基于汉字结构特征的自由格式手写体汉字切分

吕 岳,施鹏飞,张克华
(上海交通大学图像处理与模式识别研究所,上海 200030)

摘 要: 本文利用汉字的基本结构特征切分自由格式手写体汉字字符,以汉字部件作为基本要素,根据部件之间的位置关系进行合并形成完整的汉字图像.自由格式手写体汉字串常有两个字符粘连的情况,依据汉字的笔画特征可以切割粘连字符.实验结果表明,该方法对自由格式手写体汉字字符有较好的切分效果.

关键词: 手写体汉字; 结构特征; 字符切分; 自由格式
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2000) 05-0102-03

Segmentation of Free-format Handwritten Chinese Characters Based on Structure Features of Characters

LU Yue, SHI Peng-fei, ZHANG Ke-hua
(*Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai* 200030, *China*)

Abstract: A new method for segmenting free-format handwritten Chinese characters is presented in this paper. Character elements and their topological relations are utilized to merge character image. Connected handwritten characters are separated according to the features of Chinese character strokes. Experimental results show that the method has satisfactory performance for segmenting free-format handwritten Chinese Characters.

Key words: handwritten Chinese character; structure features; character segmentation; free-format

1 引言

汉字识别技术经过几十年的发展,尤其是我国科研工作者的不懈努力,已取得了长足的进步[1].目前在线手写体汉字识别和离线印刷体汉字识别都已形成商业产品,向实用化迈出了坚实的一步.然而离线手写体汉字识别离实用化要求还有一定的距离,一方面是由于手写体汉字的书写风格因人而异,以及汉字识别技术本身固有的困难;另一方面,汉字字符的切分也是汉字识别技术走向实用化的一大障碍.实践表明,汉字识别系统的正确识别率与正确的汉字切分密切相关,错误切分将导致错误识别.许多手写汉字识别系统限定将汉字工整书写在方格纸上,从而大大降低汉字切分的难度,但也使人们的书写习惯受到很大的限制,导致识别系统远离社会需求.在尽量适应人们日常书写习惯的前提下,研究自由手写体汉字的切分显得更加迫切和重要.虽然这项研究面临着诸多困难和挑战,以一些专用系统(如邮政信函的手写体汉字地址阅读、票据自动阅读及文稿的自动录入等)为突破口,将离线手写汉字识别技术推向实用大有前途.

字符切分有三类基本方法[2],其一是基于结构分析的切分,即从图像特征中寻找字符切分的规则;其二是以识别为基础的切分;还有一类是整体切分策略,即系统将字符串作为一个整体进行词识别而不是字识别.近年来,对西文字符和数字的切分人们做了大量卓有成效的工作[3,4],对汉字字符的切分主要针对印刷体文本进行处理[5],而对手写体汉字的切分鲜有研究.文献[6]对此作了有益的尝试,文中假定字符串的汉字书写比较匀称,难以满足实用要求,此外采用投影的方法切分手写体汉字,在字符间距较小甚至字符粘连的情况下显得无能为力.

就结构而言,汉字字符与西文字符和数字有较大的区别.通常汉字字符的结构比较复杂,图像可能包含几个连通元,一般需要将分离的几个连通区进行适当的合并,这在很大程度上增加了汉字字符切分的难度.汉字由位点、笔画、部件和单字四个层次组成,部件是构成汉字的最小结构单位,虽然至今尚无统一定义,但对部件的分析有益于汉字的切分和识别[7].在离线汉字的切分中,将部件与连通元之间建立有效的联系,并分析部件间的位置关系,对字符的切分有很大的帮助.本文以汉字的部件为基础,根据部件之间的上下、左右和包围关系对部件进行合并,对字符间有粘连的情况依据汉字的笔画特征切割粘连字符.实验结果表明,该方法对自由格式手写体汉字字符有较好的切分效果.

2 汉字结构特点

从语言文字学的角度出发对汉字结构进行分析,汉字由笔画、偏旁部首和单字三级组成.这一理论强调偏旁部首有

收稿日期:1999-02-08;修订日期:1999-05-10
基金项目:邮电部科研资助课题

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

特定的音或义,而不完全是从结构特征出发分析汉字. 从计算机汉字信息处理角度出发,用位点(象素)、笔画、部件和单字四个层次分析汉字,即单字由部件构成.

部件尚无确切的定义,研究者一般按照实际使用的要求来划分. 部件是构成汉字的最小笔画结构单位,其下限必须大于基本笔画,上限小于等于偏旁. 从功能上看,部件并不一定具有音和义;从存在形式上看,它是一个独立的书写单位,不管笔画多么复杂,凡是笔画串在一起的,都作为一个部件看待[8]. 根据上述定义,对《辞海》(1979 年版)收入的 11834 个规范汉字进行统计,得出 648 个部件,其中 327 个为独体字,其余 321 个为非字部件,没有音和义.

任意两个部件之间的关系有以下 9 种可能情况,如图 1 所示,其中的字例是由两部件构成的单字,更复杂的汉字可由几个这样的关系复合而成.

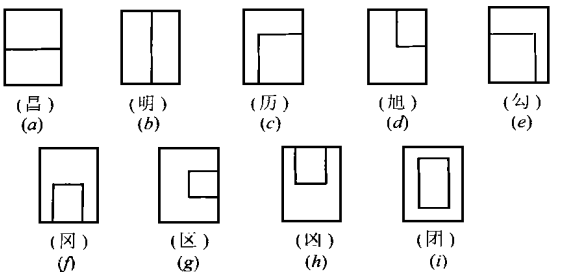


图 1　两个部件间的关系

部件有单笔(如一、乙等)和多笔之分,多笔部件的笔画之间有三种位置关系,即交叉(如十、井、也等)、交连(如人、丁、工、厂等)和分离(如彳、氵、灬、寸、心等). 从图像的角度看,单笔部件以及由交叉和交连笔画组成的多笔部件是一个独立的连通元;而由分离笔画组成的部件中有几个连通元,存在包围、上下和左右几种位置关系. 对图像的连通元进行分析比较容易,因而在定义部件时建立连通元与部件之间的联系是必要的,本文即将连通元作为部件看待.

在手写体汉字识别中,笔画常为笔段代替. 笔段有长度和方向特性,方向上有横、竖、撇、捺四种,分别对应着水平、垂直、右斜、左斜.

3　汉字字符切分

汉字字符的切分就是对每个汉字字符用一个矩形框给予定界. 汉字图像往往有多个连通元组成,手写体汉字的书写有随意性,字间的空白间隙无规则可循,很难采用西文字符和阿拉伯数字切分时常用的投影方法. 手写体汉字字符的宽高比不如印刷体汉字稳定,但是组成汉字的部件有固定的拓扑关系,所以根据部件的位置关系切分汉字可以得到较好的效果.

3.1　部件的合并

两个部件之间的位置关系,可分为三类,即上下关系(图 1(a))、左右关系(图 1(b))和包围关系(图 1(c)～(i)). 对它们分别处理:

·上下关系:合并在同一汉字内

包含关系:合并在同一汉字内

·左右关系:利用整个字符串中部件的分布信息进行合并

每一部件的位置信息包括部件的宽度 $W^{(m)}$、高度 $H^{(m)}$、左上角坐标 $(LTx^{(m)}, LTy^{(m)})$、右下角坐标 $(RBx^{(m)}, RBy^{(m)})$ 以及中心位置 $(Cx^{(m)}, Cy^{(m)})$,其中 $1 \leq m \leq M$, M 为部件总数,坐标轴的定义以图像左上角为原点,水平方向为 x 轴,垂直方向为 y 轴. 将部件 (m) 和部件 (n) 合并为部件 (k),作如下操作:

$$
\begin{aligned}
LTx^{(k)} &= \min(LTx^{(m)}, LTx^{(n)})\\
LTy^{(k)} &= \min(LTy^{(m)}, LTy^{(n)})\\
RBx^{(k)} &= \max(RBx^{(m)}, RBx^{(n)})\\
RBy^{(k)} &= \max(RBy^{(m)}, RBy^{(n)})\\
W^{(k)} &= RBx^{(k)} - LTx^{(k)}\\
H^{(k)} &= RBy^{(k)} - LTy^{(k)}\\
Cx^{(k)} &= (RBx^{(k)} + LTx^{(k)})/2\\
Cy^{(k)} &= (RBy^{(k)} + LTy^{(k)})/2
\end{aligned}
$$

首先将所有部件按中心位置的 x 轴坐标顺序排列,根据以下部件合并算法对部件进行合并.

部件合并算法

输入:按 x 方向顺序排列的部件位置信息,

输出:汉字字符的定界,及字符的平均宽度和最大高度.

步骤:

Step 1. 初始化 $p = 0$

Step 2. 对所有部件

If $Cx^{(m)} > LTx^{(m+1)}$. and. $Cx^{(m)} < RBx^{(m+1)}$

合并 (m) 和 $(m+1)$, $p++$

If $Cx^{(m-1)} > LTx^{(m)}$. and. $Cx^{(m-1)} < RBx^{(m)}$

合并 (m) 和 $(m-1)$, $p++$

Step 3. 计算整个字符串中部件合并后的平均宽度 Wa,平均间距 Ga 和最大高度 $H_{\max}$

Step 4. 对所有部件

If $W^{(m)} < \varepsilon Wa$. and. $(LTx^{(m+1)} - RBx^{(m)}) < \lambda Ga$

合并 (m) 和 $(m+1)$, $p++$

If $W(m) < \varepsilon Wa$. and. $(LTx^{(m)} - RBx^{(m-1)}) < \lambda Ga$

合并 (m) 和 $(m-1)$, $p++$

Step 5. If $p \neq 0$ go to step 1, else stop.

其中,ε 和 λ 是常数因子.

3.2　粘连汉字的切割

由于手写体汉字的随意性,常出现两个汉字粘连在一起的情况. 根据以下原则判断可能有粘连字符存在:

·$W^{(m)} > \rho Wa$

·$W^{(m)} / H_{\max} > \sigma$

ρ 和 σ 为常数.

汉字的笔画在字符中的的位置有一定的规律. 对粘连字符图像分析发现,字符中有长笔画时容易发生粘连现象,比如字符中的长捺跟后一字符相连,长撇与前一字符相连. 根据这一特点,在可能粘连的图像中进行笔画分析,提取横竖撇捺四种笔段,重点分析长撇和长捺及其在图像所处位置,从而将粘

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

连字符切割开来.

4 实验结果

以邮政信函上的手写体地址作为处理对象,由信函自动分拣机采集到真实信函上整个信封的二值化图像,经行切分获得收信人地址区的图像,再作字符切分获取独立的汉字字符图像进行地址识别. 图 2 是字符切分的过程,先提取地址区汉字字符串的部件,再根据部件之间的位置关系作适当的合并,最后切分出单个的汉字字符图像. 图 3 给出了部分地址汉字切分的结果. 图 4 是有粘连字符的切分情况.

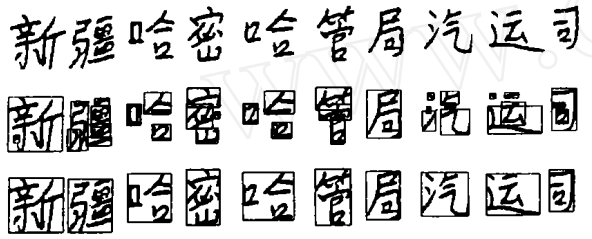


图 2 基于结构特征的字符切分

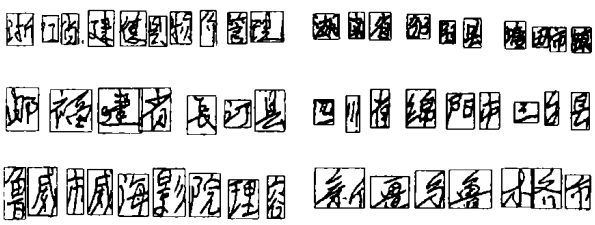


图 3 部分信函地址的字符切分结果

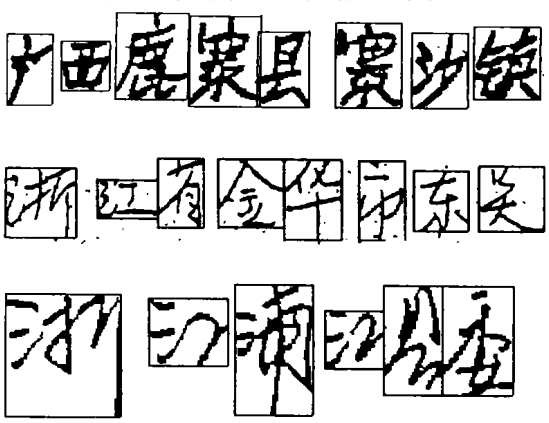


图 4 信函地址中有粘连字符的切分结果

5 结论

手写体汉字字符切分是离线汉字识别技术实用化中的关键技术之一. 本文在自由格式的手写体汉字字符切分中采用汉字的部件特征和汉字的笔画结构特征,效果较好. 但仅仅依靠汉字的结构特征尚不能完全实现自由格式手写体汉字的切分,进一步的研究,将在汉字切分过程中融入部件识别、单字识别和语义理解,可得到更好的切分效果.

参考文献

[1] 周昌乐. 手写汉字的机器识别. 北京:科学出版社,1997 年

[2] Casey R G and Lecolinet E. A survey of methods and strategies in character segmentation. IEEE trans. PAMI,1996,18(7):690～709

[3] Lu Y. Machine printed character segmentation-an overview. Pattern Recognition,1995,28(1):67～80

[4] Hu J and Yan H. A model-based segmentation method for handwritten numeral strings. Computer Vision and Image Understanding,1998,70(3):383～403

[5] Liu J and Tang Y Y. Distributed autonomous agents for Chinese document image segmentation. Int'l J. of Pattern Recognition and Artificial Intelligence,1998,12(1):97～118

[6] Hong C et al. Segmentation and Recognition of continuous handwriting Chinese text. Int'l J. of Pattern Recognition and Artificial Intelligence,1998,12(2):223～232

[7] 韩布新. 部件组合-潜在的汉字结构层次. 中文信息学报,1995,9(3):27～32

[8] 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. 上海教育出版社,1993:108～169



吕 岳 1968 年生,博士生,工程师,曾获邮电部科技进步一等奖一项,三等奖两项,发表论文十余篇,研究方向有图像处理、文字识别、文本分析及智能系统等.



施鹏飞 1939 年生,教授,博导,所长,IEEE 高级会员,中国人工智能学会理事,长期从事图像处理、模式识别及人工智能方面的科研和教学工作.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net