文章编号：2095-2163(2020)05-0206-05 中图分类号：TP391.4 文献标志码：A

基于连体段的印刷维吾尔文特征提取

贾钰峰，章蓬伟，贾园园，邵小青，刘茂霞
（新疆科技学院 工商管理系，新疆 库尔勒 841300）

摘 要：典型的印刷识别系统由预处理，特征提取，训练模型，识别器组成。本文讨论了基于连体段的印刷维吾尔文特征提取方法。结合维文书写特点提取了：孔洞数，尾点，交叉点，方向码，笔划位置特征，前后景比值等。并利用以上特征进行了印刷维吾尔文的模型训练和识别。

关键词：印刷维吾尔文；连体段；特征提取

Printed Uighur feature extraction methods based on Word-part

JIA Yufeng, ZHANG Pengwei, JIA Yuanyuan, SHAO Xiaoqing, LIU Maoxia
(Xinjiang University of science and technology Department of business administration, Korla Xinjiang 841300, China)

【Abstract】 A typical printing recognition system consists of preprocessing, feature extraction, training model and recognizer. This article discusses the printed Uighur feature extraction methods based on word-part. Combining features of Uighur, feature extraction are as follows: hole number, end point, cross-point, orientation codes, relative position features of gestures, black and white pixel ratio etc. The above features are used to train and recognize the printed Uighur model.

【Key words】 printed Uighur; word-part; feature extraction

0 引 言

关于维吾尔文印刷识别方面，相关的研究文献资料较少。但维吾尔文与阿拉伯文很相似，参考了阿拉伯文及相关印刷识别方法[1-2]：典型的识别系统模块是由预处理、特征提取、训练模型、识别器组成的，如图 1。由维吾尔文的特点得知：印刷的文字切分不论以笔划，字母还是词，切分都是相当困难的[3-5]。同时还有图像文本躁点等因素影响，如：粘连，断裂，伪字母切分等。基于连体段(WordPart)[3]的段切分是一个很好的解决方案。它能够保留出整体的完备信息，从而为提取良好的特征做准备。本文在已经预处理的基础上，对连体段提取各类特征。

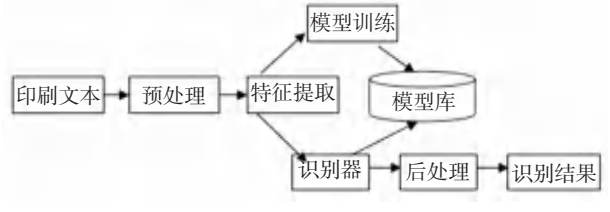


图 1 印刷维吾尔文识别系统框架

Fig. 1 Printed Uighur recognition system framework

1 维吾尔文的特点

维吾尔族的语言属于阿尔泰语系突厥语族。1938 年形成现行文字，有 32 个字母拼音自右向左横写，且有 120 多个字符形式，以词的形式表达[6]。其部分特点如下：

(1)维吾尔文字母包括主体部分和附加部分，其中 20 个字母有附加部分，它的形式为一“ ”、两“ ”、多“ ”、以及“ ”、“ ”等。附加部分在主体笔划的内部、下面或上面。如： 、 、 、 、 、 、等。

(2)维吾尔文的单词是由多个或者单个字母组合而成。根据书写规则这些字母的组合形成一个或几个前后相连接的音节或称连体段。连体段中的字母，在印刷的时候总是沿着水平线的，这条水平线被称为基线。维吾尔文的结构详情见图 2。

2 特征提取

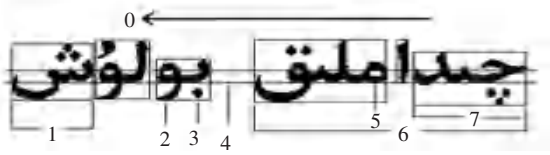
特征提取的过程就是将图像文本映射到文字独有的特征空间，以便压缩信息量，方便后续的分类、训练和识别。本文根据维吾尔文的特点，基于连体段提取出一系列常用特征，并最终建立一个特征库

基金项目：自治州科学技术基金项目(2019018)。

作者简介：贾钰峰(1986-)，男，硕士，助教，主要研究方向：图像处理与模式识别、大数据；章蓬伟(1985-)，男，硕士，助教，主要研究方向：人工智能与模式识别、大数据；邵小青(1988-)，女，硕士，助教，主要研究方向：模式识别、自动化；贾园园(1988-)，女，硕士，助教，主要研究方向：电子商务、人工智能；刘茂霞(1991-)，女，本科生，主要研究方向：电子商务。

收稿日期：2020-01-07

以备使用，现将方法介绍如下：



0.六个连体段组成两个词，方向自右向左；1.一个字母组成连体段；2.主笔划；3.主笔划下方的副笔划；4.基线和基线的上下限；5.环；6.三个连体段组成的词；7. 三个字母组成的连体段

图 2　维吾尔文结构特点说明

Fig. 2　Structural features of Uighur

2.1　笔划位置特征

维吾尔文中连体段由主体部分和附加部分组成。由图 2 可知附加笔划在主笔划的下面、上面或内部。笔划位置特征就是求出主笔划，副笔划个数，副笔划与主笔划的相对位置关系。副笔划的位置特征可以通过与基线的相对位置来识别。但有些特殊情况无法完美判断，如：ي与ي。故介绍一种基于联通域的更细致的判别方法：

选定一点作为一个种子，由图像的区域生长法[7-8]可以求得一个相同像素相互联通的区域（即笔划）。因此根据种子点就可以知道联通域，根据联通域就知道连体段笔划的数量。根据图像像素点读取的顺序，位图行内是由左往右，行间由下往上。将二值图像扫描时第一个碰到的黑色像素作为种子，再由区域生长法生长蔓延出一个笔划区域，得到一个笔划。为了防止重复计算笔划，将已经蔓延过的笔划反色（使笔划和背景同色）后继续循环寻找下一个种子点，继续生长蔓延找到下一个笔划，直至图循环完毕。该算法主要步骤：

（1）为了不破坏原图，复制初始化为连体段图像的缓冲区两个（设为 A 和 B）。

（2）对 A 图像域按由左往右，由下往上的方式依次扫描像素点。碰到黑色像素点就作为种子点记录下来。

（3）由区域生长法得出此联通域。在区域生长同时统计黑点的个数并记录黑点的坐标（黑点总个数为连体段的面积；坐标的平均值可以用来判断笔划的位置关系）。

（4）根据记录的黑点坐标，把 B 图像域相同坐标处像素值反色（白色）。此时 B 域就是除去联通域剩下的部分。

（5）将 B 域的像素值覆盖掉 A 域。

（6）重复步骤（2），（3），（4），直至循环完毕。

由此算法可统计出连体段的联通域个数，即笔划数；其中面积最大的为主笔划。其它笔划的平均纵坐标与主笔划比较，就可以判断出副笔划的上下位置特征。同理副笔划之间的平均横坐标可以判断出副笔划之间的左右位置特征，而这些特征都是全局精确特征，可以显著提高识别分类效果。连体段笔划位置特征提取步骤如图 3 所示。

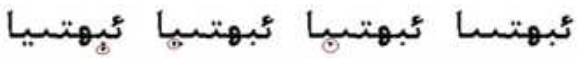


图 3　连体段笔划位置特征提取步骤

Fig. 3　Stroke position feature extraction steps of wordpart

其特征结果如下：

（1）笔划数为 8 个。面积分别为：15，16，16，815，14，82，14，16。最大面积即主笔划面积为：815。

（2）位置特征由下往上依次为：-16，-11，-11，0，-11，13，13，15。其中 0 代表主笔划；小于 0 的共 4 个，代表在主笔划下方；大于 0 的共 3 个，代表在主笔划上方。

（3）同时还可以知道副笔画之间的位置关系。

2.2　孔洞数

连体段孔洞数也是基于联通域个数的[7]。为保证孔洞数目准确性，将图像黑白像素调换后继续递归循环算法对不是笔划的区域进行联通域个数计算，如图 4 所示，可得到总的不是笔划的连通区域个数，减去和边界相交的背景区域，就可以得到连体段孔洞数数量这一特征。这也是全局精确特征。此特征在后续使用中发现非常高效。

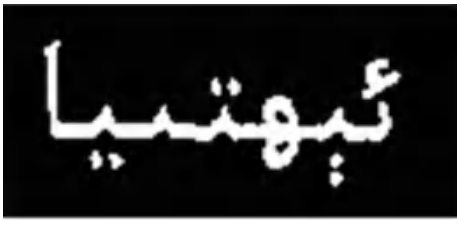


图 4　黑色联通域共有 3 块 孔洞数共有 2 个

Fig. 4　There are 3 holes in the black connecting area and 2 holes in total figure

2.3　方向码

方向码特征是字符识别中非常经典有效的特征，方向码就是把平面分成八个方向[9-10]，如图 5 所示。取出主笔划的轮廓，在轮廓上选取一个开始点和一个结束点作为一个线段，并判断这个线段归属于那个方向，然后用方向序号标识。继续重复直至循环轮廓一周后得到一组方向特征向量，这组向量就是主笔划的方向码。它具有较高稳的定性和抗干扰能力，代表主笔划的形状特征。方向码特征提取过程为：求连体段轮廓，点聚类，直线逼近，得出方向码。

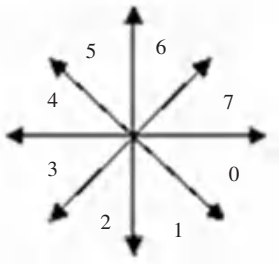


图 5 八方向示意图

Fig. 5 Eight direction schematic diagram

(1)取轮廓。取轮廓使用边界跟踪法,从图 6 得到主笔划,从主笔划选取种子点,顺时针环绕主笔划边界一周后得到连体段的轮廓[7-11]。

图 6 连体段原始图像

Fig. 6 Original image of wordpart

主笔划轮廓选取算法如下:

定义初始搜索方向为左上方,搜索到的第一个黑色像素作为轮廓种子点,找到下一个黑色像素,记录边界点坐标和个数。找不到就顺时针旋转 45°继续寻找直至找到下一个点。重复以上方法直到返回最初起始种子点为止。轮廓像素点个数就是周长特征,坐标点就是主笔划的离散形式。用此方法取得的轮廓如图 7 所示。

图 7 取出轮廓的连体段

Fig. 7 Take out the wordpart of the contour

(2)点聚类。如隔几个像素点取一个点,使轮廓点稀疏。但点聚类依然不够简化。

(3)直线逼近。在尽量保留拐点、关键点的情况下,仍对轮廓进行进一步的简化采样,将轮廓变为保持原有形状的折线,如图 8 所示。

图 8 直线逼近后的连体段

Fig. 8 The wordpart after the straight line approximation

在这里我选用了经典的 Douglas-Peucker 算法[12]实现直线逼近功能,其算法描述如如图 9 所示。

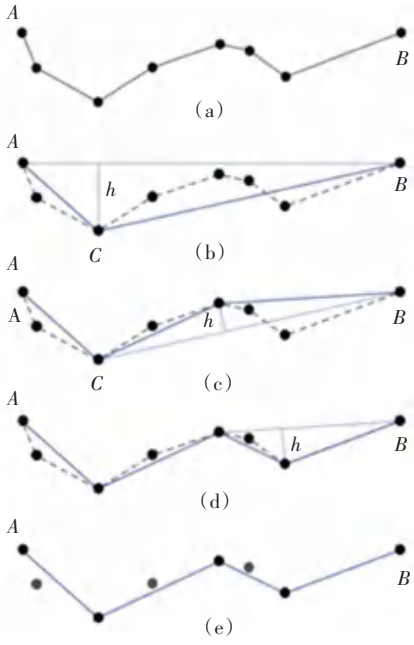


图 9 Douglas-Peucker 算法描述过程

Fig. 9 Douglas Peucker algorithm description process

①在曲线 AB 之间连接一条线段 AB,线段 AB 为曲线的弦, 如图 9(b)所示。

②求出曲线上离线段 AB 距离最远的点 C,计算其与线段 AB 的距离 h。

③若 h 小于预定阈值,则该直线段 AB 作为曲线的近似,该段曲线 AB 处理完毕。

④若 h 大于预定阈值则,利用距离最远的 C 点将曲线 AB 分为两段 AC 和 BC,线段 AC 和 BC 重复步骤 1 到 3, 直至处理完毕,如图 9(c)和图 9(d)。

⑤依次连接各个最终确定的逼近点作为连体段的近似轮廓,完成简化采样,如图 9(e)。

(4)得出方向码。根据笔划骨架中每一段线段归属的方向得出方向码特征向量。

图 8 的方向码特征向量为(假设阈值为 3,具体根据字体大小判断):4,5,5,6,1,0,0,0,7,5,5,6,0,1,0,6,1,7,5,6,0,1,2,3,3,4,2,3,3。若将每个方向的个数求和,则方向码向量降维为:6,4,1,3,5,4,2。

2.4 尾点、交叉点

尾点与交交叉点作为细化后字符识别提取的常用特征已被广泛使用[13]。前人已经研究了各种经典的细化算法,经过论证,本文采用细化算法提取此特征,得到了较好的效果[7]。在细化后的连体段图像中,交叉点一般都是在当前点的 8 邻域模板中,即以当前点 p 为中心选择一个 3×3 的模板来判别其尾点、交叉特征属性。计算 P 点的交点数的公式(1):

$$NUM = \sum_{i=0}^{6} |P_{i+1} - P_i| + |P_0 - P_7|, SUM = \sum_{i=0}^{7} |p_i|. \quad (1)$$

模板如图 10 所示。

P_3	P_2	P_1
P_4	P	P_0
P_5	P_6	P_7

图 10　8 邻域模板

Fig. 10　8 Neighborhood templates

其中，*SUM* 表示当前点 8 领域中像素点的个数。*NUM* 表示当前点 *P* 的 8 领域模板中像素值的 0,1 变化次数。由算法可以判定出：

（1）当 *NUM* = 2，*SUM* = 1 时，*P* 为尾点；

（2）当 *SUM* > 3 时为交叉点。其中，*NUM* = 6，*SUM* = 3 时，*P* 为三叉点；*NUM* = 8，*SUM* = 4 时，*P* 为四叉点。

图 11 为图 3 的主笔划细化后的情况，共有尾点、三叉以上点各 6 个。

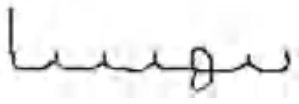


图 11　细化后的主笔划

Fig. 11　Main stroke after thinning

2.5　前后景比值

前后景比值特征是基于统计，计算出连体段矩形框中黑像素和与白像素的比值[14]。这样提取特征的优点在于计算简单，不受字符大小的影响，只要字形固定，对于印刷体就可以作为特征，如图 12 的前后景比值为 0.25。另外，印刷体维吾尔文连体段具有一定的高度和宽度，计算连体段高宽比特征值，如图 12 宽为 60，高为 56，宽高比：1.071。前后景色比值结合连体段宽高比，可以作为联合特征，针对部分连体段可获得较高的稳定性和抗干扰性。

图 12　连体段的像素模型

Fig. 12　Pixel model of wordpart

3　结束语

本文重点表述了维吾尔文连体段常用特征提取方法。在分析维吾尔文书写特点的基础上提取了如下特征：宽高比，前后景比值，孔洞数，尾点，交叉点，方向码，笔划位置特征。基于以上方法通过对四十多对样张进行批处理，建立起了基于连体段的特征库图 14 为对图 13 从右往左提取连体段相关特征的截图。

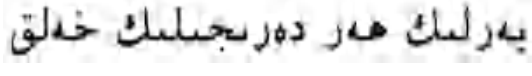
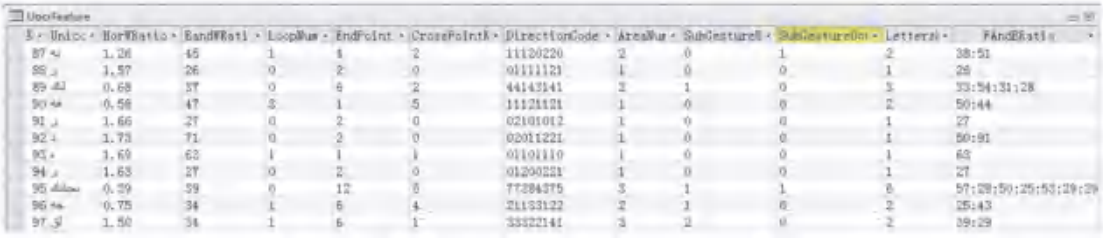


图 13　原始样张部分截取

Fig. 13　Part of the original sample



	Unico	HorWRatio	BandWRati	LoopNum	EndPoint	CrossPointN	DirectionCode	AreaNum	SubGesture0	SubGesture0n	Letters	FAndERatio
87		1.26	45	1	4	2	11120220	2	0	1	2	38:51
88		1.57	26	0	2	0	01111121	1	0	0	1	26
89		0.68	37	0	6	2	44143141	2	1	0	3	33:54:31:28
90		0.58	47	3	1	5	11121121	1	0	0	2	50:44
91		1.66	27	0	2	0	02101012	1	0	0	1	27
92		1.73	71	0	2	0	02011221	1	0	0	1	50:91
93		1.69	63	1	1	1	01101110	1	0	0	1	63
94		1.63	27	0	2	0	01200221	1	0	0	1	27
95		0.39	39	0	12	8	77284375	3	1	1	6	57:28:50:25:53:29:29
96		0.75	34	1	6	4	21133122	2	1	0	2	25:43
97		1.50	34	1	6	1	33322141	3	2	0	2	39:29

图 14　特征库截图

Fig. 14　Screenshot of feature library

图 14 为特征库的截图：从左往右每一列的特征分别为：连体段 Unicode 码，高宽比，孔洞数，尾点数，交叉点数，压缩后的 8 方向码，笔划数，基线上方副笔划数，基线下方副笔划数，连体段字母个数，前后景比值等。

将特征应用于维吾尔文字印刷识别系统，由识别器识别后证明所采用的算法思想和特征维度是有效的。但还需要进一步完备连体段的有效特征，以便抽取出那些对不同类别最为重要的特征，组合成良好且优秀的特征组合。后续还需要研究局部特征和全局特征合并训练的方法，完备字符特征集合，改进特征选取准则函数，结合分类器的改进，尽可能的提高识别率[15]。

（下转第 212 页）