维汉英混排文档识别

靳简明 王 华 丁晓青
(智能技术与系统国家重点实验室 清华大学电子工程系 北京 100084)

摘 要 维、汉、英是特点完全不同的文字。该文依据多层次语言判断和适当干预的多语言字符识别系统设计原则首次实现了维、汉、英混排文本识别系统。识别系统首先根据维、汉、英文字的各自特点实现字符块语言属性的初步判断，然后针对每种文字设计不同的字符切割算法。字符识别可信度用来判断字符语言属性和字符切分结果是否正确。实验结果表明，各种维、汉、英混排文本识别率达到96.4%以上。

关键词 混排文本识别，字符切割，字符识别，维吾尔文
中图分类号：TP391.43 **文献标识码**：A **文章编号**：1009-5896(2006)07-1188-04

Uyghur, Chinese and English Multilingual Document Recognition

Jin Jian-ming Wang Hua Ding Xiao-qing
(*State Key Laboratory of Intelligent Technology and Systems,*
Department of Electronic Engineering, Tsinghua University, Beijing 100084, *China*)

Abstract The characteristics of Uyghur, Chinese and English scripts are totally different. A Uyghur, Chinese and English multilingual document recognition system is implemented the first time based on the multilingual OCR system design principle, which includes “multi-layer character language estimation” and “suitable adjustment”. At first, the language property of each text block is estimated according to the characteristics of Uyghur, Chinese and English scripts. After that, language-oriented character segmentation algorithms are performed on text blocks, and the character recognition confidence is used to judge whether the results of character segmentation and language property estimation of a text block are right. Experimental results show the recognition accuracy of Uyghur, Chinese and English multilingual documents achieves 96.4% and above.

Key words Multilingual document recognition, Character segmentation, Character recognition, Uyghur script

1 引言

OCR(光学字符识别)系统能够把文本图像转化成对应的文本存入计算机，从而节省大量文字的录入工作。不同民族、不同文化之间的交流使得以一种文字为主的文档中出现其它语言文字的现象变得非常普遍。以我国为例，各种汉语言文档特别是科技类文档中经常混排出现英文字符，民族文字出版物上则会出现民族文字和汉字、英文混排的情况。目前单语种多字体印刷文档的识别率已经完全能够满足实际的需要，很长时间以来，汉字[1]、英文[2]、阿拉伯文(常见字体)[3]等语言的字符识别率就已经超过了98%。现有的商业OCR系统，比如THOCR(东方文字)、TypeReader(欧洲文字)以及Sakhr(阿拉伯文字)，在识别单语种文档时，能够得到非常令人满意的结果。但是在识别汉英、阿英等多语种混排文档时，现有系统就不能得到和识别单语种文档相媲美的结果了[4]。因此研究多文种混排文档识别很有必要。

混排文档的识别方法可以分为两类。第一，不判断字符语言属性，构造通用的多(双)语切割和识别算法。文献[5]根据字符的轮廓把拉丁字符和阿拉伯字符切割成基元，然后识别。文献[6]和文献[7]分别采用自组织的神经网络和聚类算法识别汉、韩、英3种文字。实际上，由于不同语言文字之间的巨大差异，很难构造统一的字符切割和识别算法。上述方法只能处理单字体小字符集的情况，距离实用有很大距离。第二，判断字符语言属性，针对不同语言使用不同的切割和识别方法。文献[8]构造的汉英混排识别系统在识别前进行语言判定，然后根据语言属性分别切割、识别汉英区域。但是该系统只进行一次语言判断，如果判断错误，就再无法得到正确识别结果。文献[4]提出了多层次语言判断和适当干预的多语言OCR系统设计原则，并利用该原则构造了实际的汉英混排识别系统[9]，取得了很好的效果。

维、汉、英是特点完全不同的文字，而在我国新疆地区，维、汉、英混排的文本非常普遍，所以识别维、汉、英混排文本对提高我国少数民族地区的信息化程度很有意义。本文以文献[4]提出的多语言OCR系统设计原则为指导首次构造了一个维、汉、英混排文本识别系统：第2节首先叙述了多语言OCR系统设计原则；第3节介绍了维、汉、英文字的基本特点；第4节介绍了实际混排系统判断字符语言属性的方法以及字符切割的关键技术；第5节用实验数据说明了实

2004-11-15 收到, 2005-03-31 改回
国家自然科学基金(60241005)和中国博士后科学基金(2004035331)资助课题

际系统的有效性；第 6 节给出了结论。

2　多语言 OCR 系统设计原则

我们在文献[4]中证明了多语言集成OCR系统识别错误率和语言混淆率D以及单个语言OCR系统的识别错误率e_i的关系，并在此基础上提出了多层次语言判断和适当干预的多语言OCR系统设计原则，从而最大程度地分层次逐步降低D和e_i，以有效识别多语混排文档。我们重述该设计原则如下(图 1)。

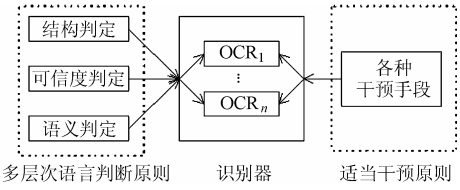


图 1 多语言 OCR 系统设计原则

Fig.1 Design principles of multi-language OCR system

第一，多层次语言判断原则。3 个层次的语言判断用来逐步降低 D：(1)结构判定：根据文本特点、字符的拓扑结构以及其他文字知识，判断字符的语言属性；(2)可信度判定：根据标记的字符语言属性，使用相应的 OCR 系统识别，接受识别可信度高的字符语言判定和识别结果，否则使用其他可能的 OCR 系统识别；(3)语义判定：通过上下文语义分析，确定极度相似字符的语言属性。

第二，适当干预原则。一般来说，识别混排文档时各个 OCR 系统面临的数据环境与系统设计时面临的数据环境是不一样的。不加干预或者干预过多都不可能得到最优的识别结果，只有适当的干预才能使系统整体识别效果达到最优。具体的干预方法和每个 OCR 系统的性能以及混排系统的应用环境有很大关系。下面是一些可以应用的干预手段：(1)区域识别：如果 OCR 系统具有字符切割功能，则把相同语言属性的文本块一次提交给 OCR 系统处理。因为提交越多的数据，系统就可以获得越准确的统计信息，也就可能得到更好的识别结果；(2)有针对性的字符切割：根据语言属性划分区域，可能出现区域内包含极少文本块或文本块跨越两个语言区域的情况。混排识别系统需要针对这些情况进行处理；(3)有针对性的字符识别后处理：针对系统具体应用环境，利用字典或语言模型，从候选字符中选出最佳结果。

3　维汉英文字特点

如表 1 和图 2 所示，维文、汉字、英文存在很大差别，所以必须针对不同的文字设计不同的字符切割和识别算法。

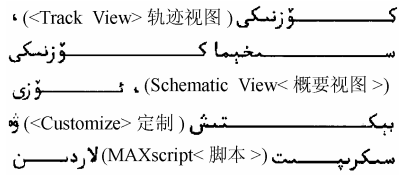


图 2 维、汉、英混排文本示例

Fig.2 Uyghur, Chinese and English mixed document

4　系统实现

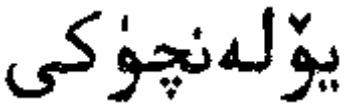
这里选用的维、汉、英 OCR 系统，都只具备单字符识别功能，也就是说只能把输入图像作为单字识别，并返回识别结果及可信度。我们设计的维、汉、英混排文本识别系统的结构与文献[4]实现的汉、英混排识别系统结构保持一致。具体识别流程如下：第一步，在切分出文字行的基础上，利用竖直投影得到文字块，自动判断每个文字块的语言属性；第二步，根据文字块标记的语言属性，进行第一次字符切分及识别；第三步，检查识别可信度，对识别结果不可信的区域，标记为其他语言属性，进行第二次字符切分及识别；第四步，根据识别可信度，合并两次识别结果。因此，该维、汉、英混排文本识别系统的关键就在于针对不同语言设计的字符切分算法，以及识别前字符语言属性的自动判断方法。

4.1　字符语言属性判断

系统在文字行的基础上，利用竖直投影得到文字块，然后根据每种语言文本的特点判断文字块的语言属性。我们首先用下面的特征(1 – 3)标记每个文字块的语言属性，然后用

表 1 维汉英文字特点

Tab.1 Characteristics of Uyghur, Chinese and English scripts

	维文	汉字	英文
文本示例		文学宝库	according
书写方向	从右往左	从左往右	从左往右
字符数目	32 个字母，每个字母 2–4 种形式(首写、尾写、中间、独立)	上万，常用字上千	26 个字母，每个字母两种形式(大写、小写)
字符特点	不等宽也不等高；由一个主体部分和 0–3 个点状附加部分组成	方块字，基本等宽等高；由一个或多个连通体组成	不等宽也不等高；一般由一个连通体组成，“i”和“j”各有一个点状附加部分
字符粘连特点	能够连接的字母总是在基线位置连接	字符之间存在间隙，一般不粘连，甚至字符内部也存在间隙	和字体和字母组合相关，包括衬线连接和字符粘连两种类型

特征(4)平滑标记的结果。

(1)文字块的宽度和高度：维文单词宽度可以很长，而汉字和英文字符的宽度短并且比较稳定，而且汉字字形近似为正方形。

(2)文字块的复杂性：汉字字符笔画复杂，英文和维文相对简单，因此可以利用竖直方向上像素 0-1 跳变数目进行区分。

(3)文字块附加部分的数目：有的维文子单词(或维文字符独立形式)和英文子串(或英文字母)非常相似，但是拥有附加部分的维文字母数目要远远超过拥有附加部分的英文字母数目。

(4)文字块语言属性的连续性：相邻文本块具有相同语言属性的概率较大，因此最后需要根据连续性对判断结果进行平滑。

4.2 维吾尔文字符切分[10]

因为维文字符通常只在基线上连接，所以只需沿着基线寻找切点。基线的位置可以通过水平投影确定。如图 3 所示，维文单词字符边界的竖直黑像素游程满足如下条件之一：(1)游程数目改变，上轮廓或下轮廓位置发生较大变化；(2)游程数目不变，但游程长度发生较大变化；(3)游程数目不变，上轮廓位置逐渐变化，导致上轮廓位置累积发生较大变化。

定义函数：

$$D(x)=\max(B_{\mathrm{Top}}-u_x,0)+\max(l_x-B_{\mathrm{Btm}},0) \quad (1)$$

其中B_{Top}和B_{Btm}分别是单词基线的上边界和下边界，u_x和l_x分别是单词第x列的上轮廓和下轮廓的位置。则所有候选切点位置x满足式(2)和式(3)或式(2)和式(4)，其中“+”和“–”分别对应字符的左边界和右边界，E是上轮廓上极小点(内凹点)的集合，H是高度阈值。

切点 x 在基线上 $D(x)\le 2$ (2)

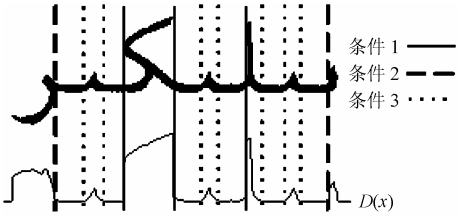


图 3 维文字符切割条件示意

Fig.3 Conditions of Uyghur characters segmentation

边界条件 1 及 2 $D(x\pm1)-D(x)>1.5H$ (3)

边界条件 3

$$\left.\begin{aligned} &D(x\pm p)-D(x)>0.75H\\ &D(x\pm1)>D(x)\\ &D(x\pm i)\ge D(x\pm i\mp1)\end{aligned}\right\} i=2\ldots p,\quad x\pm p\in E \quad (4)$$

最后，结构规则和识别信息用来合并容易被切分成多个部分的维文字符。

4.3 英文字符切分

一个文本块可能是一个英文字母或者几个粘连在一起的字符。搜索字符块上轮廓的内凹点和下轮廓的外凸点，并结合竖直投影的方法用来切割英文粘连字符。从图 4 可知：如果是衬线连接造成的粘连，则衬线连接处的竖直投影值较小；如果是非衬线连接的粘连，则字符在粘连位置处的上轮廓存在内凹点，下轮廓存在外凸点。因此上面的切割方法可以找到所有的候选英文字符切分点。但是该方法容易把“n，m，w”等字符切分成多个部分，所以和维文字符切割一样，同样需要借助识别信息去除虚假的切分点。

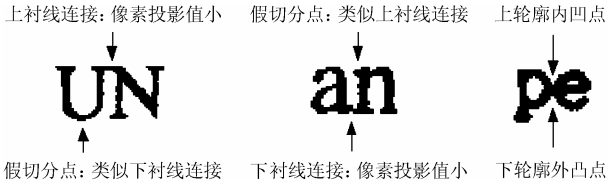


图 4 英文字符切割条件示意

Fig.4 Conditions of English characters segmentation

4.4 汉字切分

汉字之间几乎不存在粘连情况，但是字内组成部分之间也可能存在空白间距。例如“州”字可能被竖直投影成 5 个部分。因此切分汉字时，主要是通过字符的宽高比例以及相邻部分的间距，考虑几个连续部分是否可以合并，然后依据识别器的识别可信度进行确认。

5 实验结果

我们选择了维文为主，混排汉、英文字的文本图像进行测试。为了证明实际系统识别混排文本的有效性，我们还测试了纯维文文本的识别率。测试文本图像为 300DPI 的二值 TIFF 格式。测试结果参见表 2。

表 2 混排文本识别测试结果

Tab.2 Experiment results of multi-lingual text

文本类型	字符总数	混排字符数目	混排字符比例	识别率
维文	57411	0	0	98.107%
维、汉混排	13883	980	7.1%	98.891%
维、英混排	16017	2933	18.3%	96.472%
维、汉、英混排	13309	汉字 882 英文 1477 总计 2359	6.6% 11.1% 17.7%	97.092%

从实验结果可以看出，系统识别各类混排文本都保持了较高的识别率，维、汉混排文本的识别率甚至超过了纯维文文本的识别率，只是维、英混排文本的识别率略有下降。维、英混排文本识别率下降的原因主要是因为，维文字母和英文字母的结构都比较简单，而且有的字符非常相似，比如维文字符“ا”和英文字符“l”，维文字符“ە”英文字符“o”等，所以在没有字典或语义分析的情况下，非常难以区分。

6　结束语

维、汉、英是 3 种特点完全不同的文字。本文依据多层次语言判断和适当干预的多语言 OCR 系统设计原则首次实现了维、汉、英混排文本识别系统，并取得了很高的识别率，能够满足识别实际文本的要求。混排识别系统首先利用维、汉、英文字的各自特点实现字符语言属性的初步判断，然后针对不同文字使用不同的切割算法，字符的识别可信度用来判断字符语言属性和字符切分结果是否正确。因为一些维、英字符的极度相似性造成维、英混排文本的识别尚有一些缺陷，所以下面的目标是利用字典、语言模型等后处理手段进一步提高维英、维汉乃至维汉英混排文本的识别率。

参 考 文 献

[1] 刘长松, 郭繁夏, 丁晓青, 郭宏. 印刷汉字识别方法综述. 中国计算机报, 1997(663)：141−145.

[2] Rice S V, Jenkins F R, Nartker T A. The Fifth annual test of OCR accuracy. Technical Report, Information Science Research Institute, University of Nevada, Las Vegas, 1996.

[3] Kanungo T, Marton G A, Bulbul O. Omnipage vs. Sakhr： Paired model evaluation of two Arabic OCR products. SPIE Conference on Document Recognition and Retrieval VI, San Jose, CA, USA, January 27-28, 1999, 3651：109 − 120.

[4] 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(增刊)：225−230.

[5] Romeo-Pakker K, Miled H, Lecourtier Y. A new approach for Latin/Arabic character segmentation. The 3rd International Conference on Document Analysis and Recognition. Montreal, Canada, 1995: 874 − 877.

[6] Lee Seong-Whan, Kim Jong-Soo. Multi-lingual, multi-font and multi-size large-set character recognition using self-organizing neural network. The 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, 1995: 28 − 33.

[7] Chi Su-Young, Moon Kyung-Ae, Oh Weon-Geun. Recognition of large-set multilingual characters by optimal feature class reduction. The 17th International Conference on Computer Processing of Oriental Languages, HongKong, 1997: 349 − 352.

[8] Guo H, Ding X. Realization of a high-performance bilingual Chinese-English OCR system. The 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, 1995: 978 − 981.

[9] 靳简明. 汉英双语 OCR 系统集成原则及实现. 工程图学学报, 2001, 22(增刊)： 26−231.

[10] 靳简明, 丁晓青, 彭良瑞, 王华. 印刷维吾尔文本切割. 中文信息学报. 已接受.

靳简明：　男，1977 年生，博士后. 主要研究方向为图文信息处理、模式识别、图像处理.

王　华：　男，1976 年生，博士生。主要研究方向为文字识别、图文信息处理.

丁晓青：　女，1939 年生，教授，博士生导师，电子学会高级会员，中国通信学会会士. 主要研究方向为智能图文信息处理、模式识别、图像处理、文字识别、多媒体信息处理以及视频智能监测等.