fig.5 The distribution of the centers of characters generated by Fig.2

图 5 图 2 文字行的字符中心分布

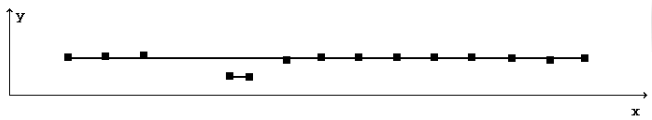


Fig.6 The distribution of the centers of characters generated by Fig.3

图 6 图 3 文字行的字符中心分布

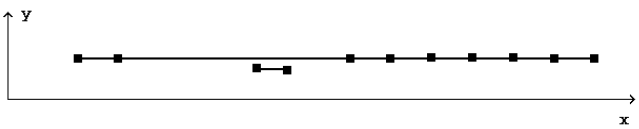


Fig.7 The distribution of the centers of characters generated by Fig.4

图 7 图 4 文字行的字符中心分布

根据图 5~图 7 中数据点的情况,采用以下步骤进行直线拟合操作:

(1) 根据式(2)由数据点确定出直线方程,根据式(3)计算数据点与直线的平均误差σ,设定阈值 H,若$\sigma<H$,则转(3),否则转(2);

$$\sigma=\left(\sum_{i=1}^{n}(b-y_i)^2\right)/n \tag{3}$$

(2) 以直线 $y=b$ 为界将数据点分为两组,并分别根据式(2)进行直线拟合操作;

(3) 如果所得汉字中心距比实际的要大,会导致一些汉字被漏掉;而如果所得汉字中心距比实际的要小,会导致一些英文字符也被合并到汉字区域中.因此,为了使拟合的直线更加精确,需要根据式(4)使用加权最小二乘法再进行一次直线拟合操作.

$$b=\left(\sum_{i=1}^{n}\omega_i y_i\right)/n \tag{4}$$

其中权重ω_i的计算规则如下:

$$\begin{cases}\omega_i=1, \text{ 当}(b-y_i)^2<H\text{时}\\ \omega_i=0, \text{ 当}(b-y_i)^2\geq H\text{时}\end{cases} \tag{5}$$

经过上述步骤后,可以拟合出一条或两条直线,如图 5~图 7 所示.这样,就可以获得汉字中心距,为后面的区域划分提供了非常有用的信息.

2.2 区域划分

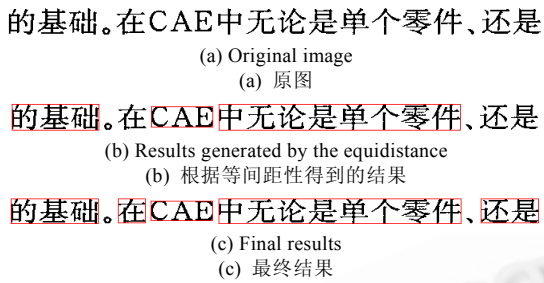
在粗切分的基础上,以第 2.1 节中介绍的方法计算出一个或两个中心距,同时将满足等间距性的区域划分出来.由于有个别汉字夹杂在英文字符中,以整体上的等间距性不能将这些汉字找出来(如图 4 中的"在"处于"."和"C"之间).针对这种情况,我们采用下面的算法进行汉字的查找操作:

算法 1.

根据第 2.3 节的区域语言判别方法,查找中文区域;
if (存在中文区域) {
D=利用第 2.1 节中介绍的方法计算中文区域中的汉字中心距;
i=1;
n=文本短行的总数;
L_x=文本短行 x 的左边缘;
R_x=文本短行 x 的右边缘;
while ($i\leq n$){
$j=i$;
取出文本短行 i;
while (($j\leq n$) && ($R_j-L_i<D$)){
if (((($i==1$)||((L_i+R_j)/2−D/2>R_j−1))
&& (($i==n$)||((L_i+R_j)/2+D/2<R_i+1)){
合并 i 到 j 之间的文本短行,并标记新生成的文本短行为汉字区域;
break;
}
j++;
}
i++;
}
}
else {
结束退出;
}

为了处理汉字倾斜的情况,上述文本短行 x 的左(右)边缘并不一定是一条竖线,也有可能是一条斜线.上述操作完成后,将连续的汉字区域合并为一个区域,同时,将不属于汉字区域的连续文本短行也合并为一个区域.

经过上述步骤后,文本行被划分为若干个区域,每个区域中含有若干个切分块,下面以此为基础进行区域语言判别.同时,根据汉字中心的等间距性,也完成了汉字的切分操作.我们以图 4 中的文本行为例,展示区域划分的全过程,如图 8 所示.



(a) Original image
(a) 原图

(b) Results generated by the equidistance
(b) 根据等间距性得到的结果

(c) Final results
(c) 最终结果

Fig.8
图 8

2.3 区域语言判别

区域语言判别以单字符语言判别作为基础,基于统计学原理,以区域为单位进行语言判别的正确率要远远高于以单字符为单位的语言判别.

我们使用文献[11]中的 Fisher 分类器进行单字符汉英判别.假设汉字判别正确率为 c,英文判别正确率为 e,一个未知区域为中文区域的概率为 a(为英文区域的概率即为 $1-a$),待判别区域用于汉英判别的字符数为 n,其中判别为汉字的字符数为 m(判别为英文的字符数即为 $n-m$).记 CR 为中文区域,ER 为英文区域.

根据贝叶斯公式:

$$P(CR\,|\,m\text{个中文})=\frac{P(m\text{个中文}\,|\,CR)\cdot P(CR)}{P(m\text{个中文}\,|\,CR)\cdot P(CR)+P(m\text{个中文}\,|\,ER)\cdot P(ER)} \tag{6}$$

其中,

$$P(m\text{个中文}\,|\,CR)=C_n^m\cdot c^m\cdot(1-c)^{(n-m)} \tag{7}$$

$$P(m\text{个中文}\,|\,ER)=C_n^m\cdot(1-e)^m\cdot e^{(n-m)} \tag{8}$$

将式(7)、式(8)代入式(6),可得:

$$P(CR\,|\,m\text{个中文})=\frac{c^m\cdot(1-c)^{(n-m)}\cdot a}{c^m\cdot(1-c)^{(n-m)}\cdot a+(1-e)^m\cdot e^{(n-m)}\cdot(1-a)} \tag{9}$$

根据文献[11]中 Fisher 分类器的结果,c=0.971,e=0.966.对于一个未知区域,中英区域的概率相等,即 a=0.5.将数据代入式(9),可得:

$$\left.\begin{aligned}P(CR\,|\,m\text{个中文})&=\frac{1}{1+\left(\frac{0.034}{0.971}\right)^m\cdot\left(\frac{0.966}{0.029}\right)^{(n-m)}}\\&\geq\frac{1}{1+\left(\frac{0.034}{0.971}\right)^m\cdot\left(\frac{0.971}{0.017}\right)^{(n-m)}}\\&=\frac{1}{1+\left(\frac{0.034}{0.971}\right)^{(2m-n)}\cdot 2^{(n-m)}}\end{aligned}\right\} \tag{10}$$

当$\begin{cases}n-m\leq 2m-n\\2m-n\geq 3\end{cases}$时,

$$P(CR\mid m\text{个中文})\geq\frac{1}{1+\left(\frac{0.068}{0.971}\right)^{(2m-n)}}\geq 99.96\% \tag{11}$$

同理可得,

$$\left.\begin{aligned}P(ER\mid m\text{个中文})&=\frac{1}{1+\left(\frac{0.029}{0.966}\right)^{(n-m)}\cdot\left(\frac{0.971}{0.034}\right)^{m}}\\&\geq \mathrm{MAX}\left(\frac{1}{1+\left(\frac{0.034}{0.971}\right)^{(n-m)}\cdot\left(\frac{0.971}{0.034}\right)^{m}},\frac{1}{1+\left(\frac{0.029}{0.966}\right)^{(n-m)}\cdot\left(\frac{0.966}{0.029}\right)^{m}}\right)\\&=\frac{1}{1+\left(\frac{0.029}{0.966}\right)^{n-2m}}\end{aligned}\right\} \tag{12}$$

当 $n-2m\geq 2$ 时,

$$P(ER\mid m\text{个中文})\geq\frac{1}{1+\left(\frac{0.029}{0.966}\right)^{2}}\geq 99.91\% \tag{13}$$

由式(11)和式(13)可以得出下列判别规则:

(1) 当 $\begin{cases}3m\geq 2n\\2m-n\geq 3\end{cases}$ 即 $m\geq\max\left(\frac{2n}{3},\frac{n}{2}+\frac{3}{2}\right)$ 时,判别为中文区域;

(2) 当 $n-2m\geq 2$ 即 $m\leq\frac{n}{2}-1$ 时,判别为英文区域;

(3) 如果不满足(1)和(2),则判别为拒绝,此时,以单字符汉英判别结果为准.

3 英文字符切分

从切分难度上来看,可以将待切分图像定义为 3 个级别.

- 0 级.相邻字符图像间可以用白竖线进行分割.
- 1 级.相邻字符图像间不粘连,但无法用白竖线进行分割.
- 2 级.相邻字符图像间存在粘连.

相应地,针对不同级别的图像,应该采用不同的切分算法.

- 0 级.竖直方向投影.
- 1 级.搜索连通体.
- 2 级.利用字符的轮廓搜索所有可能的切分点[12],生成一系列切分路径,根据英文切分评价挑选出最佳切分路径.

以往的英文字符切分方面的研究工作都是针对某个特定级别的[13],然而,在一幅待处理的文档图像中,3 种级别一般会同时存在.因此,如何协调 3 类切分算法,使系统既能利用竖直投影算法和搜索连通体算法的简单、高效,又能充分发挥搜索切分点算法对字符间粘连的复杂情况的处理能力,这是解决英文切分问题的关键.

本系统采用的控制流程如图 9 所示,英文切分评价以单词为基础,根据词典拼写式检查和识别可信度进行.通过评价决定是否需要以更复杂的方法作进一步的字符切分.

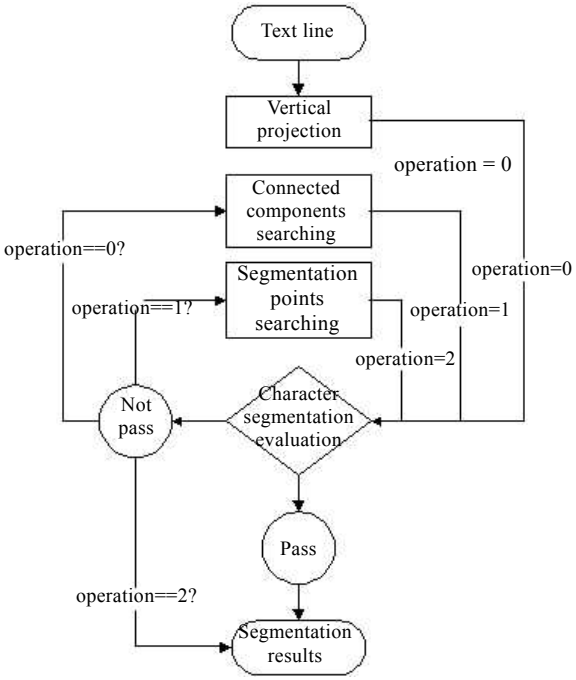


Fig.9 The control flow of the segmentation of English characters

图 9 英文字符切分控制流程

4 实验结果及分析

为了验证方法的有效性,我们根据以下标准进行采样:

(a) 采样集包括书籍和杂志;

(b) 扫描分辨率包括 300dpi 和 400dpi;

(c) 样本集涵盖各种不同的字符宽高比、字体、字号和风格的情况;

(d) 样本中英文出现率从 10%~90%.

由于数字和部分标点符号(如“!”、“?”等)与英文字符的拓扑结构非常相似,识别方法一致,因此,这里将数字、部分与英文字符拓扑结构相似的标点符号、英文字符统称为英文.其他小标点符号(如“,”,“”等等),在宽高上与汉字和英文字符有着明显区别,很容易分辨出来,因此,本实验不将它们考虑在内.

为了验证本系统的有效性,我们将文献[5−9]中提到的方法相结合,实现了另一个系统(为了叙述方便,记本系统为系统 I,另一个系统为系统 II).系统 II 与系统 I 的最主要区别是系统 II 不进行语言区域判别,而是直接以字符为单位进行语言判别.但系统 II 可以处理粘连英文字符的切分.

实验分为 4 个部分:

- 汉英语言区域分离正确率测试;
- 通过性能评价后的字符切分正确率测试;
- 系统识别率测试;
- 与单语 OCR 的比较.

4.1 汉英语言区域分离实验

系统 II 中没有汉英语言区域分离操作,因此,只给出系统 I 的测试结果,见表 1 和表 2.其中,错误 I 表示被误分到英文区域的汉字数,错误 II 表示被误分到中文区域的英文字符数.

Table 1 The result of Chinese/English region classification on magazine samples
表 1 杂志样本集上的汉英语言区域分离结果

Language	Character number	Error I	Error II	Accuracy (%)
Chinese	64 731	576	—	99.11
English	43 508	—	782	98.20
Total	108 239	576	782	98.74

Table 2 The result of Chinese/English region classification on book samples
表 2 书籍样本集上的汉英语言区域分离结果

Language	Character number	Error I	Error II	Accuracy (%)
Chinese	75 352	461	—	99.38
English	60 473	—	715	98.81
Total	135 825	461	715	99.13

实验结果表明,以区域为单位进行汉英判别达到了很好的效果,判别正确率明显高于以单字符为单位的方法,但是,比式(11)和式(13)给出的理论值要低,这主要是因为:

- 英文中夹杂着少量汉字,并且单字符汉英判别错误;
- 汉字中夹杂着少量英文字符,并且单字符汉英判别错误;
- 待判别区域在不等式(11),(13)不成立的条件,则以单字符汉英判别结果为准而引起的错误;
- 此外,在有些情况下,仅仅根据等间距性是无法确定区域边界字符的,如图 3 所示,“的”和“M”的字符中心距与汉字间的字符中心距近似相等,因此,对区域边界处字符需要进行单字符汉英判别,从而可能会引起错误.

4.2 字符切分正确率实验

测试结果见表 3 和表 4.其中,错误 I 表示被误判为英文的汉字数,错误 II 表示被误判为中文的英文字符数,错误 III 表示汉字切分错误数,错误 IV 表示英文字符切分错误数.

Table 3 Character segmentation accuracy on magazine samples
表 3 杂志样本集上的字符切分正确率结果

Language	Character number	System	Error I	Error II	Error III	Error IV	Accuracy (%)
Chinese	64 731	I	84	—	17	—	99.84
		II	179	—	76	—	99.61
English	43 508	I	—	173	—	130	99.30
		II	—	731	—	106	98.08
Total	108 239	I	84	173	17	130	99.62
		II	179	731	76	106	98.99

Table 4 Character segmentation accuracy on book samples
表 4 书籍样本集上的字符切分正确率结果

Language	Character number	System	Error I	Error II	Error III	Error IV	Accuracy (%)
Chinese	75 352	I	61	—	0	—	99.91
		II	110	—	56	—	99.78
English	60 473	I	—	162	—	187	99.42
		II	—	887	—	142	98.30
Total	135 825	I	61	162	0	187	99.69
		II	110	887	56	142	99.12

实验结果表明:

(1) 对于系统 I,通过词典拼写式检查和识别可信度评价,被误判为英文的汉字数和被误判为中文的英文字符数均有所减少.同时,根据汉字中心的等间距性进行汉字切分,可以达到很高的正确率,十多万个汉字仅有 17 个切分错误,这些错误主要是由于汉字和小标点符号离得太近引起的.英文字符的切分也达到了预期的效果,错误主要集中在粘连英文字符的切分上.

(2) 比较系统 I 和系统 II,可以看出在语言判别和汉字切分方面,系统 I 大大优于系统 II.只有英文切分方面,系统 II 略优于系统 I,这是由于:(a) 两个系统采用相同的英文字符切分算法;(b) 语言判别错误的字符数目不同,系统 II 中进入英文切分模块的英文字符数要少于系统 I.

4.3 系统识别率实验

测试结果见表 5 和表 6.其中错误 I 表示汉字的识别错误数,错误 II 表示英文字符的识别错误数.

实验结果表明,通过语言区域判别加上适当的流程控制,能够大幅度地提高系统识别率.

Table 5 Character recognition accuracy on magazine samples
表 5 杂志样本集上的系统识别率

System	Character number	Error I	Error II	Recognition accuracy (%)
I	108 239	534	402	99.13
II	108 239	677	962	98.48

Table 6 Character recognition accuracy on book samples
表 6 书籍样本集上的系统识别率结果

System	Character number	Error I	Error II	Recognition accuracy (%)
I	135825	478	530	99.25%
II	135825	568	1215	98.68%

4.4 与单语OCR系统的比较

结合表 3~表 6 中的结果,排除由于语言判别错误而导致的识别错误,可以得到单语 OCR 的识别率.表 7 和表 8 分别是杂志样张集和书籍样张集上系统 I、系统 II 和单语 OCR 识别率的比较.

Table 7 The comparison of the recognition accuracy of system I, system II and single language OCR on magazine samples
表 7 杂志样本集上系统 I、系统 II 和单语 OCR 识别率的比较

System	Chinese character number	English character number	Chinese		English	
			Character number	Recognition accuracy (%)	Character number	Recognition accuracy (%)
System I	64 731	43 508	534	99.18	402	99.08
System II	64 731	43 508	677	98.95	962	97.79
Chinese OCR	64 647	—	450	99.30	—	—
English OCR	—	43 335	—	—	229	99.47

Table 8 The comparison of the recognition accuracy of system I, system II and single language OCR on book samples
表 8 书籍样本集上系统 I、系统 II 和单语 OCR 识别率的比较

System	Chinese character number	English character number	Chinese		English	
			Character number	Recognition accuracy (%)	Character number	Recognition accuracy (%)
System I	75 352	60 473	478	99.37	530	99.12
System II	75 352	60 473	568	99.25	1 215	97.99
Chinese OCR	75 291	—	417	99.45	—	—
English OCR	—	60 311	—	—	368	99.39

其中,对于中文 OCR,中文字符总数和误识字符数的计算方法如下:

- 中文字符总数=样张中的中文字符总数–语言判别错误的中文字符数
- 中文误识字符数=本系统的中文误识字符数–语言判别错误的中文字符数

同样,对于英文 OCR,英文字符总数和误识字符数的计算方法如下:

- 英文字符总数=样张中的英文字符总数–语言判别错误的英文字符数
- 英文误识字符数=本系统的英文误识字符数–语言判别错误的英文字符数

比较结果表明,系统 I 的识别率大大高于系统 II,与单语 OCR 相比,虽然还有一些差距,但相对于系统 II 已经非常接近了.实际上,只有当语言判别错误数降到 0 时,双语 OCR 系统的识别率才能与单语 OCR 一致.

5 结 论

本文提出利用纯中文 OCR 技术和纯英文 OCR 技术构造中英文混排 OCR 系统需要解决 3 个关键问题:系统流程控制、汉英语言区域分离和英文字符切分.在第 1 节~第 3 节中,分别阐述了这 3 个问题的解决方案.实验结果表明,本文所提出的解决方案是行之有效的,系统识别率分别从 98.48%和 98.68%提高到 99.13%和 99.25%.

虽然我们所做的实验都是针对中英文混排文章的,但实际上可以将这个解决方案推广到一般的西文与东方文字混排文章的处理.

References:

[1] Among the OCR (Optical Character Recognition) products in the market, the engine developed by ExperVision won 19 out of the 20 benchmarks conducted by UNLV sponsored by the US government, Sub-title: Prof. Qingren Wang earned high marks by successfully transferring the technology developed in Nankai University to the US market place. USA: World Journal, 1993-9-17 (Headline).

[2] Rice SV, Kanai J, Nartker TA. An evaluation of OCR accuracy. Technical Report, Las Vegas: Information Science Research Institute, University of Nevada, 1993. 9−33.

[3] Rice SV, Kanai J, Nartker TA. The 3rd annual test of OCR accuracy. Technical Report, Las Vegas: Information Science Research Institute, University of Nevada, 1994. 11−38.

[4] Kanai J, Liu YC, Rice SV, Nartker TA. A preliminary evaluation of Chinese OCR systems. Technical Report, Las Vegas: Information Science Research Institute, University of Nevada, 1994. 41−47.

[5] Guo H, Ding XQ, Zhang Z, Guo FX. Realization of a high-performance bilingual Chinese-English OCR system. In: Kavanaugh M, Storms P, eds. ICDAR'95: the 3rd Int'l Conf. on Document Analysis and Recognition. Los Alamitos: IEEE Computer Society Press, 1995. 978−981.

[6] Feng ZD, Huo Q. Confidence guided progressive search and fast match techniques for high performance Chinese/English OCR. In: Kasturi R, Laurendeau D, Suen C, eds. ICPR 2002: the 16th Int'l Conf. on Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2002. 89−92.

[7] Huo Q, Feng ZD. Improving Chinese/English OCR performance by using MCE-based character-pair modeling and negative training. In: Antonacopoulos A, ed. ICDAR 2003: the 7th Int'l Conf. on Document Analysis and Recognition. Los Alamitos: IEEE Computer Society Press, 2003. 364−368.

[8] Jin JM, Wang QR. Research on multi-language OCR systems integration. Journal of Software, 2002,13:225−230 (in Chinese with English abstract).

[9] Pan WM, Jin JM, Shi GS, Wang QR. A system for automatic Chinese business card recognition. In: Antonacopoulos A, ed. ICDAR 2003: the 7th Int'l Conf. on Document Analysis and Recognition. Los Alamitos: IEEE Computer Society Press, 2003. 1138−1141.

[10] Zhu XY, Shi YF. Feedback-Based algorithm for handwritten character recognition. Chinese Journal of Computers, 2002,25(5): 476−482 (in Chinese with English abstract).

[11] Zheng YF, Liu CS, Ding XQ. Single character type identification. In: Kantor PB, Kanungo T, Zhou JY, eds. Proc. of the SPIE Document Recognition and Retrieval IX. Bellingham: SPIE—the Int'l Society for Optical Engineering, 2002,4670:49−56.

[12] Fujisawa H, Nakano Y, Kurino K. Segmentation methods for character recognition: From segmentation to document structure analysis. Proceedings of the IEEE, 1992,80(7):1079−1091.

[13] Casey RG, Lecolinet E. A survey of methods and strategies in character segmentation. IEEE Trans. on PAMI, 1996,18(7): 690−706.

附中文参考文献:

[1] ExperVision 公司研发 OCR 高科技产品,在国际同类产品 20 项评比中 19 项得第一,副标题"王庆人在南开大学研发成功技术转移来美·大放异彩".美国:世界日报,1993-9-17 (头版头条).

[8] 靳简明,王庆人.多语言字符识别系统集成研究.软件学报,2002,13:225−230.

[10] 朱小燕,史一凡.基于反馈的手写体字符识别方法的研究.计算机学报,2002,25(5):476−482.