印刷体藏文文字识别技术研究

欧 珠[1],普次仁[2],大罗桑朗杰[2],赵栋才[2],刘 芳[2],边巴旺堆[2]
Ngodrup[1],Putseren[2],Daluosanglangjie[2],ZHAO Dong-cai[2],LIU Fang[2],Bianbawangdui[2]

1.西藏大学 工学院,拉萨 850000
2.西藏大学 工学院 计算机科学系,拉萨 850000
1.School of Engineering,Tibet University,Lhasa 850000,China
2.Department of Computer Science,School of Engineering,Tibet University,Lhasa 850000,China
E-mail:ngodrup@utibet.edu.cn

Ngodrup,Putseren,Daluosanglangjie,et al.Study on printed Tibetan character recognition.Computer Engineering and Applications,2009,45(24):165-169.

Abstract: Owing to the special structure of Tibetan characters,the recognition of traditional Tibetan characters encounters the problems of low recognition rates and poor recognition effects.Through an in-depth study on features of the printed Tibetan characters,this paper develops a series of methods to increase recognition rate and improve the recognition effects of Tibetan characters even in the case of jamming.These methods include local self-adaptive binary algorithm,segmentation based on the connected domain,grid-based fuzzy stroke feature extraction and so on.The results of the experiments indicate that the methods can definitely increase the recognition rates of the printed Tibetan character recognition system and improve the ability to prevent jamming.
Key words: printed Tibetan character;segmentation;Tibetan character recognition;Optical Character Recognition(OCR)

摘 要:藏文字因其结构的特殊性,在应用传统文字识别方法进行识别时正确识别率较低,识别效果较差。在深入分析以印刷体藏文文字特征的基础上,提出了一系列可以在干扰情况下提高识别率的方法,包括局部自适应二值化算法、基于连通域的切分、基于网格的模糊笔划特征提取等。实验结果说明,这些方法可提高印刷体藏文文字识别系统的正确识别率和抗干扰能力。
关键词:印刷体藏文字符;切分;藏文文字识别;光学字符识别
DOI:10.3778/j.issn.1002-8331.2009.24.049 **文章编号:**1002-8331(2009)24-0165-05 **文献标识码:**A **中图分类号:**TP391.4

1 引言

信息处理技术在我国现代化以及信息化建设中,起着越来越重要的作用。十几年来,我国信息处理领域,在技术研究和产品开发以及产业建立上都取得了显著的成绩,目前汉字的识别系统已经达到国际较先进的水平。

我国藏文文字识别系统还处于起步阶段。随着藏文信息化进程的推进,藏文识别系统的需求更加突出,大量的藏文古籍、文献和资料需用计算机进行保存、处理和利用,而使用键盘输入要消耗大量的人力物力,为了更好地进行藏文处理,急需开发藏文文字识别软件,将藏文资料及图书输入到计算机中保存起来,以方便学习和研究。

文字识别是模式识别和人工智能领域的一个具体的研究方向,是模式识别、图像处理与文字处理技术相结合的一种新技术。一般通过特征判断(Feature Discrimination)及特征匹配(Feature Matching)的方法来进行处理。特征判断是通过文字类(例如英文和汉字)的共同的规则进行分类判别。它不需要利用各种文字的具体知识,根据特征抽取的程度分阶段地用结构分析的办法完成字符的识别。文字识别的关键就是特征提取。特征匹配的方法则是一个分类比对的过程,最优结果就是最后的识别结果。

2 藏文识别系统

2.1 藏文字特征描述

藏文字是一种拼音文字,可以被视为基本字符和基本字符通过纵向叠加而成的字符串, 由 30 个辅音字母和 4 个元音字母组成。构成一个完整藏文词素,基本单位是由藏文中的“音节分割符 tsheg bar”来确定。一个藏文词由一个或多个音节构成。每一个音节包含着“基字(root letter)”和可能跟随的如前加字(prefix)、上加字(head letter)、元音符号(vowel)、后加字(suffix)、再后加字(post suffix)。音节,通常是由音节分割符 tsheg bar 或者其他标点符号来划分的。图 1 给出了一个藏文字的各组成构建。

虽然现代藏文的字母数量不多,但因其拼写具有纵、横向拼写性,而且在拼写时有些字母出现变形,粘接也比较多,从字

基金项目:教育部科技创新工程重大项目培育资金项目(the Cultivation Foundation of the Key Scientific and Technical Innovation Project,Ministry of Education of China No.706059)。
作者简介:欧珠(1964-),男,教授,主要研究方向为计算机软件与理论、藏文信息处理。
收稿日期:2009-01-08 **修回日期:**2009-03-18

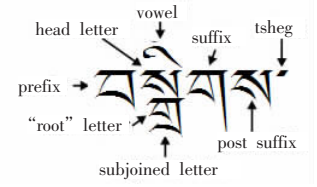

图 1 藏文字的各组成构建

符中分离出单个字母非常困难，所以在藏文识别处理中通常选取字符为基本识别单元。藏文中相似的字符多，字符相似的现象相当普遍。因此，相似字符的有效识别是藏文识别的难点所在。

2.2 印刷体藏文识别系统基本结构和运行流程

印刷体藏文识别系统一般由以下几十个模块组成，其结构如图 2 所示：

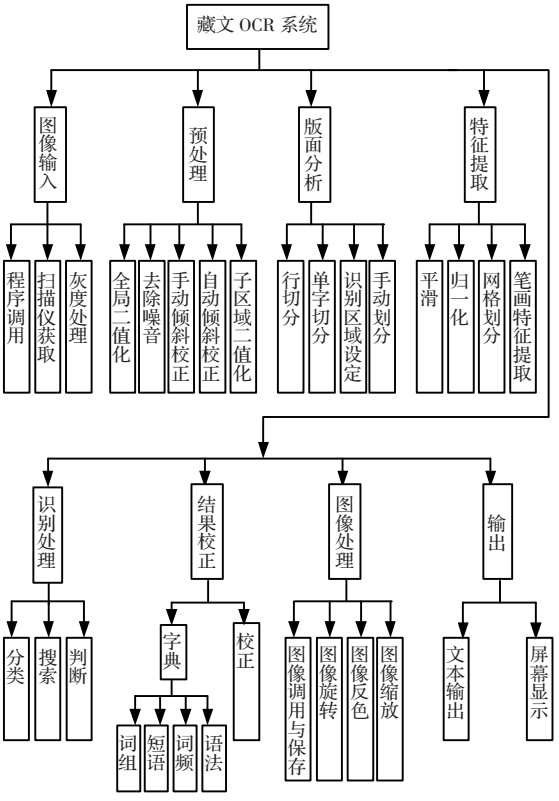

图 2 印刷体藏文识别系统结构

其运行流程如图 3 所示。

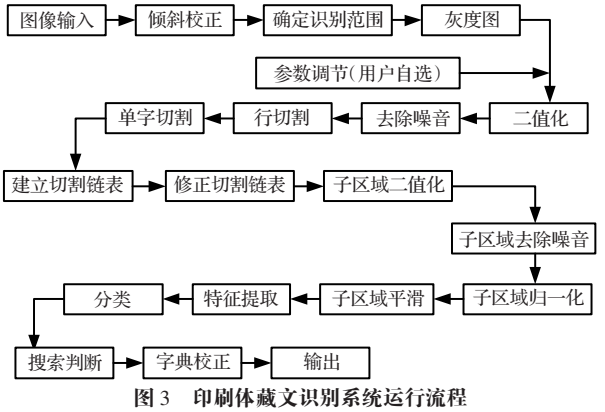

图 3 印刷体藏文识别系统运行流程

2.3 重点步骤及实现方法

2.3.1 自适应局部二值化

由于藏文文字表现形式为线条简单明显，如果二值化处理不当就会造成关键信息的缺失。如图 4 为原始图像。

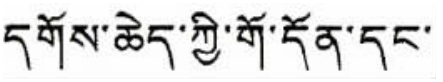

图 4 二值化前 256 色位图

采用常规方法处理，会得到如图 5 所示的二值化图像。

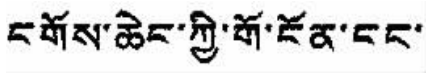

图 5 常规二值化处理后图像

其中最为代表的是藏文字"  "很容易识别成"  "，从而造成后续识别中产生大量的识别错误。

印刷体藏文识别系统中二值化采用局部自适应二值化算法，是以像素的临域信息为基础来计算每一个像素的阈值 k。设给定图像具有 L 级灰度值，对 $1<L<k$ 中的每个 k 将 $1,\cdots,k$ 分成两组，计算组 1 的像素数 $\omega_1(k)$，平均灰度 $M_1(k)$，方差 $\sigma_1(k)$；组 2 的像素数 $\omega_2(k)$，平均灰度 $M_2(k)$，方差 $\sigma_2(k)$，则：

组内方差：$\sigma_W^2=\omega_1\sigma_1^2$

组间方差：$\sigma_B^2=\omega_1\omega_2(M_1-M_2)^2$

对于一幅给定的图像可以证明 $\sigma_B^2+\sigma_W^2$=常数，因此只需求出 σ_B 的最大值，则 σ_W 自然达到最小。二值化流程如图 6 所示：

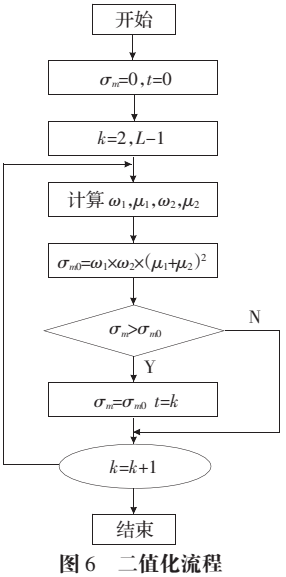

图 6 二值化流程

这种方法，在直方图中具有两个波峰（存在一个波谷）的时候，k 将取波谷处灰度值。即使不存在波谷，也可以求出最佳分界阈值 k。

印刷体藏文识别系统采用两次二值化，第一次是对整个图片全区域进行二值化，以便于进行行切分和单字切分；第二次是将已经切分后的单字符区域进行局部区域二值化，以得出单字重要信息，经过实验总结在局部二值化的过程中当二值化阈值大于 128 时，则阈值增加 10，这样获取的二值化藏文字图片更加准确。

2.3.2 基于连通域投影法的切分

藏文切分可分为行切分和单字切分。

2.3.2.1 行的切分

藏文字因为文字的特殊性导致文字上下行之间会出现重叠粘连现象，加上人为操作造成的颜色以及扫描效果的差异，

从而产生很多干扰，最典型的为下列四种情况（如图 7~图 10）：

（1）弧度干扰

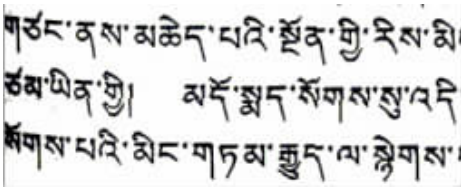

图 7　弧度干扰

这种情况是：书本扫描的时候，靠近书本中线的地方，会由于书本的弯曲产生弧度，从而导致后一行的元音部分与前一行的下元音部分产生横轴的累积重叠。

（2）粘连干扰

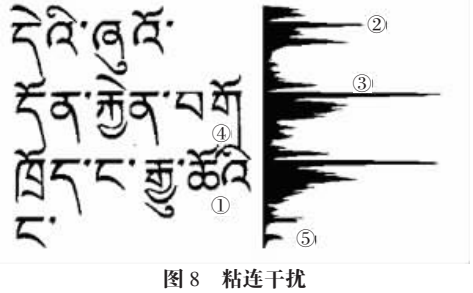

图 8　粘连干扰

图像中各点说明：

①字字粘连，宽度不统一；

②上元音与基线分离，元音处波谷最低；

③行与行之间距离较近，波谷不明显；

④行与行之间相互粘连，行的高度不统一；

⑤出现上下行文字数量变化的时候，无法判断文字数量少的那一行到底是独立文字，还是上一行部分文字的下元音。

这种情况是：行与行之间相互粘连，行内字与字之间相互粘连，并且上元音、下元音分离，从而导致行映射无明显的波谷（有的时候元音与基字结合的部分波谷值比两行间波谷值还低）。

（3）分离干扰

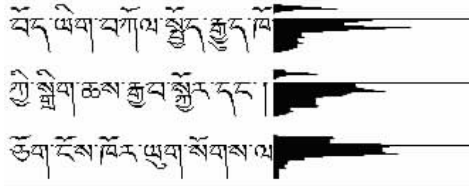

图 9　分离干扰

这种情况是：元音与主体部分相互分离，并且其距离与上一行的主体部分过于接近，在不同字号，即文字大小不相同的时候，很难判断元音部分具体是文字还是元音。

（4）各行文字个数干扰

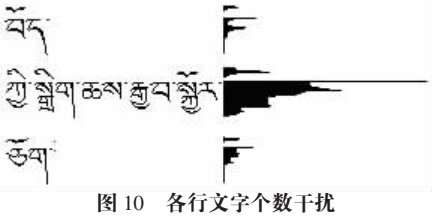

图 10　各行文字个数干扰

这种情况是：各行的文字个数不同，波峰波谷不明显。

以上四种典型的情况，通过常规的判断方法很难实现正确的切分，经过大量实验最后得出藏文行切分采用连通域投影法进行处理。

连通域投影法，首先要进行连通域搜索，即从图像最左边第一个端点 $F(x,y)$ 开始，在其周围 8 码范围内，用递归的方式查找相互连接的点的集合 $D(x,y)$，然后根据 $D(x,y)$ 确定连通区域 $Q(x,y)$。如图 11 为待识别图像，经过连通域的查找获取如图 12 的连通区域。

图 11　待识别图像

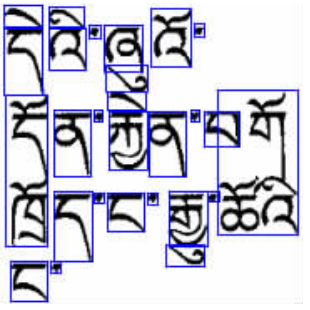

图 12　获取的连通区域

此时获取的连通域为大小不同、相互重叠、相互分离的离散区域，为获取对系统有用的区域就要进行可用性判断，判断方法为：

设连通区域高 *height* 和宽 *width*，则符合下列条件时连通域保存：

（1）*width*<*height*，标准藏文字中概率最高的情况为文字宽度小于文字高度；

（2）*height*<*width*×3，标准藏文字中概率最高的情况为文字高度小于 3 倍的文字宽度；

（3）*height*×*width*>100，识别系统接受文字的最小范围。

通过上边三个条件得出最后符合条件连通域（如图 13）。

然后对连通区域进行映射，映射的方法：将连通区域的宽度设为 1，高度除以 2，从而形成 X_n=0（n 为连通区域个数）的直线段，接着将所有的直线段对 Y 轴进行值的映射，累计 Y=Y+Y_n×M（M 为放大系数）形成新的波形（如图 14 所示）。

图 13　符合条件的连通域

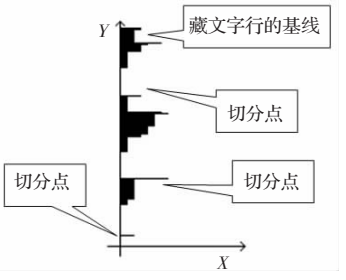

图 14　新的波形

最后按照波峰波谷位置确定切分点，从而完成行的切分。

2.3.2.2　单字切分

单字切分是在行切分的基础上进行，即将文字图像中单个文字图像提取出来，单字切分同样依据藏文字图像特点采用连通域切分法。

单字切分中遇到最典型的为下列三种情况（如图 15~图 17）：

（1）文字粘连干扰

这种情况为两个文字之间相互粘连。

（2）tsheg 点粘连干扰

这种情况为在扫描和干扰比较严重的情况下 tsheg 点与文

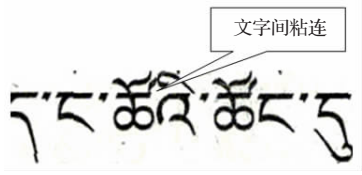

图 15 文字粘连干扰

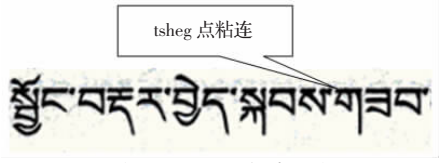

图 16 tsheg 点粘连干扰

字相互粘连。

(3)文字重叠干扰

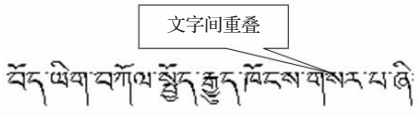

图 17 文字重叠干扰

这种情况为字与字之间并不粘连,但是在纵轴方向却相互重叠。特别是对 tsheg 点的重叠最为严重。

单字切分中连通域获取算法与行切分相同,通过连通域获取如图 18 所示。

图 18 单字切分的连通域获取

连通域获取之后,就要对区域进行整合和拆分。对于藏文字图像连通域来说,可以概括成为三种情况:上下关系、左右关系和重叠关系(如图 19 所示)。

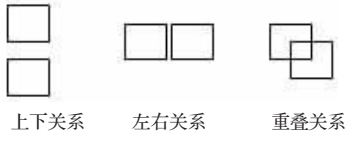

图 19 连通区域间的关系

整合拆分规则:

因为在 $D(x,y)$中包含许多代表藏文 tsheg bar、vowel 等较小区域,所以分离时首先确定 $D(x,y)$切分区域的平均宽度Dp:

$$Dp=\sum_{i=1}^{n} D_i(x,y)$$

其中 n 为 $D(x,y)$个数,然后去除所有区域宽度比 Dp 小的区域 $Dm(x,y)$,剩下的区域采用相同算法求出 Ds,Ds 可以作为 $D(x,y)$中实际文字的宽度,最后依据 $Js=Dp-Ds$,求出代表 tsheg bar 的宽度 Js。

在获取 Js 和 Ds 之后,通过下列条件进行整合拆分判断:

(1)$D(x,y)>Ds+4\times Js-3$ 时,进行区域拆分。拆分方法是将区域从中间位置一分为二,形成两个新的区域。

(2)设两个连通区域为 D_1、D_2,其边界分别为:$D_1(Left_1, Right_1)$、$D_2(Left_2, Right_2)$,则 D_1、D_2 中最左边界为:

$$MLeft=Left_1<Left_2\ ?\ Left_1:Left_2$$

最右边界为:

$$MRight=Right_1>Right_2\ ?\ Right_1:Right_2$$

当两个连通域宽度之和大于两个连通域合体后最左右两个边界的间距时,即

$$Right_1-Left_1+Right_2-Left_2>MRight-MLeft$$

则判断宽度小的区域

$$\mathrm{MIN}(Right_1-Left_1, Right_2-Left_2)$$

如果其重叠的区域大于非重叠区域时则两个连通域整合,整合方法是:将 $MLeft$ 和 $MRight$ 赋予新的区域,同时删掉原来的两个区域。如果重叠区域小于非重叠区域时则两个连通域拆分,拆分方法是在重叠区域中间位置确定 D_1 的 $Right_1$ 和 D_2 的 $Left_2$,从而达到拆分目的。

拆分整合后的效果为如图 20 所示。

图 20 拆分整合后的效果

2.3.3 基于网格的模糊笔划及轮廓特征提取

藏文特征提取采用网格设计,将切分后的单字图像归一化为 66×34 点阵信息的图片,然后平滑处理,再划分网格,最后提取各个网格中模糊笔划及轮廓特征信息。其设计主要由两个部分组成:网格划分和网格特征描述。

(1)网格划分

网格划分采用直方图的均匀等分方法进行,处理如图 21 所示。

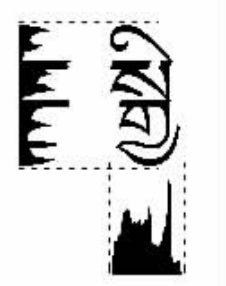

图 21 直方图的均匀等分

直方图的均匀等分就是对藏文字进行水平方向和垂直方向投影映射,然后对映射值进行平均划分,划分非均匀网格区域为 3×2,然后采用同样方法对每个划分后的网格区域再一次进行划分,每个同样划分为 3×2 非均匀网格区域,这样总的网格区域数量为 36 个网格。划分结果为如图 22 所示。

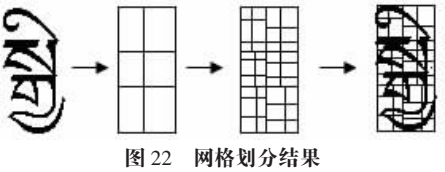

图 22 网格划分结果

(2)网格特征描述

网格特征采用笔划方向分解及笔划轮廓方向分解算法。笔划定性为黑像素笔划,笔划轮廓定性为白像素笔划。

以黑像素比划为例:设任意单个网格区域内,第一个左上角的黑像素点为 $P(x,y)$。以该点临域 8 码进行 8 个方向(如图 23 所示)上连续黑像素递归,当遇到白像素的时候,递归终止,从而获取 8 个方向长度:$f1$~$f8$。

如图 24 所示,藏文字"  "的图像,则点 P 沿 8 方向进行黑像素递归,获得 $f1$、$f2$、$f3$、$f4$、$f5$、$f6$、$f7$、$f8$。

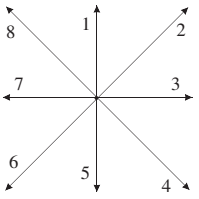

图 23　8 个方向

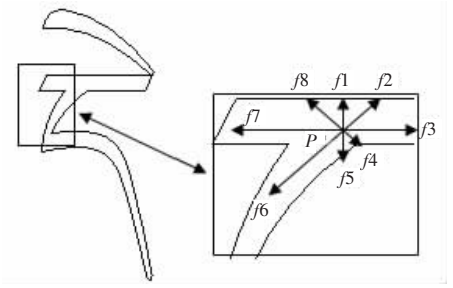

图 24　点 P 沿八个方向递归黑像素

由于扫描或字体变化等因素的干扰及影响,藏文在处理方向分解过程中采用模糊定义,即对藏文中横、竖、撇、折、弧度等笔划的具体方向按如下公式进行模糊处理:

$D1=f1+f5+f2+f6+f8+f4$

$D2=f2+f6+f3+f7+f1+f5$

$D3=f3+f7+f4+f8+f2+f6$

$D4=f4+f8+f1+f5+f3+f7$

处理之后得出的 4 方向为对应笔划的模糊方向,这样处理可以很好地去除干扰,然后再判断 $D1$~$D4$ 中最大的值,累积到 $G1$~$G4$ 中:

Max($D1$,$D2$,$D3$,$D4$)=$D1$ 则 $G1$++;

Max($D1$,$D2$,$D3$,$D4$)=$D2$ 则 $G2$++;

Max($D1$,$D2$,$D3$,$D4$)=$D3$ 则 $G3$++;

Max($D1$,$D2$,$D3$,$D4$)=$D4$ 则 $G4$++;

最后求出每个方向的分布,其中 n 为网格内黑像素的个数。

$$F1=\frac{G1}{n}\quad F2=\frac{G2}{n}\quad F3=\frac{G3}{n}\quad F4=\frac{G4}{n}$$

同理按照该方法对白像素笔划进行相同处理,从而得到 $F5$、$F6$、$F7$、$F8$ 数据。

一个网格包含 8 个特征数据,共计 36 个网格,这样单字特征数据量共计 288 维。经过实验 288 维数据在印刷体藏文识别系统中识别效率和识别时间效果最佳。

2.3.4　识别处理

识别处理采用距离分类器方法。常用的一些距离测度有:欧氏距离、加权距离、城市块距离等,印刷体藏文识别系统采用一种加权误差均衡距离,定义两个特征矢量 $\boldsymbol{X}$、$\boldsymbol{Y}$ 的距离函数为:

$$f_i(\boldsymbol{X},\boldsymbol{Y})=\sum_{i=1}^{n}[w_i(x_i-y_i)^2+\varepsilon w_i^2]$$

$$w_i=\frac{n}{(\sigma_i+a)\sum_{i=1}^{n}\frac{1}{(\sigma_i+a)}}$$

σ 是方差,ε 为 10,a 为 8。序列中距离 f 最小的结果为最后识别出的结果字符。

3　测试结果

随着国家标准藏文编码字符集(基本集、扩充集 A、扩充集 B)的制定,标准藏文编码在各种印刷件中大量应用,通过以上算法设计、优化、整合的印刷体藏文识别系统对以国家标准编码字符集基本集(部分)和扩充集 A(全部)共计 1 652 个文字进行测试,在扫描件无干扰的情况下平均识别率≥98%,平均识别效率为每秒钟识别 260 个藏文字,在扫描件存在常见干扰的情况下平均识别率≥92%,平均识别效率为每秒钟识别 260 个藏文字。具体数据参见如表 1、表 2:

表 1　各样本测试的平均识别率

数据	样本品质		
	无干扰	常见扫描干扰	
		划区域识别	整篇幅识别
训练样本	99.39%	95.76%	94.55%
测试样本	98.78%	93.93%	92.74%

表 2　各样本测试的候选字平均识别率

数据	样本品质					
	无干扰		常见扫描干扰			
			划区域识别		整篇幅识别	
候选字范围	5 字	10 字	5 字	10 字	5 字	10 字
训练样本	99.52%	99.70%	96.97%	97.58%	95.61%	96.47%
测试样本	98.97%	99.21%	95.16%	95.76%	93.64%	94.98%

测试过程中无干扰例图如图 25:

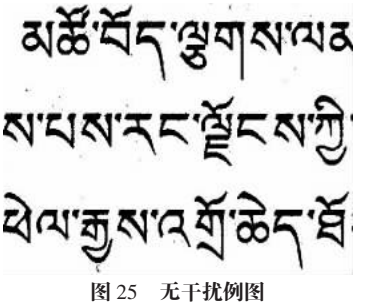

图 25　无干扰例图

常见扫描干扰例图如图 26:

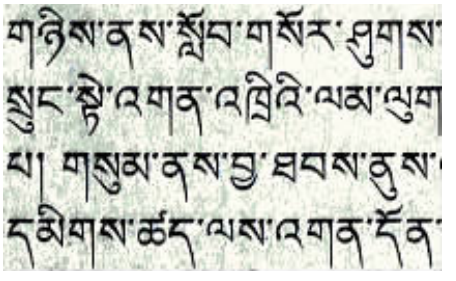

图 26　常见扫描干扰例图

4　结束语

通过对大量印刷体藏文字的研究和实验,总结了一套适合藏文印刷体文字识别的方法和手段,并将这些方法整合到一起提出了完整的识别方案,并进行优化和研究,以实现更高效的识别系统。目前该方案已经应用到"教育部 2006 年度高等学校重大工程项目培育基金项目'藏文文字识别技术研究及其实现'"中,取得了很好的效果。

参考文献:

[1] 王浩军,赵南元,邓钢铁.藏文识别的预处理[J].计算机工程,2001(9).

(下转 172 页)