

# 多字体印刷维吾尔文字符识别系统的研究与开发

哈力木拉提<sup>1)</sup> 阿孜古丽<sup>2)</sup>

<sup>1)</sup>(新疆大学信息科学与工程学院 乌鲁木齐 830046)

<sup>2)</sup>(北京科技大学信息学院知识工程研究所 北京 100083)

**摘要** 该文介绍了维吾尔文的特点及维吾尔文字符识别系统,针对维吾尔文的连体结构,重点讨论了解决过程中的技术难点. 其中利用投影分离出连体段中的字母,采用边切分边识别的方法,对文本图像进行了切分,分类,提取外围特征,并通过样张的训练,使维吾尔文字符的识别获得了较满意的结果.

**关键词** 维吾尔文; 字母切分; 字符识别; 连体字母段; 外围特征

中图法分类号 TP391

## Research and Development of a Multifont Printed Uyghur Character Recognition System

Halmurat<sup>1)</sup> Aziguli<sup>2)</sup>

<sup>1)</sup>(College of Information Science and Engineering, Xinjiang University, Urumqi 830046)

<sup>2)</sup>(Institute of Knowledge Engineering, Beijing University of Science & Technology, Beijing 100083)

**Abstract** This article introduces the application of the Uyghur sentence characteristic and Uyghur character recognition system to the connected structure of Uyghur. Authors discuss especially those aspects that solve the technological difficult points. The system utilizes projection to isolate the letters of the connected character field. Then they adopt the method of side segmentation and side recognition to perform the segmentation, classification, and extraction of outer features for the Uyghur text image, and the training of the prospectus for Uyghur. Using this character recognition algorithm, we can get relatively satisfying results.

**Keywords** Uyghur; character segmentation; character recognition; connected character field; outer feature

## 1 维吾尔文的特点

多字体印刷维吾尔文字符识别(multifont Uyghur character recognition)是指利用计算机来识别印刷在纸上的多种字体维吾尔文(通过扫描输入),并将文本文字的图像信息转化为计算机可以直接处理的文字代码形式,同时让计算机完成文本的自动输入.

维吾尔文属于阿尔泰语系突厥语族. 维吾尔文以及在中国新疆地区使用的哈萨克文、柯尔克孜文等文种都借用了阿拉伯文和部分波斯文字符. 维吾

尔文由 32 个字母组成,而且有 120 多个字符形式. 其特点如下:

(1)维吾尔文的书写方向为从右到左,行向为从上往下.维吾尔文字母有 4 种不同的书写形式:只有尾部与下一个字母相连的“首写形式”、首尾与相邻字母连接的“中间形式”、只有首部与上一个字母相连的“尾写形式”和首尾与相邻字母都不相连的“独立形式”.在维吾尔文中使用何种书写形式是根据字母在字中所处的位置来确定.完全不同于汉字、英文等.

(2)维吾尔文的词是由一个或多个字母组成. 根据书写规则,这些字母可能前后相连形成一个或几

个连体字母段或称连体段。无论是印刷体还是手写体，在连体字母段中，字母是沿着某一水平线相连的，这种水平线被称为基线。

(3)许多维吾尔字母主体相同，仅以上、下点标记说明字母的不同，有一点、二点、三点标记，但与字母主体上、下不粘连。比如：字母ب، پ، ئ، ئى，它们的字母主体为ب، 在字母主体上方标有، ، ئ， ئى的

4种特殊标记的字母为元音字母。

(4)字母不等宽。不仅不同的字母可能不等宽，而且某字母的四种形式也不等宽。甚至在文字排版时，可用直杠符号填入字母之间，使一行文本均匀分布。

由于维吾尔文的这些特点给字母的识别增加了一定的困难，尽管如此也可以从中提取不少有用的信息。用图1来进一步描述维吾尔文的特点。

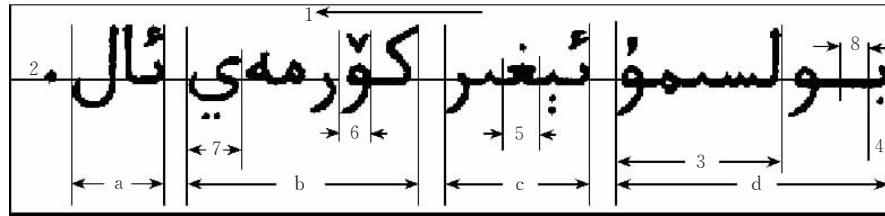


图1 维吾尔文的特点(其中：1为从右向左的书写方向；2为基线；3为五个字母的连体字母段；4为首写形式；5为中写形式；6为尾写形式；7为独立形式；8为插入的直杠；a为两字母、两部分构成的词；b为六字母、四部分构成的词；c为五字母构成的词；d为七字母、两部分构成的词)

## 2 维吾尔文识别系统构成

维吾尔文识别系统大体上可以分为两部分：识别部分和学习部分，如图2所示。识别部分是将印刷在纸上的多种维吾尔文字体经扫描仪输入作为二值文件，

并由该系统读入送至预处理部分，进行预处理。经过预处理后，有些文字模式成为规范化的二值数字信息点阵，对该二值化点阵，抽取一定的特征后，和储存在字典中已知的标准字符特征匹配判别，可识别出输入的未知字符。对于经匹配判别的结果进行后处理纠错，进一步改善识别结果，这是系统识别部分的工作过程。

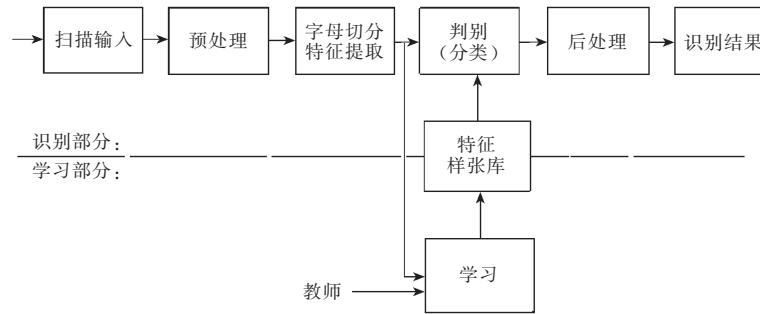


图2 维吾尔文识别系统框图

系统的学习部分是根据多个未知子样(一个文字的不同字模称子样)抽取出来的模式或特征表达形式来构造、修改、充实特征库，并用于识别的字典中，形成系统的学习功能。

## 3 维吾尔文识别系统预处理

维吾尔文识别系统预处理包括行切分、字切分、平滑、规范化等。切分的任务是把维文字符块从文本图像中提取出来。由于维文具有连体和字母不等宽特点，字母切分的准确与否直接影响字符的识别。所以，在预处理中把重点放在切分上。

### 3.1 行切分

设文本的二值图像为 $f(i, j)$ ，文本图像的大小

为 $Cx \times Cy$ ，用式(1)表示

$$f(i, j) = \{0, 1\} \quad (1)$$

式中列数 $j=1, 2, 3, \dots, Cx$ ，行数 $i=1, 2, 3, \dots, Cy$ ，行切分的目的是从一幅文本图像中，计算出一行文字像素的上下界，从而得到文本行。定义 $f(i, j)$ 在 $i$ 行上的水平投影公式为

$$H(i) = \sum_{j=1}^{Cx} f(i, j) \quad (2)$$

其中 $i=1, 2, 3, \dots, Cy$ ， $H(i)$ 反映了文本图像按行累积分布情况。分析 $H(i)$ 的分布规律，图像的水平投影为零的区域对应于图像行间的空白区，就可以获得文本行的个数，即 $n$ 个文本行， $h_1, h_2, \dots, h_n$ ，而 $h_i = H(i)$  ( $i=t+1, t+2, \dots, b$ )。

一般文本行间都存在明显的间隔，所以，如果 $i$ 行图像处于行间隔，则 $H(i)$ 为0。但有些维吾尔文

字母上方或下方都有特殊标记,与字母主体有间隙。因此需设定阈值,来判定是否作为文本行行间隔,避免字母主体与标记的分离。计算文本行间以及字母标记和字母主体间的间隙高度  $B > 0$ ,当  $B$  大于阈值,则判定为文本行行间隔,否则认为是标记与字母主体的间隙,属于同一文本行。

### 3.2 字切分(列切分)

利用文本行的垂直投影公式(3)进行列切分,根

据  $V(j)=0$ ,在一文本行里分离出字母的独立形式或几个字母连成的连体字母段。同时,我们可以很方便地将字母间或字母段之间的空白宽度记录下来,将最小的宽度记为  $w$ ,如果空白宽度比  $w$  大得较多(例如 2 倍以上),则可判定为词与词之间空白或词字与符号之间的空白,否则为字母间或连体段间空隙。这个信息告诉我们,词内只能有维文字母存在。如图 3 所示。

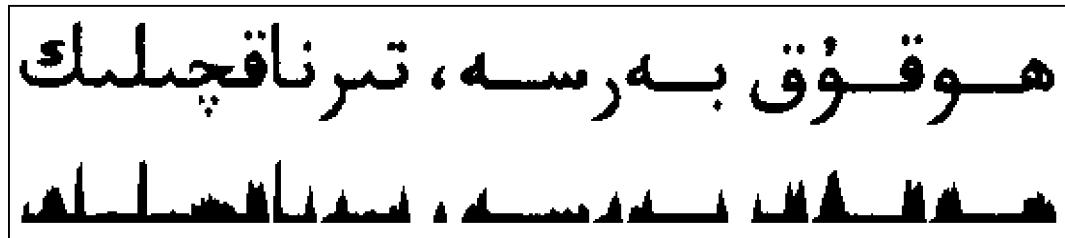


图 3 文本行的垂直投影

设垂直投影公式为

$$V(j) = \sum_{i=t}^b f(i, j) \quad (3)$$

式中  $j=1, 2, \dots, C_x$ ,  $t$  为文本像素行的上界,  $b$  为文本像素行的下界。

### 3.3 连体段中的字母切分

由于是基于字母的识别,所以,需要从连体段切分出字母。维吾尔文不像汉文,每个汉字之间都有空隙,利用这一空隙可将汉字分开。如何在连体字母段中较准确地将字母切分出来,是维吾尔字符切分中比较棘手的问题。沿着基线书写是维吾尔文的一大特点,在一行文本图像中,这一特点表现为像素集中分布于某水平线周围。因此,水平投影的最大值就是文本行的基线:  $baseline = \max(H(i))$  ( $i = 1, 2, 3, \dots, Cy$ )。如图 3。

通过行切分和字切分,切分出字母的独立形式或连体字母段。首先由公式(3)得到连体字母段( $j$ =

$s, \dots, e, s$  为起始像素,  $e$  为结尾像素)的垂直投影值。从图 5 看,它基本上能反映出连体段中字母所在的位置,投影函数  $V(x)$  的峰值中间接近基线的值,是字母与字母在基线上的连接点。这也是我们试图作为切点的最佳位置。用公式  $V(j) = (V(j-1) + V(j) + V(j+1))/3$  去掉粗糙的边缘(见图 5(b)),因为峰值比公式(4)  $av$  的值大,当投影函数  $V(x) < av$  时,把两个峰值间的中点作为切点(图 4)。由切点将字母从连体段中逐一分离出来(图 5)。

$$av = (1/NP) \sum_{n=1}^{NP} V(n) \quad (4)$$

式中  $n$  为连体段的像素;  $av$  为平均值。



图 4

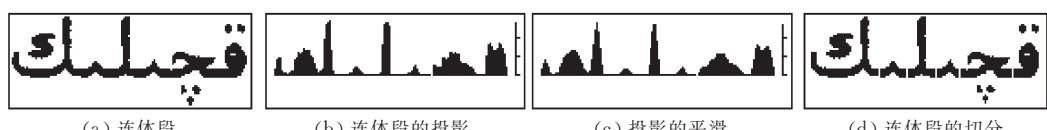


图 5 字母切分

由于连体段字母的首写和中间形式变化不是太大,这种切分方法对这两种形式的切分非常有效。对于字母尾写形式的切分和独立形式的处理,在本文的字符识别中讨论。

### 3.4 切分时字母的粗分类

维吾尔文字符的笔画少,为识别所提供的信息较少。分析行切分、字切分、连体段切分的结果,会发现切分后的字母本身就提供了分类信息,这些信息

又反映了维吾尔文字符自身的特点,只要进行切分,就有字母的三种形式(首写形式、中间形式和尾写形式)存在。所以,在进行连体字母段切分的同时,就可以做粗分类。字母的独立形式属于独立类。连体字母段的切分结果,见图 5(d),从右向左第一个字母和尾部字母分别归类于首写类和尾写类,而介于其间的字母都归于中间类。这种粗分类可缩小查寻范围,提高识别中的匹配速度和准确性。

## 4 识别系统特征提取及识别字典

切分后得到单个字母的图像(或点阵),首先要进行规范化处理,把切出的图像归到 $48\times 48$ 点阵的几何中心,为特征提取做准备。

维文字母结构简单,字符的边缘含有丰富的信息,由于内部笔画少,可以说一个字母特征几乎都表现在其边缘上。

在实际文本中,由于不同的字体、不同的字号,即使是同一字体字号的字母也不等宽和不等高,并且切点的位置也不可能精确到两字母原来的连接处,这些都会强烈地影响上述特征的准确提取,所以,在设计维文字母识别系统时,利用统计识别的方法,抽取外围特征,这种特征能较好地反映字型结构特点,对区分不同字母很有效。方法如下:

对字母的左右边缘,采用均匀提取特征,对上下边缘,由于字母中部特征比较稳定,采用以中部边缘为引导,对靠近两边的特征适当取舍,这样即使系统抗干扰能力得到加强,又能很好地反映字母边缘的变化。

外围特征的抽取过程,以左端外围特征为例:

1. 把归一化后的字母点阵图像分为 $n$ 行;
2. 计算每一行从图像左端向右扫描得到最初与文字笔画相交的长度,作为字母左边缘的外围特征。

利用同样的方法可以求出其它右、上、下3个边缘的外围特征。但上下特征提取集中在中部。如图6所示。

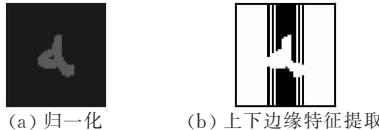


图6 特征提取

识别字典是字符识别系统的重要组成部分,其功能根据学习样张抽取字符分类特征,建立识别字典。学习的最终目的是为识别服务,识别时,是抽取待识别文字特征,与字典中的分类特征进行比较,从而做出属于哪一类的判决,通过学习可以不断改善和补充字典,提高系统的识别性能。

通过对样张的学习来建立维吾尔文识别字典,样张里包括维文符号、数字、某些字母的主体和字母。

学习的过程是抽取字符外围特征,通过聚类分析,建立或改善和补充识别字典,在学习时采用多次不同扫描参数分别扫描输入,以便使识别系统适用面更宽,对实际中不同质量的文本都能有较高的识别率,在设计识别系统时,已对当前较为流行的北大方正排版系统维文部分的8种字体(每种字体的7种字号,共56个样张)进行了学习。这8种字体为红旗白体、红旗黑体、红旗细黑、新红旗白体、书黑、报白、报黑、书白。由于学习功能的建立,可根据需要增

加系统所能识别的维文字母样张。

## 5 维吾尔文识别系统的识别方法

和其它模式识别一样,维文文字识别的基本思想也是匹配判别。抽取代表未知字母模式本质的表达式(如各种特征)和预先存储在机器里的标准字母模式表达式的集合(字典)逐一匹配,用一定的准则进行判别,在机器存储的标准维文字母模式表达式的集合中,找出最接近输入字母模式的表达式,该表达式对应的字母就是识别结果。

为了简便和提高运算速度,我们选用了分类判别准则 Minkowsky 距离,令 $q=1$ ,即绝对值距离。

$$D(\mathbf{X}, \mathbf{G}) = \left[ \sum_{i=1}^m |x_i - g_i|^q \right]^{1/q} \quad (5)$$

式中 $\mathbf{X}$ 表示切分后未知字母的特征向量, $\mathbf{X}=(x_1, x_2, \dots, x_m)$ ; $\mathbf{G}$ 为字典中某一标准文字特征向量, $\mathbf{G}=(g_1, g_2, \dots, g_m)$ 。

利用距离准则来判别时,当切分后未知字母的特征矢量 $\mathbf{X}$ 和字典中某一标准文字的特征矢量 $\mathbf{G}_j$ 相同时,则 $D(\mathbf{X}, \mathbf{G}_j)=0$ 。所以,分出计算切分后未知字母特征矢量 $\mathbf{X}$ 和字典中所有标准文字特征矢量 $\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_q$ 之间的距离( $D(\mathbf{X}, \mathbf{G}_1), D(\mathbf{X}, \mathbf{G}_2), \dots, D(\mathbf{X}, \mathbf{G}_q)$ ),求出最小值 $D(\mathbf{X}, \mathbf{G}_j)$ ,且 $D(\mathbf{X}, \mathbf{G}_i) \leq \delta$ ( $\delta \geq 0$ ,是预定的常数),即可判别出输入未知文字属于哪一类。当此类中,如果只有一个字母,就识别出未知字母是字典中匹配的特征向量 $\mathbf{G}_i$ 的哪个字母。如果是一批,就选出在此类中出现频率最高的字母。

字母识别的正确与否依赖于字母的正确切分。由于维文字母在词中的变化,以上的切分方法有时产生误切,一是字母的独立形式被作为连体段切分;二是连体段尾部的尾写形式也作为中间形式被切分,针对这种情况,我们提出了在维吾尔文字符切分中采用边切分边识别的方法。

有些独立形式的宽度比连体段还要宽,做垂直投影时,它同样具备做切分的条件,如:**ئ**和**ە**,切分时,前者有3个切点将被分割成4段,识别为3个字母。显然,这是由于对不容许切分的独立形式,进行了切分,导致识别错误。一般来说,维文字母的宽度要比它所在的文本行的高度要小,根据这一特征,首先,把连体段宽度小于文本行高度的连体段作为独立类来识别,如果没有被识别,再作为连体段处理。见图7(a)。这种方法,可以使这些不容许切分的独立形式得到辨认。

在连体段中,衔接在尾部的尾写形式,由于尾写形式的多样性,同样有切点存在。比如字母**ۇ**就有4个或5个切点,如果按每个切点去识别字母,将导致同样的识别错误。所以,我们以识别结果作为指导

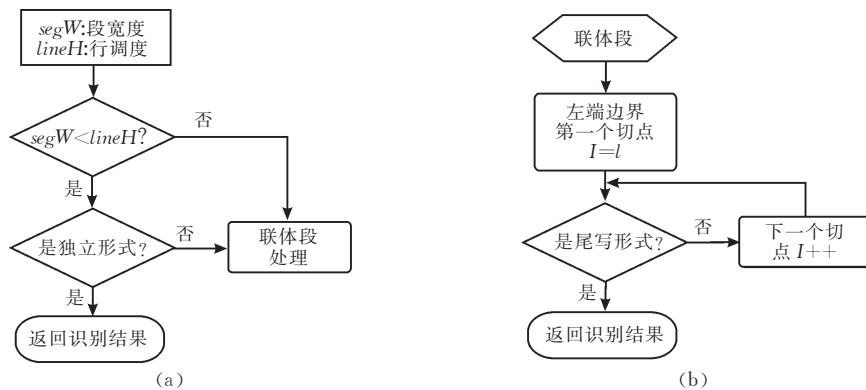


图 7 边切分边识别过程

来切分,首先,连体段左端为尾写字母的左边界,向右以第 1 个切点作为右边界开始试切,切出一块与尾写类进行匹配,如果不是尾写形式,再以第 2 切点位置作为右边界试切,进行第 2 次识别,这一过程最多可以到第 5 个切点,因为在尾写形式中,最多有 5 个切点. 见图 7(b). 这样,可以使那些不易被切分的尾写字母得到较好的识别.

实验表明这种方法是有效的,特别是对独立形式的识别非常有效.

## 6 结 论

通过 56 种各种字体字号样张进行学习,建立起多字体、多字号的维吾尔文字符识别系统. 利用垂直投影分离出连体段中的字母,采用边切分边识别的方法,对文本图像进行了切分、分类、提取外围特征、综合全局信息和局部信息的分析和应用,使维文字符的识别获得了比较满意的结果. 对于质量较好的



**Halmurat**, born in 1959, associate professor. His research interests include pattern recognition and information process technology of the minority of Xinjiang.

### Background

Uyghur, Kazak and Kyrgyz are the most widely used languages in Xinjiang, China. They all borrowed numbers of Alphabetic letters from Arabic and Persian and therefore belong to Arabic Script System.

In 1996, as the visiting scholar, author implemented the elementary research on printed Uyghur character recognition, in the THOCR lab of professor Ding Xiao-Qing, Tsinghua University. The research included the segmentation, recognition and specimen page training, built the basic framework and established the foundation to the later research on the

维文文本,识别率为 90%以上.

### 参 考 文 献

- Zhao Bo-Zhang, Zhang Song-Zhi. Chinese Information Processing. Beijing: China Astronautics Publishing House, 1990 (in Chinese)  
(赵泊璋, 张松芝. 中文信息处理技术. 北京: 中国宇航出版社, 1990)
- Wu You-Shou, Ding Xiao-Qing. Chinese Character Recognition Mode and Practice. Beijing: Higher Education Press, 1993 (in Chinese)  
(吴佑寿, 丁晓青. 汉字识别原理方法与实践. 北京: 高等教育出版社, 1993)
- Sherif Sami El-dabi, Refat Ramsis, Alandin Kamel. Arabic character recognition system: A statistical approach for recognizing cursive typewriting text. Pattern Recognition, 1990, 26 (5): 485~495
- Abuhaiba I. S. I., Ahmed P.. Restoration of temporal information offline Arabic handwriting. Pattern Recognition, 1993, 26 (7): 1009~1017
- Bian Zhao-Qi. Pattern Recognition. Beijing: Tsinghua University Press, 1993 (in Chinese)  
(边肇祺. 模式识别. 北京: 清华大学出版社, 1993)

**Aziguli**, born in 1969, associate professor. Her research interests include data-mining, pattern recognition.

Uyghur Character recognition. In 1998, this research project received the support of Xinjiang Oil field Company, CNPC. In 2000, "Multi Font Uyghur Character Recognition System" passed the check and accept.

In May 2002, with the request of professor Ding Xiao-Qing, Department of Electronic Engineering, Tsinghua University, author attended the Uyghur language study and development section of the "China Minority Character Recognition Research".