doi:10.3772/j.issn.1002-0470.2013.07.009

识别和几何信息融合的维吾尔文联机手写单词分割①

玛依热·依布拉音②*** 地里木拉提·吐尔逊** 艾斯卡尔·艾木都拉③***

(*武汉大学电子信息学院 武汉 430072)
(**新疆大学信息科学与工程学院 乌鲁木齐 830046)
(***新疆大学软件学院 乌鲁木齐 830046)

摘 要 通过对手写维吾尔文字中的字母连接特点的深入研究,提出了一种有效的基于动态规划的联机手写单词分割方案:首先,去掉单词中的附件部分后,通过分析主要笔划书写轨迹的形状,找出潜在的过分割点并合并被切分成的基本块与对应它的附加部分,得到基本字母片段序列;然后,对相邻的基本片段进行组合形成切分候选网格;再利用单字母分类信息和基于切分块的几何信息进行融合,采用动态规划算法来进行评价分析,从而动态分割并寻找出最优的分割路径。对于联机单词样本进行的实验证明,上述方案对于维吾尔文单词的分割有很好的效果。

关键词 联机手写维吾尔文,单词分割,字母识别,动态规划

0 引 言

近年来,随着智能设备的广泛使用,人们对快速、高效的文字输入的需求迅速增加。手写识别为文字输入提供了一个建立接口的重要组成部分[1]。而且,字符分割是对单词进行识别的关键步骤。目前,从联机手写识别的研究领域来看,英文和汉字文字的研究已经很成熟[2-5],而维吾尔文字的研究才刚刚起步。字母的有效分割是维吾尔文单词识别的一个关键环节。然而,联机维吾尔文字的分割研究至今还非常少见。分割的正确与否在很大程度上决定单词识别系统的性能。分割的目的是在整段文字中分离出待识别的字母,为下一步的特征提取、分类做好准备。错误的切分必然导致错误的识别,因此联机手写维吾尔文分割技术的研究对联机手写识别技术的发展有着很大影响。维吾尔文字是一种在新疆少数民族地区广泛使用的语言文字,研究它的分割方法具有重要意义。维吾尔文字是一种拼音文字,其书写方式与汉文和西文有很大不同,它有很多位于字母上方或下方的附加笔划[6]。字母是维吾尔语文字结构的最基本构件。维吾尔文字的书写方向为从右到左,每个词中所有字母连着写。每一个字母在一个词的词首、词中和词尾所取的字形不一样。每个字母依据在单词中的位置确定使用何种形式。这些特点给维吾尔文字的识别带来很大的困难。维吾尔文印刷体字符识别有一些比较成熟的研究成果[7]。但是,维吾尔语文字识别的研究相对滞后。

根据在字符切分中的单词切割和字符识别之间的关系,目前主要有两类切分方法——不基于识别的切分和基于识别的切分[8-11]。前者又称为先切分后识别的方法,它是算法原理最简单的一类切分方法。它首先把单词或连写字符切割成单字,然后识别单字,因此系统的识别结果对切分无反馈。后者又称为切分同时识别的方法,该类系统将单词或连写字符切割成基元,然后借助单字符识别结果来指导基元进行合并[12-16]。基于识别的切分方法,其预切分算法的性能和切分假设的置信度计算都将影响最终的切分结果。后者利用了字符的识别信息,切分的准确性优于前者。基于识别的切分通常包括两个步骤:“过分割”和“切分路径选择”。“过分割”将输入的单词图像切分成基元。基元是组成整体字符图像的单元,它或者是一个整体字符的图像,或是一个整体字符图像的某部分。基于这些基元,“切

① 国家自然科学基金(61263038)和教育部新世纪优秀人才支持计划(NCET-10-0969)资助项目。
② 女,1981 年生,博士生;研究方向:文字识别,智能图文信息处理;E-mail:mayire401@gmail.com
③ 通讯作者,E-mail:askarhamdulla@gmail.com
(收稿日期:2012-10-11)

分路径选择”步骤采用动态规划算法找到最佳的切分路径。本文采用的策略是基于识别的切分,因为在切分的环节能够引入识别的反馈信息,对提高切分的正确率有帮助。本文针对维吾尔文单词中字母粘连的特点,提出了一种有效的基于动态规划的联机手写单词分割方案,即首先去掉单词中的附件部分后,通过分析主要笔划书写轨迹的形状,找出潜在的过分割点并合并被切分成的基本块与对应它的附加部分,得到基本字母片段序列,然后,对相邻的基本片段进行组合形成切分候选网格,并结合利用单字母分类器输出和基于切分块的几何信息,采用动态规划算法来进行评价分析,从而动态分割并寻找出最优的分割路径。

1 维吾尔文字及其手写体的特点

维吾尔文是在新疆少数民族地区广泛使用的官方语言文字,它借用了阿拉伯文和部分波斯文字母。维吾尔语文字是一种拼音文字,其书写方式与汉文和西文有很大不同。维吾尔文由 32 个字母组成,字母是维吾尔语文字结构的最基本构件。维吾尔文字母形体因独写或在词首、词中、词尾的位置不同而略有不同。每个字母一般都有两种或者四种书写形式。每个字母根据在单词中的位置来确定使用何种形式,其书写形式可以分为四类:独写形式,尾写形式,首写形式和中写形式。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写形式;首部与上一个字母相连的尾写形式;首尾与相邻字母都不相连的独写形式。

维吾尔文字书写时字母连续流畅,自右向左书写。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,字母是沿着一条水平线相连的,这种水平线被称为基线。一个单词包含一个或多个连体段。每个连体段也是由一组字母,或一个字母组成。一个单词内部的字符可能在上重叠。某些字体中,一个尾写(独立)形式字符也可能和另一个首写(独立)形式字符在基线上方或者下方粘连。32 个维文字母中有 20 个字符包含附加部分,附加部分包括不同数目和位置的点以及四种基本形状“ﻩ”、“ﻫ”、“ﯙ”、“ﺍ”。附加部分与字母主体上、下不粘连。字母ﻫ和字母ﻩ的附加部分不是点笔划。带有点笔划的维吾尔文字母很多,点笔划有一点、二点、三点三种标记。

2 过分割

过分割的算法流程如图 1 所示,当有新的单词时,首先去掉单词图像中的附加部分并只对主要连体段部分进行预切分。预分割步骤中,潜在切分点被检测并主要部分可能被碎成非常小的一系列基本片段,所以用基于规则的过滤方法删除额外的切分点并保留初步过分割点。然后利用初步切分块和附加部分的形状知识,合并初步切分块和对应它的附加部分,得到一些候选分割点。

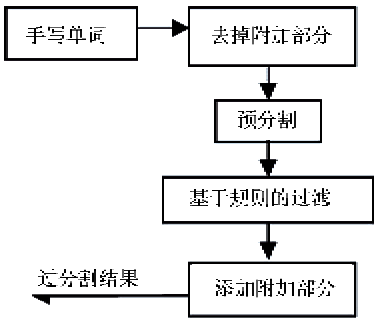

图 1 过分割算法流程图

2.1 去掉附加部分

手写的轨迹数据是根据笔尖在手写板上的运动轨迹按时间顺序获取,所以不但能获得每个点的坐标信息,而且能得到它们的时间序列、笔划数等信息。在手写维吾尔单词过程中,笔尖运动轨迹由它在手写板上的 x,y 坐标和“落笔”、“抬笔”的状态来描述,将笔尖在手写板上的运动轨迹分隔为笔划序列。不管笔划的书写顺序,每个单词包括的笔划可分为主笔划和附加笔划。所以,笔划可能代表连体段的主要部分,也可能是单个字母的主要部分,或有时甚至是一个点。检测并去掉附加部分是一个重要步骤。为了区分维吾尔单词的主要部分和附加部分,同时满足以下条件:第一,对每个笔划进行计算下面的几何特征:每个笔划的宽度、高度、宽高比。如果这些值小于预定的阈值,那么这个笔划属于附加部分。第二,如果这个笔画的书写方向为从右到左($x_i - x_{i+1} > 0$),那么这个笔划属于附加部分。

2.2 预切分

在本节中,去掉附加部分后,主要连体段部分中的潜在切分点被检测候选切分块序列,作为下一个模块中分析并额外的切分点被过滤,然后重建字母

附加部分。从字母基本片段序列中生成选择最优分割的输入。生成候选分割路径的基本步骤如下:

(1) 确定候选分割点:计算从每一个坐标序列点 P_i 到它的下一个点 P_{i+1} 的倾斜角度($\overline{P_iP_{i+1}}$ 与水平线之间的角度)。如果这个角度 α 小于 $\frac{\pi}{6}$ (实验参数值)并且书写方向为从右到左 ($x_i - x_{i+1} > 0$),则该坐标点被称为候选分割点。

(2) 处理重叠:利用空间信息删除一些不满足条件的错误候选分割点。如果过某个初分割点的直线与从$(90-\alpha)°$ 到 $(90+\alpha)°$的弧上任有一个交叉点(也可以说,如果从$(90-\alpha)°$ 到 $(90+\alpha)°$的范围内有覆盖点),则从初步分割点组中删除这个分割点并其余的点都保留。α 是通过实验确定经验值为 25°。

(3) 生成分割线段:通过连接成连续的分割点形成切分线段并确定最终分割点。计算每两个初始分离点之间的水平距离,如果每两个相邻初步分割点差 $g_i - g_{i+1} < 8$ (g_i 表示分割点的横坐标值),则连接成分割点形成切分线段,否则该两个点属于不同的切分线段。

(4) 定位分割点:以上步骤中,本算法可以找到 K 个切分线段。每个切分线段的中间位置就被判断为分割点,从而得到最终的分割点系列 $S_i(i = 1, 2, \cdots, k)$。

2.3 基于规则的过滤

在本步骤中,通过基于规则的方法过滤删除一些不满足条件的错误分割点。

规则 1:根据基线删除多余的分割点。先根据主笔画点序列计算该主笔画的基线,然后从上一步确定的切分点中删除不在基线和基线附近的分割点。这里所提到的基线是指对主笔画点序列进行水平投影后,投影值最大的那条线。如果每个分割点对应的纵坐标 Y 值与基线值的差大于 10(实验参数值),则判断为该分割点不是正确的分割点并删除。

规则 2:如果两个被建议的分割点之间的距离小于一个预定义的阈值(5 为经验值),那么删除该分割点。

2.4 添加附加部分

分割点被检测之后,需要重新把附加部分分配给所属于的切分块,一般这些切分块是字母的主体部分,或是主体部分相连的段。附加部分的正确合并对于切分块识别率的影响非常大。如果不把附加部分分配给这些主体笔划,那么有些原切分块不能单独构成一个字母,因为很多字母的主体笔划是相似或是相同的,唯一区分它们的是它们的附加部分。本文通过计算附加部分和切分块之间的重叠度来判定附加部分的归属。重叠度可利用它们边界框的大小和位置信息计算得到。假设边界框用上下左右边界的坐标表示,那么附加部分的边界框表示为 (x_d^l, x_d^r, x_d^t, x_d^b),切分块的边界框表示为 (x_s^l, x_s^r, x_s^t, x_s^b)。假定 $x_s^l < x_d^l$,如果 $x_d^l < x_s^r$,那么它们相重叠。重叠度和跨度分别表示如下:

$$overlap = x_s^r - x_d^l$$
$$span = max(x_s^r, x_d^r) - x_s^l \quad (1)$$

归一化的重叠度计算如下:

$$normOverlap = \frac{1}{2}\left(\frac{overlap}{width_1} + \frac{overlap}{width_2}\right) - \frac{dist}{span} \quad (2)$$

其中 $width_1$ 和 $width_2$ 分别表示每个切分块的宽度,$dist$ 表示它们中心的水平距离。

单词原始图与切分结果如图 2 所示。

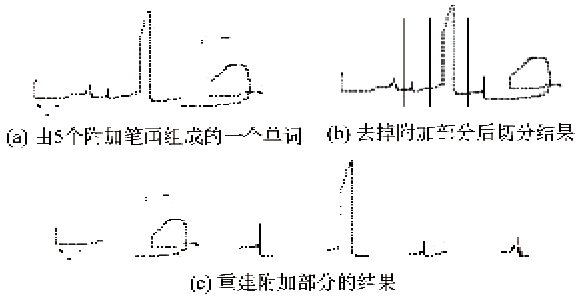

图 2 单词原始图与切分结果

3 基于动态规划的分割

联机手写维吾尔文单词切分的算法流程如图 3 所示,主要由过分割模块、单字母识别模块、基于切分块的几何信息模块和基于动态规划的分割模块等四个模块构成。在过分割基础上,结合利用单字母识别器信息和基于切分块的几何信息,动态规划算法在候选分割点中寻找出最优切分路径。

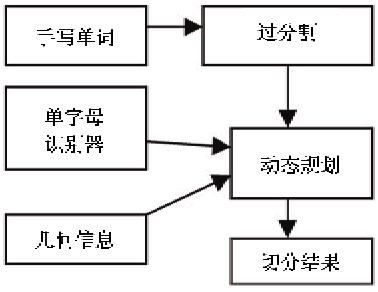

图 3 联机手写单词分割算法流程图

利用动态规划方法解决切分问题的关键在于代价函数的设计,该函数用于描述将一个或多个切分块进行合并时所需的代价。由于本文针对的是维吾尔文单词的切分,因此由每条路径得出的候选切分块,送入维吾尔文字母识别器进行识别后,识别器将给出识别结果。根据切分块的识别信息,得到最终对应整个单词的代价函数。由于在手写单词中被分割出来的字母切分块的结构信息有助于降低错误分割率,因此,本文方法在代价函数中引入基于切分块的几何信息,并与单字母识别器信息进行线性加权,以提高分割性能。

3.1 单字母识别信息和切分块的几何信息

单字母识别主要包括预处理(轨迹平滑和归一化)、特征提取、降维、分类等几个步骤。在联机手写单词分割中,每个过切分后得到的候选切分块的点序列输入到单字母识别器之前都被表示成一个 200 维的特征向量:首先使用伪二维矩归一化(pseudo-two-dimensional moment normalization,P2DMN)方法[17]将其中每个点的坐标进行变化,然后使用连续基于坐标归一化的特征提取(normalization cooperated feature extraction,NCFE)方法[18]提取 8 方向的特征,接着使用 Fisher 线性判别分析(Fisher linear discriminant analysis,FLDA)将上一步提取出来的特征向量维数降到 120 维。而单字母识别器则使用包含 128 个类别的改进二次判别函数(Modified quadratic discriminant function,MQDF)分类器[19]。

切分块的几何信息是指切分候选字符模式相对于整个单词的高度、宽度、位置等信息。手写维吾尔单词中,组成该单词的每个字母大小不保持一致。因此,切分块的高和宽与其平均值的偏差可以作为分割中的重要几何信息。对切分块的几何信息,对每一个字母类别建立一个模型,所用的分类器是 MQDF,使用的特征包括经过字母所在单词归一化了的字母高度、宽度、高宽比、对角线的长度、字母中心与所在单词中心在 y 方向的距离等。

MQDF 是基于高斯概率密度的二次判别函数 QDF 的一种修正版本,通过使用一个常数代替协方差矩阵中较小的特征值,这些特征值对应的特征向量也因此可以忽略不计。这样既降低了存储空间和计算复杂度,也改进了分类性能。记输入模式为 d 维特征矢量 $\boldsymbol{x} = (x_1, x_2, x_3, \cdots, x_d)^{\mathrm{T}}$。假设每类样本服从高斯分布 $p(\boldsymbol{x} \mid \omega_i) = N(\mu_i, \sum_i)$,$\mu_i$ 和 $\sum_i$ 分别表示均值和协方差矩阵,再假设各类先验概率相等,判别函数可以表示为

$$-2\log p(x \mid \omega_i) = (x - u_i)^{\mathrm{T}} \sum_i^{-1} (x - u_i) + \log \left| \sum_i \right| \quad (3)$$

协方差矩阵可以表示为 $\sum_i = \Phi_i \Lambda_i \Phi_i^{\mathrm{T}}$,其中 $\Lambda_i = diag[\lambda_{i1}, \cdots, \lambda_{ik}, \cdots, \lambda_{id}]$ 的对角元素是协方差矩阵的特征值,Φ_i 是由协方差矩阵的特征向量为列向量构成的正交矩阵。用一个常数代替较小的特征值,Λ_i 表示为 $\Lambda_i = diag[\lambda_{i1}, \cdots, \lambda_{ik}, \delta_i, \cdots, \delta_i]$,得到 MQDF 判别函数为

$$f(x, \omega_i) = \sum_{j=1}^{k} \frac{1}{\lambda_{ij}} [(x - u_i)^{\mathrm{T}} \varphi_{ij}]^2 + \frac{1}{\delta_i} \left(\| x - u_i \|^2 - \sum_{j=1}^{k} [(x - u_i)^{\mathrm{T}} \varphi_{ij}]^2 \right) + \sum_{j=1}^{k} \log \lambda_{ij} + (d - k) \log \delta_i \quad (4)$$

3.2 基于动态规划的路径搜索

经过上述的过分割过程,可以得到一系列切分点以及对应的字母基本片段序列,分别用 $\{P_0, P_1, \cdots, P_N\}$ 和 $\{S_0, S_1, \cdots, S_{N+1}\}$ 来表示。过分割的结果往往具有一定的冗余度,过分割后字符可能会被切分成多个部分。一些基本片段可能包含一个单一的维吾尔文字母,还有一些可能只包含字母的一部分,需要做进一步的合并。根据过分割结果构造一个切分候选网格。如图 4 所示,切分候选网格中每一个节点对应一条切分路径,每一条边对应一个候选字符模式。所有候选字符模式构成切分候选网格,网格中从起点到终点的一条路径就对应单词的一种切分方式。

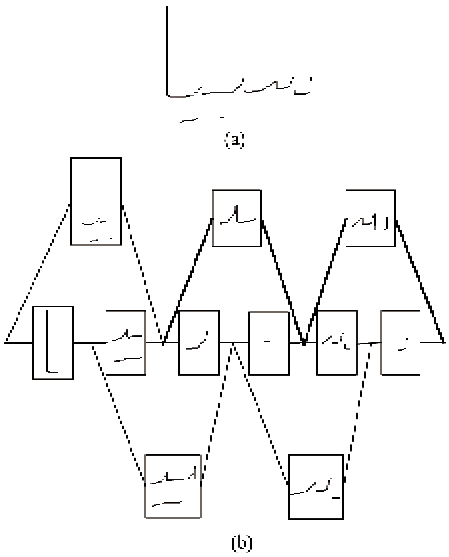

图 4 切分候选网格,粗黑线表示正确的切分路径

基于动态规划的方法,对可能的组合进行评价,搜索最优的评价路径。被输入的单词过切分成基本片段序列,然后组合成候选字符模式。候选字符模式就是指序列中连续片段的组合。本文中,需要将候选字符模式输入到 MQDF 单字母分类器进行识别和验证,以确定正确分割路径。同时,为了较准确地搜索出最佳切分路径,考虑了基于切分块的几何信息。单词切分就是找到最优的基本片段块组合方式,即搜索最优的路径。如图 4 所示,粗黑线表示动态规划算法搜索得到的最佳分割路径。

4 实验结果

4.1 数据库

目前没有公开的维吾尔文手写单词数据库,所以,联机手写数据的采集及整理是维吾尔联机手写单词识别系统中的首要任务。本系统利用汉王手写板获取不同手写者写的单词轨迹信息。手写的轨迹数据是根据笔尖在手写板上的运动轨迹按时间顺序获取,这样我们不但能以"X"和"Y"坐标的形式获得每个点的坐标信息,而且能得到它们的时间序列、笔划数等信息。虽然最终系统处理的是实时、在线的轨迹数据,但是在做实验的过程中,这些轨迹数据文件可以被代替为那些用手写板提供实时轨迹数据的手写者的手写样本。这样,第 i 个单词 W_i 的坐标序列通过下面的公式来描述:

$$W_i = \{(x_{i0}, y_{i0}), (x_{i1}, y_{i1}), \cdots, (x_{ik}, y_{ik}), \cdots, (x_{in}, y_{in})\} \quad (5)$$

在这里 $i = 1,2,\cdots,300$,也就是本文研究中采集的 300 个维吾尔文单词,(x_{i0}, y_{i0}) 和 (x_{in}, y_{in}) 分别表示某个字符 W_i 的第一个和最后一个点的坐标值。单词图像信息将被保存为二进制文件格式,它包括坐标点、坐标总长度、单词的总笔画数和抬笔/落笔等信息。

本实验采集维吾尔文手写单词样本用手写板收集 。手写过程中,要求手写者按照平常习惯的书写方式书写。通过采集得到三个不同书写人写的共 900 个单词样本,包括 5694 个字母。被采集的 300 个单词中,每个单词包括的字母数不等,从 2 到 15 个字母,如表 1 所示。实验中,第一个数据集样本是比较整齐的样本,第二个数据集是以正常风格写的样本,第三个数据集属于自由风格。

表 1 测试集上的单词数

字母数(N)	$2 \leqslant N \leqslant 5$	$6 \leqslant N \leqslant 10$	$11 \leqslant N \leqslant 18$	单词总数	字母总数
单词数	65	203	32	300	1898

4.2 实验结果

为了测试本文提出的分割算法的性能,在上一节中给出的三个不同数据集上做了实验。一般,依据切分算法所应用的识别系统的最终识别结果来衡量大样本集的分割效果。而通过人工观察和统计来给出小样本集上的字符切分的评价。本文用对应小样本集的统计评价方法。本文以召回率(R),精度(p)和 F 度量来评价性能,它们被定义为:

$$R = \frac{\text{检测到的正确的分割位置}}{\text{真实的分割位置总数}} \times 100\%$$

$$P = \frac{\text{检测到的正确的分割位置}}{\text{检测到的分割位置总数}} \times 100\%$$

$$F = \frac{2PR}{P+R} \times 100\%$$

表 2 中给出了代价函数中只考虑单字母识别信息的,基于动态规划的分割方法在不同数据集上的实验结果。结果表明,与过分割方法比较,代价函数只包括单字母识别信息的基于动态规划的分割方法提高了切分精度。

表 2 基于单字母识别信息的分割方法在不同数据集上的性能

	正确分割点总数	过分割点总数	召回率(%)	精度(%)	F 度量
数据集 1	1726	2398	90.94	71.97	80.31
数据集 2	1702	2521	89.68	67.72	77.36
数据集 3	1635	2578	86.15	63.41	72.84

表 3 中给出了结合考虑单字母识别信息和几何信息的分割方法在不同数据集上的实验结果。在代价函数中引入基于切分块的几何信息,并与单字母识别器信息进行线性加权,以提高分割性能。结果表明,在未应用语言模型的情况下,与表 2 比较,表 3 中用到的分割方法提高了切分精度(71.97%→73.54%,67.72%→70.18%,63.41%→67.37%)。实验结果表明,该方法对不同大小的手写字母都能得到满意的效果。

表 3 结合单字母识别信息和几何信息的分割方法在不同数据集上性能

	召回率(%)	精度(%)	F 度量
数据集 1	95.66	73.54	82.76
数据集 2	94.38	70.18	80.77
数据集 3	93.49	67.37	78.24

切分精度的提高将会有助于提高单词识别率的性能。错误分割的主要原因在于以下两个问题:一个是字母附加部分的错误分配;另一个是过分割的问题。

5 结 论

本文提出了一种有效的基于动态规划的联机手写维吾尔文单词分割算法,该方法为今后的维吾尔文单词识别工作奠定基础。维吾尔文字中的粘连和重叠性质,延迟笔划等问题对分割带来不少的困难。本文中,首先用过分割方法把输入的单词切分成基本片段序列,候选字符模式就是序列中连续片段的组合,所有候选字符模式构成切分候选网格,网格中从起点到终点的一条路径就对应单词的一种切分方式。单词切分就是找到最优的基本片段块组合方式,即搜索最优的路径。用维吾尔字联机手写单词数据的初步实验结果表明,本文提出的分割算法具有良好的性能和效果。今后的工作是改进分割方法提高切分精度,并把该分割算法应用到单词识别系统,进一步测试分割性能。

参考文献

[1] Plamondon R, Srihari S N. On-line and off-line handwriting recognition: a comprehensive survey. *IEEE Trans Pattern Analysis and Machine Intelligence*, 2000, 22(1): 63-85

[2] Jaeger S, Manke S, Reichert J, et al. Online hand-writing recognition: The NPen + + recognizer. *International Journal on Document Analysis and Recognition*, 2001, 3(3): 169-180

[3] Khorsheed M S. Off-line Arabic character recognition: a review. *Pattern Analysis and Applications*, 2002, 5(1): 31-45

[4] Liu C L, Koga M, Fujisawa H. Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading. *IEEE Trans Pattern Analysis and Machine Intelligence*, 2002, 24(11): 1425-1437

[5] Wang Q F, Yin F, Liu C L. Handwritten Chinese text recognition by integrating multiple contexts. *IEEE Trans Pattern Analysis and Machine Intelligence*, 2012, 34(8): 1469-181

[6] 买买提沙地克. 基础维吾尔语. 乌鲁木齐: 新疆人民出版社,1992. 25-41

[7] 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. 计算机学报, 2004, 27(11): 1480-1484

[8] Casey R G, Lecolinet E. A survey of methods and strategies in character segmentation. *Pattern Analysis and Machine Intelligence*, 1996, 18(7): 690-706

[9] Benouareth A, Ennaji A, Sellami, M. Arabic handwritten word recognition using HMMs with explicit state duration. *Journal on Advances in Signal Processing*, 2008:1-13

[10] Cheung A, Bennamoun M, Bergmann N W. A new word segmentation algorithm for Arabic script. *Digital Imaging Comput Technical Application*, 1997, 431-435

[11] Lu Y, Shridhar M. Character segmentation in handwritten words-an overview. *Pattern Recognition*, 1996, 29(1): 77-96

[12] Tseng Y H, Lee H J. Recognition-based handwritten Chinese character segmentation using probabilistic viterbi algorithm. *Pattern Recognition Letters*, 1999, 20(8):791-806

[13] Hong C, Gareth L, Wu Y M. Segmentation and recognition of continuous handwriting Chinese text. *International Journal on Pattern Recognition and Artificial Intelligence*, 1998, 12(2):223-232

[14] Tseng L Y, Chen R C. Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming. *Pattern Recognition Letters*, 1998, 19(8): 963-973

[15] Gao X, Lallican P M, Giard-Gaudin C V. A two-stage on line handwritten Chinese character segmentation algorithm based on dynamic programming. In: Proceedings of the 8th International Conference Document Analysis and Recognition, Seoul, Korea, 2005. 735-739

[16] Han Z, Liu C P. A two-stage handwritten character segmentation approach in mail addresses recognition. In: Proceedings of the 8th International Conference Document Analysis and Recognition, Seoul, Korea, 2005. 111-115

[17] Liu C L, Koga M, Sako H, et al. Aspect ratio adaptive normalization for handwritten character recognition. In: Proceedings of the Advances in Multimodal Interfaces—ICMI 2000, Beijing, China, 2000. 418-425

[18] Liu C L, Zhou X D. Online Japanese character recognition using trajectory-based normalization and direction feature extraction. In: Proceedings of the 10th International

Workshop on Frontiers in Handwriting Recognition, La Baule, France, 2006. 217-222

[19] Kimura F, Takashina K, Tsuruoka S, et al. Modified quadratic discriminant functions and the application to Chinese character recognition. *Pattern Analysis and Machine Intelligence*, 1987, 9(1): 149-153

Segmentation of online uyghur handwritten words by integrating recognition and geometric information fusion

Mayire Ibrayim[*,**], Dilmurat Tursun[**], Askar Hamdulla[***]

([*]School of Electronic Information, Wuhan University, Wuhan 430072)

([**]College of Information Science and Engineering, Xinjiang University, Urumqi 830046)

([***]College of Software, Xinjiang University, Urumqi 830046)

Abstract

Considering that correct and efficient segmentation of Uyghur words into characters is crucial to the successful recognition of Uyghur words, this paper presents a novel character segmentation method using dynamic programming in online cursive Uyghur handwriting to enable many connected characters separation in cursive Uyghur handwriting. The method consists of the steps below. Firstly, after removing delayed strokes from the handwritten words, potential breakpoints are detected from concavities and ligatures by temporal and shape analysis of the stroke trajectory, and reconstruct delayed strokes and a sequence of primitive segments are obtained. Then, by merging the neighboring blocks, candidate segmentation paths are created. Further, paths are evaluated by the character recognition and geometric information, and a dynamic programming algorithm is used to find the best segmentation point for each character. The preliminary experiments on an online Uyghur word dataset demonstrate that the proposed method can achieve good performance in segmenting cursive handwritten Uyghur characters.

Key words: online Uyghur handwriting, word separation, character recognition, dynamic programming