•人工智能及识别技术•　　**文章编号：1000—3428(2012)18—0147—04**　　**文献标识码：A**　　**中图分类号：TP391**

维吾尔文手写体联机识别中延迟笔画处理方法

皮桂林，赵　晖

(新疆大学信息科学与工程学院，乌鲁木齐 830046)

摘　要： 延迟笔画处理是维吾尔文联机手写体单词识别中的难点，样本数据格式为一串具有时序性的序列，造成字母主笔画和延迟笔画在序列位置上多半不连续，很大程度上增加了延迟笔画处理和单词识别的难度。为此，提出一种维吾尔文延迟笔画处理方法，其中包括延迟笔画查找与投影，实验结果证明该方法具有良好的效果，单词识别准确率达到 93.71%。

关键词： 维吾尔文；延迟笔画；手写体识别；联机识别；隐马尔可夫模型

Delay Stroke Processing Method in Online Uyghur Handwriting Recognition

PI Gui-lin, ZHAO Hui

(College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

【Abstract】 Processing delay stroke is a difficulty in online Uyghur handwriting recognition. Because sample data format is a timing sequence, it makes letter's main stroke and delay strokes uncontinuous in sequence. It greatly increases the difficulty of processing delay stokes and recognizing handwriting word. To solve this uncontinuous problem, this paper proposes a delay stroke processing method which contains seeking and projection of delay strokes. Result shows this method is robust and recognition accuracy rate is up to 93.25%.

【Key words】 Uyghur; delay stroke; handwriting recognition; online recognition; Hidden Markov Model(HMM)

DOI: 10.3969/j.issn.1000-3428.2012.18.040

1　概述

随着电子技术的飞速发展，人与电子产品交互对象越来越多，如手机、手写电脑。维吾尔语属阿尔泰语系突厥语族，是维吾尔族日常交际工具，新疆其他少数民族群众也有通晓维吾尔语的，如乌兹别克族，实现维吾尔文联机手写识别是满足新疆少数民族地区日益增长的物质文化需求的需要。

现代维吾尔文以阿拉伯文字为基础，共 32 个字母，每个字母有 2 种或 4 种书写形式[1]，字母笔画由主笔画和延迟笔画两部分组成，其中主笔画不可缺少，维吾尔文中还包含一个一个元音前置符。维吾尔文书写时沿基线从右往左书写，先写主笔画，再写延迟笔画，连体段一般作为整体书写，例如单词“ياخشى”，字母书写顺序依次为“يـ”、“ا”、“خـ”、“ش”，其中，点为延迟笔画，“يا”和“خشى”是连体段。

手写体识别根据获取数据来源不同分为脱机识别和联机识别[2]，联机识别的数据是坐标点序列，通过电子设备采集的笔尖先后经过的坐标点，每个点的数据由三元组(x,y,z)构成，x、y、z 分别表示点的横坐标、纵坐标、是否是笔画起始点。由于坐标点采集顺序和连体段作为整体书写的特点，因此字母的主笔画与延迟笔画在点序列中不连续，很大程度上增加了延迟笔画处理和手写单词识别的难度。

从目前研究现状来看，维吾尔文手写体识别中还没有很好的延迟笔画处理方法，延迟笔画的处理一直以来都是维吾尔文联机手写体识别的难点。因为现代维吾尔文与阿拉伯文很相似，所以在处理维吾尔文的延迟笔画时，可以参考阿拉伯文的处理方法。

文献[3]分析了阿拉伯文中主笔画和延迟笔画的特点，所给出的区分方法有限，文献[4]提出通过使用投影方法来处理延迟笔画，但是该方法不能直接应用到维吾尔文上。为此，本文结合阿拉伯文来分析维吾尔文中主笔画和延迟笔画书写的特点，在数据点序列中标记主笔画和延迟笔画对应的序列，找出延迟笔画在主笔画中最佳对应位置，改进文献[4]中的方法，把延迟笔画序列投影到最佳对应位置，解决字母主笔画与其延迟笔画序列不连续问题，从而为维吾尔文单词的识别打下良好的基础。

2　维吾尔文笔画特点

联机手写维吾尔文数据有两方面特性：图像性和时序性，其中，前者体现在样本点序列所代表的二维几何空间上，后者则体现在点被采集的先后顺序上。为了找出哪些

基金项目： 国家自然科学基金资助项目(60962005)；新疆维吾尔自治区自然科学基金资助项目(200821124)

作者简介： 皮桂林(1987－)，男，硕士研究生，主研方向：模式识别；赵　晖(通讯作者)，教授、博士

收稿日期： 2011-11-14　　**修回日期：** 2012-01-11　　**E-mail：** zhaohui@xju.edu.cn

是主笔画，哪些是延迟笔画，可以从这2个特性进行分析。

维吾尔文中绝大部分字母都有延迟笔画，很多字母与其他字母的差别主要通过延迟笔画的不同来区分，如表1所示，所列出的每组字母的差异主要体现在延迟笔画的笔画数、形状及与主笔画的相对位置等特征上，其中，延迟笔画种类主要包括“.”、“s”、“ﻋ ء”、“ˇ”、“,”、“ˈ”、“/”等。图1是一个典型的维吾尔文单词，从图像上看，共有10个延迟笔画(方框中)，单词中每个连体段的主笔画与基线相交，比较高，外接矩形面积也很大；而延迟笔画位于基线上下，其几何重心与基线距离较远，没有环，宽高都较小。

表1 书写相似字母表

相似字母组号	独写形式	首写形式	中写形式	尾写形式
1	ن ت پ ب	ﻳ □ ﭘ ﻧ ﺗ ﭘ ﺑ ﺋ	ﺋ ﻳ □ ﭘ ﻧ ﺗ ﭘ ﺑ	ن ت پ ب
2	چ خ ج	ﭼ ﺧ ﺟ	ﭼ ﺧ ﺟ	غ خ چ ج
3	قفۋۈۆۇو	ﻗ ﻓ	ﻗ ﻓ ﻏ	ف ۋ ۈ ۆ ۇ و ق
4	ڭ ك ل	ﯖ ﮔ ﻛ	ﯖ ﮔ ﻛ	ڭ ك ل
5	ش س	ﺷ ﺳ	ﺷ ﺳ	ش س
6	ې ي ى			ي ى ې
7	ژ ز ر			ژ ز ر

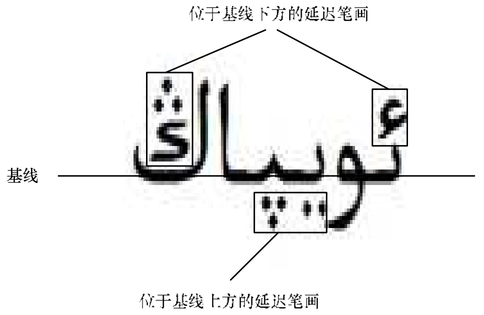


图1 维吾尔文单词

另外，在不同字母进行组合时，主笔画的形状根据字母在单词中的位置不同选择不同的形状，而延迟笔画基本上不受组合影响改变形状。换句话说，延迟笔画能在不同单词中保持固定的书写形状，主笔画则可能随着相邻字母的不同其形状也发生改变。

由于电子设备采集的数据是根据时间顺序来记录的，因此对于单词的各个部位来说，它们的出现也是有先后顺序的。按照维吾尔文的书写习惯，先写主笔画，再写延迟笔画，如果是连体段，则先连续书写完连体段中各个字母的主笔画，再写各个字母的延迟笔画，延迟笔画与相应的主笔画在序列中位置不连续。对于采样率一定的设备，写延迟笔画时笔尖经过的距离很短，如果书写速度均匀，则设备收集的点数也很少，而主笔画则相反，因此，在点序列中，主笔画的序列长度即采样点的数目一般都比延迟笔画多很多，特别是与“.”、“,”和“ˈ”这类短笔画的延迟笔画相比。

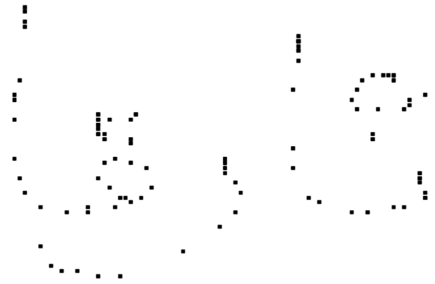
综上所述，可以得到如表2所示的主笔画和延迟笔画的特点。

表2 主笔画与延迟笔画特点比较

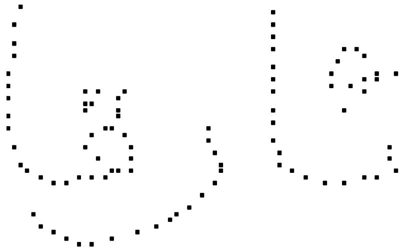
特点	主笔画	延迟笔画
笔画点序列点数	多	较少
是否被基线穿过	是	否
外接矩形面积	较大	较小
高度	较高	较低
重心离基线距离	较近	较远
是否有环	不确定	没有
宽高比	不确定	一般为0.25~4

3 延迟笔画查找

处理延迟笔画首先要区分哪些笔画序列属于主笔画，哪些属于延迟笔画，并根据延迟笔画查找结果标记相应点序列的归属。在查找延迟笔画之前，需对样本进行预处理[5-6]，以减少电子设备采集数据引入的误差，如高度不一致、采样点稀疏，对样本进行预处理的步骤依次为去孤立点、去重复点、坐标平移、线性归一化、插点、重采样和平滑，图2是预处理前后的对比效果。



(a)预处理前



(b)预处理后

图2 延迟笔画预处理前后对比

维吾尔文书写随意，但都按沿基线从右往左书写的原则，其中，延迟笔画分别位于基线的上方和下方，为了准确查找延迟笔画，根据该特点，需先查找单词的基线。在收集的样本点序列中每个点的信息由三元组(x,y,z)表示，x和y分别表示横坐标和纵坐标且都是整数，z记录该点是否是笔画刚书写的位置。在进行预处理之后，单词笔画的上边界和下边界分别用$y=y_{\max}$和$y=1$表示，函数$h=g(y)$表示点序列中纵坐标为y的点的总数。由于基线位置与经过样本中点最多的直线位置相同，因此$h=g(y)$取得最大值时对应的纵坐标为基线所在位置。通过计算区间$[1,y_{\max}]$中所有整数对应的函数$h=g(y)$的值，可以得到基线位置。

维吾尔文单词笔画只有主笔画和延迟笔画2种，每一笔要么是主笔画，要么是延迟笔画，因此，根据表2中列

出的特点，设置相关的优先顺序和阈值来判断笔画的归属。在判断是否满足某个特点时，使用设定阈值的方法，阈值根据所有样本的适用情况选取，图 3 为延迟笔画查找结果，方框中的笔画为延迟笔画，其他为主笔画。

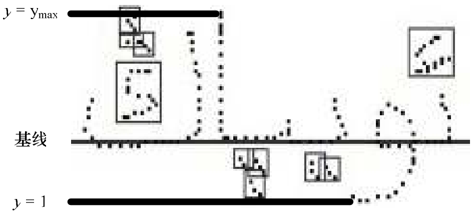


图 3　延迟笔画的查找

对于查找到的延迟笔画，除了标记它是延迟笔画之外，还需记录它与基线的相对位置，即位于基线的上方还是下方，这对后面的投影过程有很大帮助。判断标准为：如果延迟笔画重心在基线上方，则延迟笔画在基线上方，否则在下方，因为极个别延迟笔画因书写习惯不同与基线有交点。

4　延迟笔画处理

文献[4]对延迟笔画处理的方法是使用投影，即在主笔画中找到延迟笔画与相应主笔画在垂直方向上的最佳对应位置，然后把延迟笔画和添加的虚拟点序列一起插入该最佳位置。但是该方法应用在维吾尔文上存在很多问题，其延迟笔画查找方法不够准确，用延迟笔画起点向主笔画引垂线，使最佳位置往往与主笔画有偏离，在没有位置和有多个位置可选择时，处理方式也不好。针对以上问题，本文在维吾尔文中先使用延迟笔画重心向主笔画引垂线，再判断垂线与主笔画相交的交点数目，通过该数目确定投影最佳位置，规则如下：

(1)没有交点时。当延迟笔画在基线上方时，最佳位置选择在延迟笔画左下方且离其重心最近的点。当延迟笔画在基线下方时，最佳位置选择在延迟笔画左上方且离其重心最近的点。

(2)只有一个交点时。直接选择该交点作为最佳位置。

(3)至少有 2 个交点时。当延迟笔画在基线上方时，最佳位置位于交点中离延迟笔画重心最近的点。当延迟笔画在基线下方时，最佳位置位于交点中离基线距离最近的点。

找出最佳位置关系之后，进行投影所需数据及其对应关系如图 4 所示，此时，可以把延迟笔画序列移动到最佳位置之后，与对应的字母主笔画序列放在一起。为了使每个延迟笔画序列放入对应主笔画序列中与该字母的其他延迟笔画分开，采用在延迟笔画的首尾插入虚拟点的方法，其插入方式为通过起点和终点向最佳位置引直线段，按照重采样时两点间距离的阈值在该直线段上取点。投影后笔画序列表示的图像如图 5 所示，其中，箭头方向所代表的是各个点在样本序列中依次出现的顺序；数字表示箭头先后顺序。经过投影之后，每个字母的主笔画和延迟笔画都在一段连续的笔画序列中，从而解决了笔画不连续的问题，对维吾尔文的联机手写识别模型建立有很大帮助。

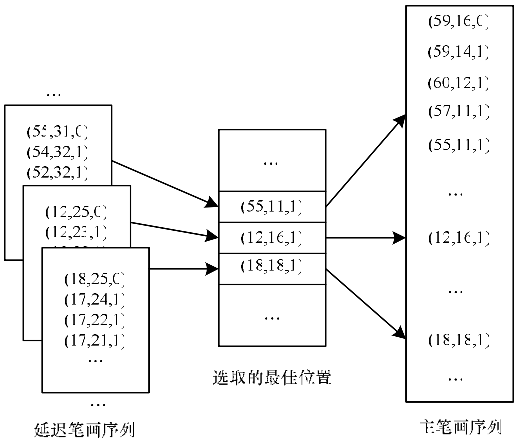


图 4　投影所需数据对应关系

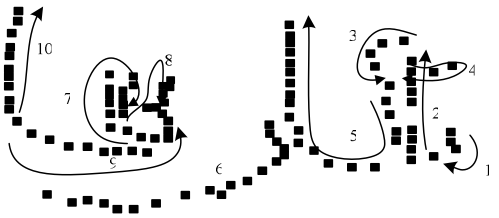


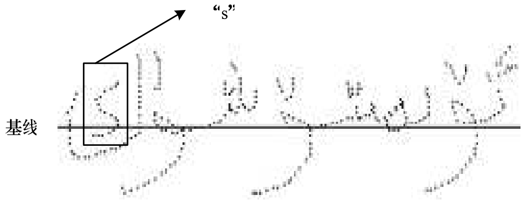
图 5　投影后的点序列顺序

5　实验与结果分析

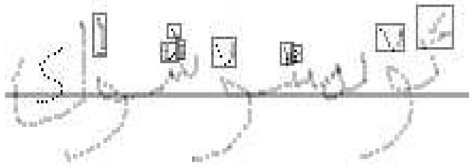
收集维吾尔文单词样本的工具是数码笔和数码本，使用者用数码笔在数码本上书写单词，数码笔则把笔尖经过数码本上的点坐标依次记录下来。共收集 110 个人手写的 9 个或 10 个不同单词，书写时要求连体段一笔书写和延迟笔画不简写，共收集 1 088 个单词样本，去除书写错误的单词，如缺少延迟笔画、字母丢失、笔误，剩余 1 066 个。这些单词中共包含 6 184 个延迟笔画，通过延迟笔画查找，共查找出 6 177 个延迟笔画，查找准确率为 99.8%。对查找出的延迟笔画进行投影，投影正确的延迟笔画数为 6 114，准确率为 98.9%。

在查找延迟笔画的过程中，带点的延迟笔画序列长度很短，一般不超过 6 个，很容易进行判断。另外像“ء”、“ˇ”、“ʾ”、“ˈ”和“ˊ”这样的延迟笔画，它们都位于基线上方，不仅其重心离基线很远，而且其最低点也离基线很远，判断也相对容易。但是延迟笔画“s”的判断则比较麻烦，因为它满足主笔画的绝大多数特征，很容易被当做主笔画，另外，由于人们书写习惯的差异，它会被书写得靠近基线甚至与基线相交，如图 6(a)所示的单词，进行延迟笔画查找之后得到图 6(b)结果，延迟笔画“s”未被判断出来。为避免该错误判断，需引入与基线无关的几何特征，由于“s”只出现在字母“ك”和“ڭ”中，其位置在主笔画“ل”外接矩形的内部，因此通过该特征可以把它从主笔画中分离出来，如图 6(c)所示。在对延迟笔画进行处理的过程中，投影位置的选择非常重要，直接关系到结果的好坏。位于连体段中间的字母，对其延迟笔画进行

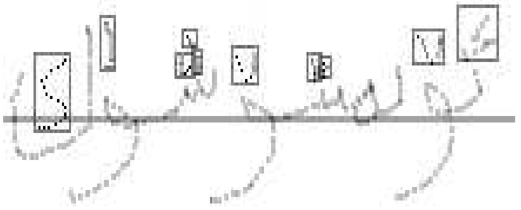
投影都比较准确，因为它们能准确查找最佳位置进行投影。然而位于连体段的第1个字母，由于受前一个字母影响，因此位置查找比较复杂，如图7(a)所示的延迟笔画1在右边先书写的字母上找到最佳位置，图7(b)中延迟笔画2和3也受右边字母影响有2个投影位置供选择。针对前者出现的情况，延迟笔画在基线上方，可设置查询位置纵坐标的下限；后者延迟笔画在基线下方，最佳位置选择离基线最近的点。



(a)书写中的延迟笔画“s”

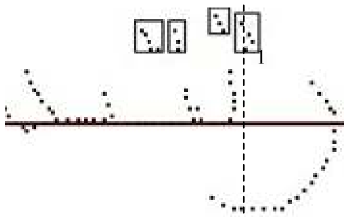


(b)使用在主笔画外接矩形内部特征前

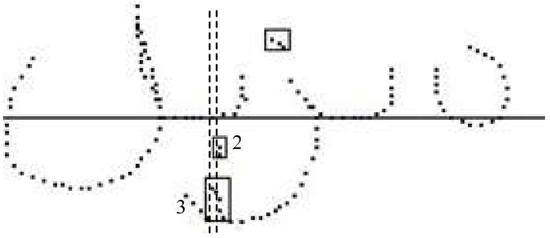


(c)使用在主笔画外接矩形内部特征后

图6 延迟笔画“s”的查找



(a)基线上方受影响的延迟笔画



(b)基线下方受影响的延迟笔画

图7 投影位置受影响的延迟笔画

通过使用这种延迟笔画处理方法，使字母的主笔画和延迟笔画序列出现在同一段序列中，处理后的序列仍然具有时序性，从而很方便地为字母的每种书写形式建立HMM模型[7]，再利用决策树来识别单词。对收集的样本提取环特征和方向特征[8]，训练每个字母书写形式的HMM模型，用包含2 000词汇量的决策树识别1 058个样本单词，分别输出1个~10个最佳识别结果(NBest)，识别准确率变化情况如图8所示。图8中阿拉伯文处理方法是使用处理阿拉伯文方式处理维吾尔文的延迟笔画，维吾尔文处理方法使用本文的处理方式，实验表明使用本文延迟笔画处理方法比直接使用阿拉伯文的处理方法能得到更高的识别准确率，但是由于收集的样本不足以体现所有人书写维吾尔文单词的特点，因此该方法还需进一步完善，如基线查找和预处理。

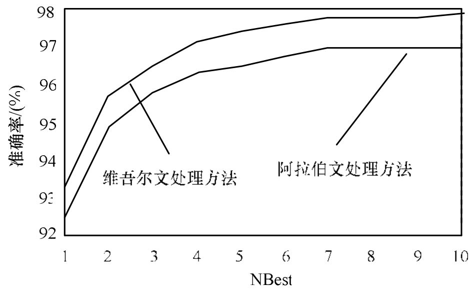


图8 识别准确率变化曲线

6 结束语

在维吾尔文单词的联机手写体识别中，怎样处理延迟笔画是一个非常重要的问题，本文提出一种维吾尔文手写体延迟笔画处理，解决了单个字母主笔画和延迟笔画在单词序列中不连续的问题，为单个字母建立模型打下良好的基础。

参考文献

[1] 国家质量监督局. GB21669-2008 信息技术维吾尔文、哈萨克文、柯尔克孜文编码字符集[S]. 2008.

[2] Plamondon R, Srihari S N. On-Line and Off-line Handwriting Recognition: A Comprehensive Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1): 63-84.

[3] Abdelazeem S, Eraqi H M. On-line Arabic Handwritten Personal Names Recognition System Based on HMM[C]//Proc. of ICDAR’11. Beijing, China: [s. n.], 2011: 1304-1308.

[4] Biadsy F, El-Sana J, Habash N. Online Arabic Handwriting Recognition Using Hidden Markov Models[C]//Proc. of the 10th International Workshop on Frontiers in Handwriting Recognition. La Baule, France: [s. n.], 2006: 3278-3286.

[5] Ahmed H, Azeem S A. On-line Arabic Handwriting Recognition System Based on HMM[C]//Proc. of ICDAR’11. Beijing, China: [s. n.], 2011: 1324-1328.

[6] 何 松, 戚建宇. 基于改进BP神经网络的手写体数字识别[J]. 计算机科学, 2011, 38(10): 204-205.

[7] Rabiner L R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[8] Bai Zhenlong, Huo Qiang. A Study on the Use of 8-directional Features for Online Handwritten Chinese Character Recognition[C]//

Proc. of ICDAR'05. Seoul, Korea: [s. n.], 2005.

编辑 张正兴