基于模板匹配的印刷维吾尔文字符识别研究

陈 卿[1],袁保社[1],李 晓[1],任宏宇[2],张建华[3]
(1.新疆大学 信息科学与工程学院,新疆 乌鲁木齐 830046;
2.94537部队气象台,河南 许昌 461101;
3.新疆公众信息产业股份有限公司,新疆 乌鲁木齐 830046)

摘 要:维吾尔文字的连笔书写及字型变化的一些特征给识别带来一定的困难并会影响到识别的正确率。在分析了维吾尔文单词的组词规律及其字型结构特征基础上,采用一种基于区域分割模板匹配的识别方法,通过建立标准维吾尔文字母图像模板库,并与通过预处理所获得的待识别维吾尔文字母图像进行匹配。对一些相似度高且难区分的维吾尔文字母则采用提取这些相似字符的附属笔画部分的图像并对其按笔画的连通性、交叉性以及形态等特征进行附属笔画判定的方法来确定这些相似字符,从而较准确地实现了对维吾尔文印刷字符的识别。实验识别率达到94%。

关键词:印刷体;维吾尔文字;特征提取;模板匹配;字符识别
中图分类号:TP391.43 **文献标识码**:A **文章编号**:1673-629X(2012)04-0119-04

Printed Uyghur Character Recognition Based on Template Matching

CHEN Qing[1], YUAN Bao-she[1], LI Xiao[1], REN Hong-yu[2], ZHANG Jian-hua[3]
(1. School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China;
2. Meteorological Observatory, 94537 Troops, Xuchang 461101, China;
3. Public Information Industry Co., Ltd. of Xinjiang, Urumqi 830046, China)

Abstract: The link writing and font variation in Uighur text bring about certain difficulties to Uighur identification. In the analysis of the rules how Uighur words are constructed and structural characteristics of the font, a standard character images template library can be established. And then match the collection of character images with the standard template. To those similar words among witch just have a slight difference to each other a secondary identification of the attachment image should be conducted to achieve a more accurate recognition of printing Uighur characters. Experiment recognition rate reaches to 94%.

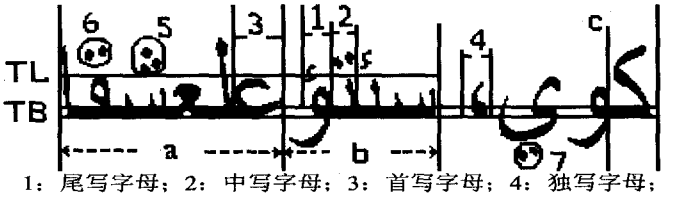
Key words: print; Uyghur character; feature extraction; template matching; character recognition

0 引 言

维吾尔文是中国主要少数民族文字之一,开展UOCR(Uyghur Optical Character Recognition)技术研究是中文多文种信息处理系统的迫切需要。目前国内成熟的印刷体汉字识别技术为UOCR的研究提供了很好的技术基础。

维文字母由主体笔画和附属笔画这两部分构成[1],维文字母一般具有独写、首写、中写和尾写这四种不同书写形式,且维吾尔文字的书写规则是从右至左连续书写的[2]。此外,每个字母的长度和高度都不一样,属于不等宽字体(不像汉字,汉字属于定宽字体)[3,4],这些都给维文识别带来很大的技术难度。

文中在研究了维吾尔文字符结构特征(如图1所示)基础上,提出了特征提取和模板匹配相结合的方法来进行识别,实验结果验证了其有效性。



1:尾写字母;2:中写字母;3:首写字母;4:独写字母;5、6和7均为附属笔画;TB:文字基线;a和b均为字母组成的连通体;c:两字母之间的连接处;

图1 维文字符结构特征

1 字符特征提取

输入维吾尔文字符图像经过预处理以及字母切分后得到若干印刷体维吾尔文的字母图像,提取字母的

收稿日期:2011-08-27;**修回日期**:2011-11-29
基金项目:工信部2009年度电子信息产业发展基金项目(工信部财[2009]453)
作者简介:陈 卿(1987-),男,硕士研究生,研究方向为中文信息处理;袁保社,教授,研究方向为中文信息处理。

宽高比、书写形式、附属笔画数以及环等特征信息可提高识别过程的准确率。

1.1 宽高比特征

在印刷体字符中，不同字符的宽度或者高度通常是不相等的，因此将待识别字母的宽高比作为特征之一对字母的识别具有一定价值。在提取特征时可以对每个字母对应的像素矩阵从上、下、左、右四个方向搜索来确定边界从而得到其高度 H_i 和宽度 W_i，宽高比 A_i 计算公式如下：

$$A_i = W_i/H_i, |A_i - MA_j| \leqslant T_m \quad (1)$$

其中 MA_j 表示模板字母图像的宽高比，T_m 为一个合适大小的阈值。若 A_i 与 MA_j 的差的绝对值大于 T_m，则判定待识别字母图像与模板字母图像宽高比不同，否则判定两字母图像的宽高比相同。

1.2 附属笔画特征

维吾尔文中很多字母都是由主体笔画和附属笔画（上附属和下附属）[5,6] 构成，有如‘ ’、‘ ’、‘ ’、‘ ’、‘ ’以及‘ ’等附属笔画。通过从字母图像基线行 BL 左端开始对所有基线行上的黑像素点进行其 8-邻域点的蔓延，将所有蔓延到的黑像素点（即主体笔画部分）置为白像素点，再结合 BL 的位置统计出基线以上部分的上附属笔画数和基线以下部分的下附属笔画数。

1.3 书写形式特征

维文字母一般具有独写、首写、中写和尾写四种书写形式，如‘ ’、‘ ’、‘ ’和‘ ’分别对应字符‘ ’的独写、首写、中写以及尾写形式。从上述几个字母可以看出字母的独写形式和尾写形式具有很高的相似度，当引入书写形式特征后很容易就可以将两者区分开来。数组 $R[n]$ 和 $L[n]$ 分别表示一行维文字符经切分后每个字母的左右边界所在列，Style $[n]$ 表示第 n 个字母的书写方式，对字母的书写形式的判断可遵循以下公式：

若 $n \neq 0$ 且 $R[n]$ 和 $L[n]$ 满足括号中条件时，

$$\text{Style}[n] = \begin{cases} 0 \ (\text{if } R[n] \neq L[n-1] \text{ and } R[n+1] \neq L[n]) \\ 1 \ (\text{if } R[n] = L[n-1] \text{ and } R[n+1] \neq L[n]) \\ 2 \ (\text{if } R[n] = L[n-1] \text{ and } R[n+1] = L[n]) \\ 3 \ (\text{if } R[n] \neq L[n-1] \text{ and } R[n+1] = L[n]) \end{cases} \quad (2)$$

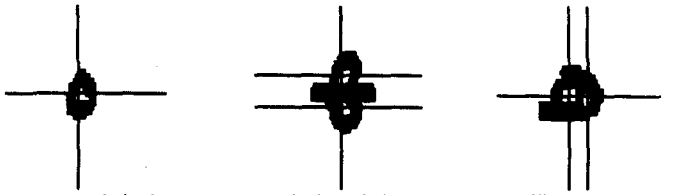
若 $n = 0$ 且 $R[n]$ 和 $L[n]$ 满足括号中条件时，

$$\text{Style}[n] = \begin{cases} 0 & (\text{if } R[n+1] \neq L[n]) \\ 1 & (\text{if } R[n+1] = L[n]) \end{cases} \quad (3)$$

其中 0 表示独写，1 表示首写，2 表示中写，3 表示尾写。

1.4 环结构特征

判断字母中是否具有环结构[7,8] 以及存在几个环能更好地识别带环字母，在维文字母中具有环结构的字母有‘ ’、‘ ’、‘ ’、‘ ’等。一次穿透从像素点矩阵角度可以看作是黑点→白点→黑点这一遍历过程，而黑点→白点→黑点→白点→黑点这一遍历过程则为二次穿透。



(a) 一次穿透　(b) 竖向二次穿透　(c)横向二次穿透

图 2 带环字母横向与纵向穿透示例

图 2 所示为三种环结构被穿透后的情况，假设穿透次数为 Z_{ij}（其中 $i = 1$ 和 2 时分别表示横向与纵向穿透，$j = 1$ 和 2 时分别表示一次穿透与二次穿透），若 $(Z_{11}, Z_{21}) = (1,1)$ 则说明该字母中存在一个环，若 $(Z_{11}, Z_{22}) = (2,1)$ 或者 $(Z_{12}, Z_{21}) = (1,2)$ 则说明该字母中存在两个环。

2 模板匹配

字符特征提取主要针对的是对字母的个体特征，根据提取的字符特征为每个字符建立模板库，在对未知字符进行识别时同样要用到这些特征。对待匹配图像与模板图像进行区域分割模板匹配[9]，以基线为边界将图像分割为上下两个区域再对这两个区域分别匹配识别。

假定待匹配字母图像的宽高分别为 Width 和 Height，将待匹配图像与模板图像归一化为 48×48 的二值点阵图像，图像像素点的像素值为 0 时表示白点而像素值为 1 时则表示黑点。在对字符图像进行归一化时，其基线位置会发生变化，即有在经过归一化的二值图像基线位置为 $BL_t = (BL \times 48) \div Height$。$D_k$ 和 $C_{m上/下}$ 分别表示两像素点是否匹配（0 表示不匹配，1 表示匹配）以及两图像（待匹配图像与模板图像）基线以上部分或基线以下部分的成功匹配的点的个数；C_m 表示两幅图像作整体匹配时匹配成功的点的个数；T_{n1}、T_{n2} 为一合适的阈值，其作用是防止因 C_m 满足最大但是局部区域匹配率 $C_{m上}$、$C_{m下}$ 达不到阈值要求而造成的误识别；max 用于记录待匹配字母图像与模板图像匹配率的最大值。D_k、$C_{m上/下}$、C_m 计算公式如下：

$$D_k = \begin{cases} 0 & (|f_{上/下}(i,j) - F_{上/下}(i,j)| = 1) \\ 1 & (|f_{上/下}(i,j) - F_{上/下}(i,j)| = 0) \end{cases} \quad (4)$$

$$C_{m上/下} = \sum_{k=0}^{宽\times高} D_k \quad (5)$$

$$C_m = \sum_{K=0}^{48\times48} D_k \tag{6}$$

式(4)中$f_{上}(i,j)$、$f_{下}(i,j)$、$F_{上}(i,j)$、$F_{下}(i,j)$分别表示从待匹配图像基线上下两部分和模板图像基线上下两部分各个像素点到其对应的像素值的映射。

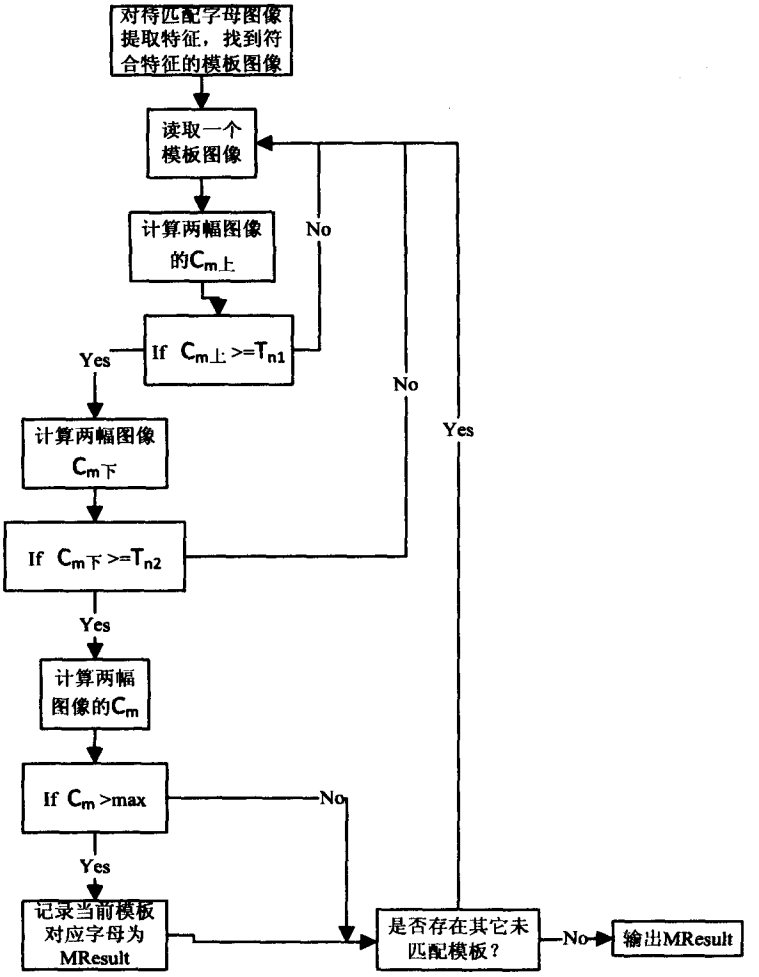


图 3　模板匹配流程

模板匹配过程如图 3 所示,首先提取待匹配字母图像特征,再根据提取的特征从标准模板库中筛选出符合这些特征的标准模板子集;获取该子集中的一个模板图像与待匹配图像进行匹配,若经计算待匹配图像与当前模板图像基线以上区域的匹配率 $C_{m上}$ 小于 T_{n1},则获取下一模板图像重新匹配,否则开始计算两图像基线以下区域的匹配率 $C_{m下}$;若 $C_{m下}$ 小于 T_{n2} 则同样获取下一模板图像重新匹配,否则计算两幅图像的整体匹配率 C_m;若 C_m 大于 max,则 max 记录其值并用 MResult 记录当前模板图像对应的维吾尔文字母并获取下一模板字母图像,重复上述步骤直至不存在未进行匹配的模板图像,此时的 max 即为待匹配图像与模板图像的最大匹配率,而 MResult 则记录了满足条件 $C_{m上}$ 不小于 T_{n1}、$C_{m下}$ 不小于 T_{n2} 并且整体匹配率 C_m 为 max 记录的最大值的模板图像所对应的维吾尔文字母[10,11]。通过这种方法可极大提高模板匹配过程中的运算效率且使得最终识别结果具有较高的准确性。

3　相似字鉴别

维文字符中通常会存在一些主体笔画相同、附属笔画相似的相似字子集,如图 4 所示。对于这类相似字符由于附属笔画的信号相对于主体笔画往往较弱且附属笔画之间的差异细微,尤其是当附属笔画受到噪声干扰时极易产生误识别的情况。相似字鉴别的目的就是通过区分相似字的附属笔画来区分相似字[12]。

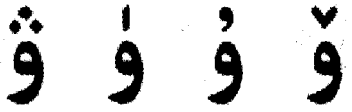


图 4　维吾尔文相似字示例

假设附属笔画的高度为 H_c,宽度为 W_c,$f(i,j)$ 是从附属笔画图像各个像素点到其对应的像素值映射,则对相似字符的附属笔画的识别则遵循以下几个规则:

(1)连通性:对于点状附属笔画如'ۥ'和'ۣ',利用连通域分析确定其点的个数及各点的相对位置。

(2)交叉性:对于具有交叉特征的附属笔画如'ۦ',假定图像中间行与中间列的位置分别为 $\frac{W_c}{2}$、$\frac{H_c}{2}$,从第 $\frac{W_c}{2}$ 列按行由上至下搜索满足条件

$$\begin{cases} f(i,\frac{W_c}{2}) = 0 \\ f(i,a) = 1 \text{ and } f(i,b) = 1 \end{cases}$$ 或

$$\begin{cases} f(i,\frac{W_c}{2}) = 1 \\ f(i,\frac{W_c}{2}-1) = 0 \text{ or } f(i,\frac{W_c}{2}+1) = 0 \\ f(i,c) = 1 \text{ and } f(i,d) = 1 \end{cases}$$ 的点,其中 $i \in [0,H_c-1]$,$a \in [0,\frac{W_c}{2})$,$b \in (\frac{W_c}{2},W_c-1]$,$c \in [0,\frac{W_c}{2}-1)$,$d \in (\frac{W_c}{2}+1,W_c-1]$,若附属笔画图像中至少存在一行满足上述条件则说明存在交叉。

(3)形状信息:在附属笔画图像的第 H_c、$\frac{H_c}{2}$、0 行分别计算其左右两端黑点的距离,通过分析这三个距离长短变化可对一些附属笔画进行判定。如附属笔画为'ۣ'和'ۦ',两者就具有明显的形态变化,其中附属笔画'ۣ'每行左右两端黑像素点的间距自上而下

呈现出递增的趋势，而附属笔画‘ ’则与之相反，这样，根据其形状信息就可以把‘ ’和‘ ’这两者区分开来。

(4)水平/垂直投影：在对诸如‘ ’和‘ ’这两个附属笔画进行投影后特征明显，如图 5(a)中所示‘ ’的水平投影时的行宽以及垂直投影时的列高变化细微或基本保持不变，而图 5(b)中所示‘ ’水平投影时的行宽以及垂直投影时的列高则出现剧减的情况。图 5(c)、5(d)则分别表示两个附属笔画的垂直投影。

(a) 水平投影 (b) 水平投影 (c)垂直投影 (d) 垂直投影

图 5 两种附属笔画的水平投影及垂直投影

4 结束语

文中的算法通过提取维吾尔文字母特征量，对待识别的字符进行带有冗余的分类(粗分类)，力求在保证分组准确的基础上尽可能缩小字符匹配范围，以提高识别的过程的运算速度。同时分组后的模板匹配算法从各组中识别出最后结果，保证了识别的正确率。经实验测得在含有 400 多个样本的测试集上的识别率达到 94%。实验结果表明两者的结合使系统具有较高的识别正确率与识别速度。

参考文献：

[1] 袁保社，吾守尔·斯拉木. 一种手写维吾尔文字母识别算法[J]. 计算机工程，2010，36(2)：186-188.

[2] 靳简明，王 华，丁晓青. 维汉英混排文档识别[J]. 电子与信息学报，2006，28(7)：1188-1191.

[3] 王 华，丁晓青，哈力木拉提. 多字体多字号印刷维吾尔文字符识别[J]. 清华大学学报(自然科学版)，2004，44(7)：946-949.

[4] 靳简明，丁晓青，彭良瑞，等. 印刷维吾尔文本切割[J]. 中文信息学报，2005，18(5)：76-83.

[5] Amin A，Mari J F. Machine recognition and correction of printed Arabic text[J]. IEEE Transactions on Systems，Man and Cybernetics，1989，19(5)：1300-1306.

[6] Al-Muallim H，Yamaguchi S. A method of recognition of Arabic cursive handwriting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，1987，9(5)：715-722.

[7] Olivier C，Miled H，Romeo-Pakker K，et al. Segmentation and coding of Arabic handwritten words [C]//Proceedings of 13th International Conference on Pattern Recognition. Vienna，Austria：[s. n.]，1996：264-268.

[8] Romeo-Pakker K，Miled H，Lecourtier Y. A new approach for Latin/Arabic character segmentation[C]//Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montréal，Canada：[s. n.]，1995：874-877.

[9] 吴伟伟，王小红，周亚南. 字符识别中两种改进的模板匹配算法[J]. 传感器世界，2008，14(6)：35-37.

[10] 崔 政，李 壮. 两种改进的模板匹配识别算法[J]. 计算机工程与设计，2006，27(6)：1083-1085.

[11] 张 晶，李志敏，黄 凡. 一种改进的自适应模板匹配法[J]. 微计算机信息，2008，24(9)：166-167.

[12] 哈力木拉提，阿孜古丽. 多字体印刷维吾尔文字符识别系统的研究与开发[J]. 计算机学报，2004，27(11)：1480-1484.

(上接第 118 页)

Rules[C]//Proceedings of 20th International Conference on Very Large Database. Santiago Chile：[s. n.]，1994：487-499.

[2] 范 明，孟小峰. 数据挖掘：概念与技术[M]. 北京：机械工业出版社，2001.

[3] 程舒通，徐从富. 关联规则挖掘技术研究进展[J]. 计算机应用研究，2009，26(9)：3210-3213.

[4] Savasere A，Omiecinski E，Navathe S. An Efficient Algorithm for Mining Association Rules in Large Databases[C]//Proc. 21st Int'l Conf. on Very Large Databases. San Francisco：Morgan Kaufmann，1995：432-444.

[5] Cheung D W. Maintenance of Discovered Association Rules in Large Database：An Incremental Updating Technique[C]//Proc. of 1996 Intl. Conf. on Data Engineering. [s. l.]：IEEE Computer Soc. Press，1996：106-114.

[6] Weiss A. Computing in Clouds[J]. ACM Networker，2007，11(4)：18-25.

[7] 刘 鹏. 云计算[M]. 北京：电子工业出版社，2010.

[8] Dean J，Ghemawat S. Mapreduce：simplified data processing on large clusters[C]//Proceedings of the 6th Symposium on Operating System Design and Implementation. San Francisco，California，USA：[s. n.]，2004：137-150.

[9] Rajaraman A，Ullman J D. Mining of Massive Data[M]. Stanford：[s. n.]，2010.

[10] Venner J. Pro Hadoop[M]. [s. l.]：Apress，2009.

[11] White T. Hadoop：The Definitive Guide[M]. [s. l.]：O'Reilly Media，Yahoo! Press，2009.

[12] Ghemawat S，Gobioff H，Leung S. The Google Filesystem [C]//Proc. of ACM Symposium on Operating Systems Principles. Lake George，NY：[s. n.]，2003：29-43.