基于候选特征笔画和多类阈值的手写汉字切分①

马建平[1], 汪庆锋[1], 陈 渤[2], 陈 强[3]

[1](浙江工业大学 计算机科学与技术学院, 杭州 310023)
[2](浙江商业职业技术学院 应用工程学院, 杭州 310053)
[3](广东第二师范学院 计算机科学系, 广州 510310)

摘 要: 通过分析汉字的常见结构, 鉴于汉字与汉字之间的距离和构成汉字的部件之间的距离的显著差异性, 提出一种基于候选特征笔画和多类阈值的手写汉字切分方法. 首先从构成手写汉字的笔画集合中提取候选特征笔画, 根据候选特征笔画将手写汉字预切分, 然后利用基于间距阈值的部件组合规则对过切分的汉字部件进行组合, 最后采用基于单字宽度阈值的粘连汉字判断规则搜索粘连汉字, 对粘连汉字进行递归切分. 实验表明, 该方法对连续手写汉字的切分准确率较高, 具有一定的实用性.

关键词: 汉字结构; 手写; 特征笔画; 阈值; 切分

Segmentation of Handwritten Chinese Characters Based on the Candidate Characteristics Strokes and Multi-Class Threshold

MA Jian-Ping[1], WANG Qing-Feng[1], CHEN Bo[2], CHEN Qiang[3]

[1](School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)
[2](School of Application Engineering, Zhehuabg Vocational College of Commerce, Hangzhou 310053, China)
[3](Department of Computer Science, Guangdong University of Education, Guangzhou 510310, China)

Abstract: With the analysis of the common structures of Chinese characters, a handwritten Chinese characters segmentation method which based on the candidate characteristics strokes and multi-class threshold is proposed in view of the significant difference between the distance between Chinese characters and the distance between the components that constitute a Chinese character. Firstly, the candidate characteristics strokes are extracted from the collection of strokes that constitute handwritten Chinese characters, the handwritten Chinese characters are pre-segmented according to the candidate characteristics strokes. Then, the over-segmentation components of Chinese characters are combined by using a components combination rule based on the spacing threshold. Finally, the adhesions Chinese characters are searched by using an adhesions Chinese character judgment rule based on the single character width threshold and recursive segmented. The experiments show that the segmentation accuracy of continuous handwritten Chinese characters is quite high by this method, having certain practical.

Key words: structure of Chinese characters; handwriting; characteristics stroke; threshold; segmentation

随着移动通信产业的高速发展, 移动终端[1]已经迈入了全触屏时代, 手写输入[2]对移动终端的支持提升了移动终端的综合优势, 为用户带来了全新的人机交互体验方式. 手写输入的识别率较高, 可以在输入完毕后快速的识别出用户期望的结果. 目前, 手写输入只针对单个汉字的输入和识别, 下一个汉字需要在上一个汉字识别结束后才能输入和识别, 输入效率较低, 不符合人们日常生活中使用纸笔连续输入汉字的交互方式[3].

因此, 对连续手写汉字进行正确切分并对切分后的汉字逐个识别是解决目前手写输入存在的缺陷的有效途径. 目前, 手写字符切分有三类基本方法[4]: 第一

① 基金项目:国家自然科学基金(61075118,60907032);浙江省教育厅科研项目(Y201122614);2011 年广东省现代信息服务业发展专项资金(13090);2011 年东莞市现代信息服务业发展专项资金竞争性项目(DG201101);2012 年省高等院校学科建设专项资金(2012KJCX0079)
收稿时间:2012-10-27;收到修改稿时间:2012-12-08

类是基于结构分析的切分, 即从手写字符的投影图像中寻找特征信息, 从特征信息中寻找字符切分规则, 文献[5-7]中使用该类方法对手写字符的投影图像进行分析, 提出了相应的字符切分规则; 第二类采用先切分, 然后按规则组合的方式对字符进行切分, 文献[8-10]中均提出了各自的组合规则, 将切分后的字符部件按照规则进行组合, 实现手写字符的切分; 第三类采用一种整体切分策略, 将手写字符作为一个整体进行词识别, 避免字符切分.

通过分析各类切分方法的优劣性以及移动终端自身的物理特性, 本文采用第二类方法对手写汉字进行切分. 以汉字的结构特征为基础, 根据汉字与汉字之间的距离和构成汉字的部件之间的距离的显著差异性, 提出一种基于候选特征笔画和多类阈值的手写汉字切分方法, 对手写汉字进行预切分、过切分部件组合、粘连汉字递归切分, 以实现连续手写汉字的正确切分.

1 汉字结构分析

从计算机处理汉字信息的角度来说, 汉字被分为位点、笔画、部件和单字四个层次[11]. 其中部件是构成汉字的最小笔画结构单位, 目前没有确切的定义, 本文基于文献[12]中的描述, 将部件定义为由单个或者多个笔画构成的独立存在单位, 也即一个部件中必须至少包含一个笔画.

两个部件之间的位置关系主要可以分为三类: 上下、左右、包围, 其中包围关系还可以细分为 7 种. 图 1 中显示了两个部件之间的 9 种可能的位置关系, 所有的汉字就是根据这些基本的位置关系由单个或者多个部件构成的.

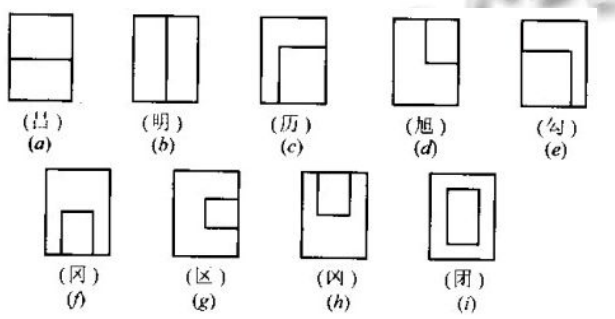

图 1 两个部件之间的位置关系

2 手写汉字切分

2.1 预切分

汉字与汉字之间的距离和构成汉字的部件之间的距离差异性通过新汉字的首个笔画和同一汉字新部件的首个笔画两者之间的不同定位措施来体现. 一般来说, 用户会将新汉字的首个笔画输入位置定位到远离前一个汉字的区域, 而将同一汉字新部件的首个笔画输入位置定位到靠近前一个部件的区域. 图 2 中分别显示了上述两种情况下对应的首个笔画, 本文用红色矩形框对它们进行了标识.

图 2 用户手写输入汉字习惯

上述两种情况下的首个笔画定位方式存在较大的差异性, 但是与前一个汉字或者部件之间都会存在距离(一般是指左右距离, 因为大多数人是按照从左往右, 从上往下的手写习惯输入汉字的), 唯一的差别就是距离的大小. 因此, 本文将构成手写汉字的笔画集合中与前一个部件之间存在左右距离的笔画定义为候选特征笔画, 提取符合该定义的候选特征笔画, 然后根据候选特征笔画将手写汉字预切分.

记构成手写汉字的笔画集合为 *S*, 预切分后的部件集合为 *C*, 定义一个临时的部件对象 *tempC*. 手写汉字预切分的算法如下:

(1) 遍历 *S*, 依次获取一个笔画对象 *s*, 计算 *s* 在移动终端屏幕上的左边界值;

(2) 若 *s* 为 *S* 中的首个笔画, 清空 *tempC* 中的所有笔画, 再将 *s* 添加到 *tempC* 中, 继续执行(1), 否则计算 *tempC* 在移动终端屏幕上的右边界值;

(3) 比较 *s* 的左边界值和 *tempC* 的右边界值, 若 *s* 的左边界值不大于 *tempC* 的右边界值, 将 *s* 添加到 *tempC* 中, 否则 *s* 为本文定义的一个候选特征笔画, 将 *tempC* 作为一个预切分后的部件添加到 *C* 中, 清空 *tempC* 中的所有笔画, 再将候选特征笔画 *s* 添加到 *tempC* 中, 作为新部件的首个笔画;

(4) 遍历 *S* 结束后返回 *C*, *C* 中存放的即为根据候选特征笔画进行预切分后的部件集合, 其中部分的部件已经为完整的汉字, 剩余的为过切分的汉字部件, 需要对过切分的汉字部件进行组合.

2.2 过切分部件组合

根据 2.1 中的描述, 提取的候选特征笔画可以分

为两类: 一类用于区分不同的汉字, 是正确切分汉字的依据, 本文将这类笔画定义为特征笔画; 另一类用于区分同一个汉字的不同部件, 导致了汉字的过切分, 本文将这类笔画定义为非特征笔画.

上述的两类笔画与前一个部件之间的左右距离具有显著的差异性, 针对这一特性, 本文提出一种基于间距阈值的部件组合规则. 记间距阈值为 *T*, *c*1、*c*2 为 *C* 中任意的两个相邻部件, 组合规则定义如下:

如果 *c*2 的左边界值与 *c*1 的右边界值的差值大于 *T*, 那么 *c*1、*c*2 为不同汉字的部件(*c*1、*c*2 也有可能都为完整的汉字), *c*2 的首个笔画为一个特征笔画, 不需要将 *c*1 和 *c*2 进行组合.

否则 *c*1、*c*2 为同一个汉字的不同部件, *c*2 的首个笔画为一个非特征笔画, 导致了 *c*1、*c*2 的过切分, 需要将 *c*1 和 *c*2 进行组合.

根据组合规则的定义, 间距阈值 T 的取值优劣会影响两类笔画的区分, 影响过切分部件的组合, 因此, 选取一个合理的间距阈值是关键. 由于用户的手写输入习惯存在差异性, 不能使用一个固定的间距阈值来区分这两类笔画, 因此, 本文令间距阈值 $T=\mu*Da$, 其中 μ 为一个常数, *Da* 为部件集合 *C* 中相邻部件之间的平均间距. 采用这种方式定义的间距阈值 *T* 会随着 *C* 的不同而动态变化, 自适应的调整合理的取值, 时刻体现用户的手写输入习惯, 以实现最佳的过切分部件组合效果.

记 2.1 中获取的集合 *C* 中的部件个数为 *lc*, 组合后的部件集合为 *MC*, 定义 *Da*, 初始化 *Da*=0, 定义计数游标 *num*, 初始化 *num*=0. 过切分部件的组合算法如下:

(1) 遍历 *C*, 依次获取两个相邻的部件 *c*1、*c*2, 计算 *c*2 的左边界值与 *c*1 的右边界值的差值;

(2) 判断差值, 若差值大于零, 将这个差值累加到 *Da* 中, 同时累加计数游标 *num* 的值;

(3) 遍历结束后将 *Da* 的值除以 *num*, 计算本文所需的 *Da* 值, 同时赋值间距阈值 *T*;

(4) 再次遍历 *C*, 依次获取两个相邻的部件 *c*1、*c*2, 计算 *c*2 的左边界值与 *c*1 的右边界值的差值, 若差值不大于 *T*, 表明 *c*1、*c*2 为本文定义的两个过切分的汉字部件, 其中 *c*2 中的首个笔画即为本文所定义的一个非特征笔画, 需要将这两个相邻部件进行组合;

(5) 将过切分的汉字部件 *c*1、*c*2 进行组合, 组合完毕后将 *c*1 置为 *null*, *c*2 用于表示组合后的新部件(新部件可能为一个完整的汉字, 也有可能仍然为某个汉字的部件);

(6) 所有过切分的汉字部件组合完毕后再次遍历 *C*, 将 *C* 中不为 *null* 的部件添加到 *MC* 中, 实现 *null* 部件的筛选;

(7) 遍历 *C* 结束后返回 *MC*, *MC* 中存放的即为经过组合后的部件集合, 实现了初步的手写汉字切分.

2.3 粘连汉字递归切分

用户在移动终端上手写输入汉字时也会存在个别汉字之间间距较小的情况, 当将这类间距与间距阈值 *T* 进行比较时, 会将相邻的汉字判断为同一个汉字的两个不同部件并且被组合, 导致粘连汉字的产生. 因此, 本文同样将用户的手写习惯作为研究依据, 提出一种基于单字宽度阈值的粘连汉字判断规则. 记单字宽度阈值为 *WT*, *c* 为集合 *MC* 中的任意一个部件, 粘连汉字的判断规则描述如下:

如果部件 *c* 的宽度大于 *WT*, 那么 *c* 为一个粘连汉字, 需要将 *c* 进行递归切分.

否则部件 *c* 为一个单字, 无需对 *c* 进行切分.

同样的, *WT* 的取值优劣会影响粘连汉字与单字的区分, 比照间距阈值 T 的定义方式, 本文令 $WT=\xi*Wa$, 其中 ξ 为一个常数, *Wa* 为部件集合 *MC* 中所有对象的平均宽度. *WT* 同样的会随着 *MC* 的不同而动态变化, 自适应的调整合理的取值, 时刻体现用户的手写输入习惯, 达到最佳的区分效果.

记 2.2 中获取的集合 *MC* 中的部件个数为 *mcl*, 递归切分后的单字集合为 *G*, 定义 *Wa*, 初始化 *Wa*=0, 定义一个临时的粘连汉字对象 *tempG*. 粘连汉字的递归切分算法如下:

(1) 遍历 *MC*, 依次获取一个部件对象 *mc*, 计算 *mc* 的宽度, 并将宽度值累加到 *Wa* 中;

(2) 遍历结束后将 *Wa* 的值除以 *mcl*, 计算本文所需的 *Wa* 值, 同时赋值单字宽度阈值为 *WT*;

(3) 再次遍历 *MC*, 依次获取一个部件对象 *mc*, 计算 *mc* 的宽度, 若宽度值不大于 *WT*, 表明 *mc* 为一个单字, 直接将 *mc* 添加到 *G* 中, 继续执行(3), 否则 *mc* 为一个粘连汉字, 需要对其进行递归切分, 令 *tempG=mc*;

(4) 提取 tempG 的特征笔画序号 *index*, 通过比较 *tempG* 中相邻笔画之间的左右间距, 取左右间距最大的两个相邻笔画中的右边笔画(也即特征笔画)的序号;

(5) 根据 *index* 将 *tempG* 切分成部件 *c*1、*c*2，若构成 *tempG* 的笔画集合长度为 *length*，那么 *c*1 中包含 *tempG* 中序号从 0 到 *index*－1 的笔画集合，*c*2 中包含 *tempG* 中序号从 *index* 到 *length*－1 的笔画集合；

(6) 计算 *c*1 的宽度，若宽度值大于 *WT*，表明 *c*1 仍为粘连汉字，对其递归切分，令 *tempG*=*c*1，执行(4)、(5)，否则 *c*1 为一个单字，将 *c*1 添加到 *G* 中；

(7) 计算 *c*2 的宽度，若宽度值大于 *WT*，同 *c*1，令 *tempG*=*c*2，执行(4)、(5)，否则 *c*2 为一个单字，将 *c*2 添加到 *G* 中；

(8) 遍历 *MC* 结束后返回 *G*，*G* 中存放的即为经过递归切分粘连汉字后的单字集合，实现了最终的手写汉字切分.

3　实验结果及分析

为说明本文提出的手写汉字切分方法的有效性，笔者对用户在移动终端上手写输入的汉字进行了切分实验，实验所用的设备为 HTC Sensation(G14)智能手机，该设备使用的系统版本为 Google Android 2.3，处理器频率为 1.2GHz. 考虑到该设备屏幕尺寸的有限性，本文将单行的手写汉字作为切分对象.

本文中所定义的常数 μ 和 ξ 是通过分析用户的手写输入习惯而得到的，会根据用户的不同而发生改变，在切分实验中各参数的取值分别为：μ =0.85，ξ =1.5，图 3 显示了手写汉字切分的整个过程.

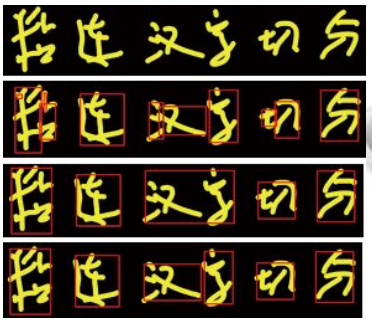

图 3　手写汉字切分过程

从图 3 可以看出，本方法首先获取用户输入的单行手写汉字，将手写汉字预切分成部件集合，对过切分的部件进行组合，对粘连汉字进行递归切分，最后得到了正确的切分结果.

为了说明本文方法的普遍适用性，图 4 给出了其余的部分切分实验结果，其中包括切分成功和切分失败的例子.

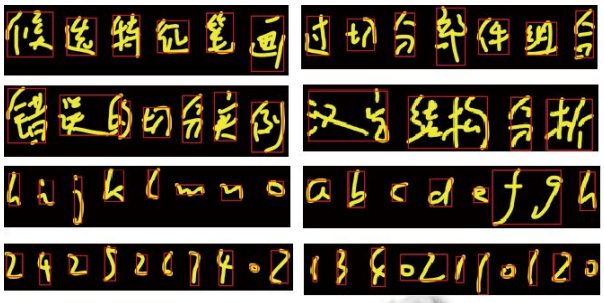

图 4　其余的部分切分实验结果

从图 4 可以看出，本文提出的切分方法具有一定的实用性，部分字符之间间距过小使得粘连字符没被有效的递归切分是导致切分失败的主要原因. 此外，本文提出的方法对单行手写英文字母和手写阿拉伯数字的切分也具有同样的效果，经过预切分、过切分部件组合、递归切分粘连字符后都获得了正确的切分结果，验证了本文方法的普遍适用性.

笔者对实验结果进行了统计分析，实验中，单个汉字的输入形式可以为草书或正楷，汉字与汉字之间不允许连笔. 表 1 记录了具体的实验结果，比较了直方图投影分割法和本文方法对单行手写汉字的切分正确率.

表 1　手写汉字切分正确率比较

单个汉字输入形式	切分次数	切分正确率	
		直方图投影法	本文方法
均为草书	200	75%	97%
草书和正楷	200	70%	93%
均为正楷	200	63%	91%

从表 1 可以看出，当单字输入形式均为草书时，本文方法的切分正确率达到了 97%，已经具有了一定的实用性；当单字输入形式均为正楷时切分正确率略微有所下降，为 91%，但是仍然保持着较高的正确率. 同时，在相同的单字输入形式下，本文方法的切分正确率均高于直方图投影分割法，验证了本文方法的实用性和有效性.

4　结语

鉴于目前移动终端手写输入存在的缺陷，提出一种基于候选特征笔画和多类阈值的手写汉字切分方法，对单行的手写汉字进行预切分、过切分部件组合、粘连汉字递归切分. 实验结果表明，本文提出的方法具有较高的切分正确率，并且对其它类型的手写字符也

有同样的切分效果, 具有一定的普遍适用性和有效性, 为移动终端实现连续手写汉字的输入和识别提供了有效的参考依据.

参考文献

1 Knudsen MB, Pedersen GF. Spherical outdoor to indoor power spectrum model at the mobile terminal. IEEE Journal on Selected Areas in Communications,2002,20(6):1156–1169.

2 Lee J, Chung Y. Design of a wireless handwriting input system for mobile devices. Proc. of the 9th International Symposium on Consumer Electronics. Macau: IEEE Computer Press, 2005,222–225.

3 韩勇,须德,戴国忠.MST 在手写汉字切分中的应用.软件学报,2006,17(3):403–409.

4 Casey RG, Lecolinet E. A survey of methods and strategies in character segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence,1996,18(7):690–706.

5 罗佳,王玲.基于凹凸特性的非限制粘连手写数字串切分.微计算机信息,2007,23(25):275–276.

6 马瑞,夏永泉,杨静宇.基于背景分析的手写数字切分方法.计算机科学,2007,34(1):198–200.

7 赵姝岩,郭捷,施鹏飞.基于笔画分析和背景细化的粘连手写汉字切分.上海交通大学学报,2003,37(9):1434–1437.

8 Hong C, Loudon G, Wu Y, Zitserman R. Segmentation and recognition of continuous handwriting Chinese text. Pattern Recognition and Artificial Intelligence, 1998,12(2):223–232.

9 Lin YuT, Chen RC. Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming. Pattern Recognition Letters, 1998,19(8):963–973.

10 Tseng LY, Chuang CT. An efficient knowledge based stroke extraction method for multi-font Chinese characters. Pattern Recognition,1992,25(12):1445–1458.

11 吕岳,施鹏飞,张克华.基于汉字结构特征的自由格式手写体汉字切分.电子学报,2000,28(5):1–3.

12 傅永和.汉字结构和构造成分的研究.现代汉语用字信息分析.上海:上海教育出版社,1993.108–169.

(上接第 142 页)

参考文献

1 Deb K.An efficient constraint handling method for genetic algorithms.Computational Methods for Applied Mechanical Engineering,2000,18(9):311–318.

2 Michalewicz,Z,S choenauer,M.Evolutionary algorithms for constrained parameter optimization problems.Evolutionary Computation Journal,1996,4(1):1–32.

3 张晶,翟鹏程.惩罚函数法在遗传算法处理约束问题中的应用.武汉理工大学学报,2002,24(2):56–59.

4 荣喜民,安智宇.非线性规划的混合遗传算法.系统工程与电子技术,2003,25(5):621–624.

5 刘伟,刘海林.基于外点法的混合遗传算法求解约束优化问题.计算机应用,2007,27(1):238–240.

6 周永华,李鹏,毛宗源.一种新的混合杂交方法及其约束优化中的应用.计算机工程与应用,2006,27(6):48–51.

7 刘淳安.解非线性约束规划问题的新型多目标遗传算法.计算机工程与设计,2006,27(5):756–757.

8 余新华,孙作龙.带约束函数优化问题的新算法.武汉理工大学学报,2002,24(5):13–16.

9 李秀梅,刘华毅,徐景德.一种新的遗传算法求解约束优化问题.计算技术与自动化,2003,22(1):17–20.

10 林丹,李敏强,寇纪淞.基于遗传算法求解约束优化问题的一种算法.软件学报,2001,12(4):628–632.

11 敖友云,迟洪钦.一种求解约束函数优化问题的遗传算法.燕山大学学报,2005,29(4):294–297.

12 戴庆,申静波.基于遗传算法的运输问题最优解研究.天津理工大学学报,2008,24(3):43–45.