第16卷 第6期
2006年6月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 6
Jun. 2006

关于手写汉字切分方法的思考

邵 洁,成 瑜
(南京航空航天大学,江苏 南京 210016)

摘 要:汉字切分是汉字识别系统中必不可少的组成部分,但由于手写体汉字的书写多变而随意,极大地增加了汉字分割的难度。文中回顾了近十年来脱机手写体汉字分割的发展历程及在发展中涌现的一些主要类型的切分方法,分析了每一类方法的优缺点及其包含的各个分支。最后,参考各种方法的优缺点,对今后的手写体汉字分割发展方向进行了展望。

关键词:手写体汉字分割;直方图;基于识别的分割
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2006)06-0184-03

A Survey of Methods in Handwritten Chinese Character Segmentation

SHAO Jie, CHENG Yu
(Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: Chinese character segmentation has a significant role in off-line handwritten Chinese character recognition system. This paper provides a review of the methods and strategies in character segmentation. It contrasts merits and weaknesses of these methods, and especially introduces the recognition-based segmentation. In the end, some suggestions are listed to improve the development of Chinese character segmentation.

Key words: handwritten Chinese character segmentation; projection; recognition-based segmentation

0 概 述

脱机手写体汉字识别是模式识别领域中一个极具挑战性的难题[1]。汉字分割是将扫描图像中的汉字句段分解成孤立汉字的过程,是汉字识别系统中影响识别效果的重要因素。在脱机单字识别日趋走向成熟的今天,汉字识别系统作为产品推向社会成为可能,它将在信函分拣、银行支票识别、统计报表处理以及手写文稿的自动输入等诸多方面发挥巨大的作用。然而,手写体汉字的书写随意性很大,相邻汉字之间的位置关系也复杂多样,因此,汉字切分成为识别系统中一个不可避免的步骤,是现阶段自由体汉字识别走向实用阶段的重要障碍之一[2]。

手写体汉字的书写可能产生如下6种位置排列情况[3](如图1所示):

①正常:汉字各自分开独立为整体;

②粘连:汉字的某一笔在一点或几点与相邻汉字接触;

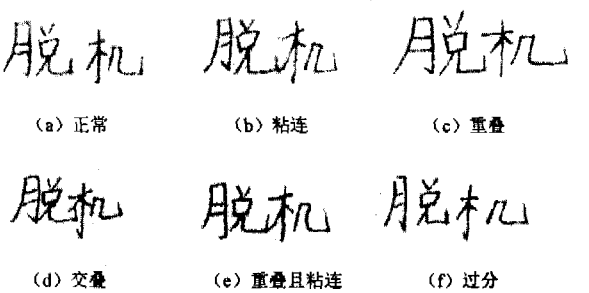
③重叠:汉字间无接触,但无法用垂直分割线分割;

④交叠:两个汉字共享某一部分像素区域,不仅仅个别几点相连;

⑤粘连且重叠:粘连与重叠情况并存;

⑥过分:汉字左右部分间距过大或汉字内部出现笔画断裂。

其中,交叠不常见,但非常难处理。重叠和粘连的情况非常常见,但可以寻找到很多相关算法。过分是较少被提及的一种情况,因为很少出现,所以也没有什么专门针对的算法。若书写中几种情况同时并存,就大大增加了分割的难度,是今后研究手写体汉字切分的重点和难点。



(a) 正常 (b) 粘连 (c) 重叠

(d) 交叠 (e) 重叠且粘连 (f) 过分

图1 手写体汉字书写位置关系

文中参考了1995年至今脱机手写体字符分割的主要相关文献,回顾和总结了脱机手写体汉字近年来的发展。

1 基于汉字整体认识的分割方法

在西文字符分割中,自20世纪50、60年代起,就有一种寻找字符间空白区域以划分不同字符的分割思想,直方图投影分割法可以说是这种思想的延续和方法的更新,是最早被广泛使用的一种汉字分割方法。这类方法算法简

收稿日期:2005-10-09
作者简介:邵 洁(1981-),女,江苏南京人,硕士研究生,主要研究方向为图像处理与计算机视觉;成 瑜,教授,研究方向为图像处理与计算机视觉。

单,运行速度快,并且对间距较大规整书写的汉字也有相当好的分割效果。现在,直方图投影分割[4]被广泛用于无粘连手写字符或印刷体字符的粗分割[5~8]。一旦字符粘连或重叠,垂直投影图常常在最佳分割点上表现为最小投影值,在文献[5,6]中,还采用了宽度递归算法。首先通过垂直投影将相互分开的字符分离,通过计算已分字符的高宽比划分粘连字符。当相邻汉字重叠或粘连部分只有一、两笔时,使高宽比为平均值的分割点往往恰好在投影最小点的周围,两者共同作为分割依据协调考虑。宽度递归算法来源于观察人们的书写习惯,汉字书写宽度总是相对固定的。它还有许多相关变通的方法,用来应对不同使用环境下的变化。对于过分汉字,可以利用高宽比值划分,但正确率不高。对于正常书写字符,在平均高宽比决定点周围寻找投影最小点画分割线往往是最佳选择。

直方图投影与宽度递归法结合分割字符,简单且快速,拓宽了可分汉字的类型,提高了分割率。但它们只适用于均衡字体或印刷手写体汉字的分割,一旦笔画宽度改变、汉字相互重叠或粘连,就无法产生很好的效果。

2　像素跟踪法

像素跟踪是数字图像处理领域的基本方法之一,它利用二值图像的特殊性,跟踪黑像素得到汉字笔画,是一种倾向于结构分解的分割方法。像素跟踪法对于笔画宽度和汉字间的位置都不敏感,它非常适合无粘连字符的切分。通过判断近邻像素连通区域的相互关系、大小及比率,可以简单而精确地区分单个字符。

连通域分析法(Bounding Box)[9]采用3*3模板,首先在汉字图像上找到一个黑像素点作为模板中心,再追寻中心点相邻黑像素点逐一移动,直至无可移动点,称为跟踪笔画,由此产生原始连通域。开始这种方法被用于手写体邮政编码和脱机手写英文的分割[9],取得了很好的效果,可以达到90%以上的分割正确率。在直方图投影法与像素跟踪法的分割比较中,采用直方图投影法的错误率是像素跟踪法的两倍,而时间却是后者的四倍。与邮编和手写英文不同的是,一个汉字字符常常包含多个连通域,则需要采用模糊匹配算法将其合并。

(a) 正确　　(b) 错误

图2　连通域单元和穿越算法的切分

针对连通域法无法分割粘连汉字的问题,王琳琬等[10]提出了基于连通域单元和改进穿越算法的汉字切分。用 $m-1$ 条直线将粘连单元横向分成高度均匀的 m 格,从左向右进行列扫描,计算每一列上交叉点与前景像素点重合的点的个数,寻找最佳切分点。从实验结果看,该算法不仅可以切分粘连字符,还解决了部分粘连位置在竖直方向上可能存在多个笔画的问题,但对于粘连过紧密的字符仍不能正确切分(如图2所示)。

3　基于汉字笔画结构的分割方法

虽然以上两类方法较好地解决了无粘连汉字的分割,但粘连汉字的分割依然是困扰人们的难题。汉字是由具有一定排列规律的笔画组合而成的,通常情况下,使用以识别为基础的判别规则可以将不属于同一汉字的笔画剔出,以及区分粘连笔画和单一笔画。因此,采用笔画提取再合并的方法可以从另一个角度解决笔画粘连问题。

Tseng[11]的文章中提到了一种笔画连接盒(Stroke Bounding Box)的动态算法。首先利用四方向法提取汉字的横竖撇捺,并将其分长短笔画共得到八类,每一类笔画都采用外接矩形描述,称为笔画连接盒。将过长或不合理的笔画分割开,最后利用基于识别的动态算法合并连接盒,分割效果如图3所示。文中提及,在由50个人手写的900个汉字和100个印刷体汉字的组成的样本测试中显示,对于重叠或粘连字有99%的正确分割率,对粘连且重叠的汉字为96%,过分汉字为95%。

图3　笔画连接盒算法效果

另一种基于笔画的分割方法[12]通过黑游程跟踪法提取笔画,笔画提取前计算了黑像素游程宽度、字符的平均宽度(通过垂直投影法得到)和高度。首先从图像中寻找到一条黑游程,作为笔画的开始,然后对该黑游程进行逐行跟踪,在当前黑游程的下一行左右的一定范围内,找到所有的黑游程,并根据已有的游程平均宽度和游程直线拟合得到的笔画方向,确定归入该笔画的黑游程,并确定出下一行的跟踪范围,直到找不到新的游程,跟踪结束,得到一个笔画。从图像中提取的笔画分别用外接矩形和凸包描述,采用一定的算法合并。在文中,由于分割对象是信函地址,笔者将此方法作为预切分的手段,之后采用基于识别的最优路径动态规划算法决定分割路径。

基于笔画提取的分割方法在很大程度上依赖于笔画的提取优劣程度,至今,笔画提取主要有3种方法,分别基于二值图像、细化图像和汉字轮廓。基于二值图像的笔画提取省略了对图像的进一步处理,所以分割时大多采用这种笔画提取法。然而,这种笔画提取的方法还没有达到很高的水平,对于横不平竖不直的汉字提取效果不佳。因此,这种方法的分割正确率也受到限制,适用范围不够广。此外,笔画先提取后合并使算法过于复杂,将其作为垂直投影后的细切分可以更简单省时。

4　基于识别的分割方法

在尝试了诸多基于结构的分割方法后,人类视觉的感知模式越来越成为模式识别领域人们研究的重点。人们用眼睛可以识别各种文字和图像,而忽略其中的变形或干

扰,即使不识字的孩童也能够将独立的字从字符串中提取出来。基于结构的分割方法是有效的,却不是最终发展方向和解决方法。理论和实践都证明基于识别的统计分割方法是汉字分割的新的出路[2]。

背景细化(如图 4 所示)是一种可靠而直观的字符切分方法。在文献[7,8,13,14]中均有提及。细化是数字形态学方法,在文献[15]中有详细描述,但将背景细化协助分割始见于手写体数字分割[13]。为了得到一个比较完整的背景轮廓,需要将字符归一化到四周只留下较小空白区域的矩形位图中。首先图像二值化,再采用 Hilditch 的细化算法[15]提取背景区域的骨架。文献[13]给出了背景骨架的分段和特征点定义。利用细化背景分割字符的关键是找到背景骨架中的交叉点和拐点。无粘连字的分割路径是背景骨架中的一段,其中不存在端点,是由交叉点,拐点及两点间连线构成。因此,只要寻找图像顶部背景骨架和底部背景骨架的交叉点间的连线,辅以汉字平均宽度为依据,就可确定无粘连字的最佳切分路径。粘连字的切分稍微复杂一些。首先需要对字符本身进行处理,比如细化[7]或轮廓提取[8]。第二步记录汉字本身粘连笔画间的候选交叉点 a,依据汉字平均宽度确定顶部背景骨架线的交叉点为候选分割线起始点,沿不大于 90 度角的方向向下跟踪背景骨架至 a 点附近上方背景骨架端点。同理确定底部背景骨架线交叉点并向上搜索至 a 点附近下方背景骨架端点。最后连接上下两端点确定分割线。

(a) 背景细化 (b) 分割路径特征点

图 4 汉字背景细化

Viterbi 运算法则[8,16]是一种可以获得最优分割路径的动态算法,它可以产生非线性分割路径,因此非常适合于重叠字符的分割。Tseng[16]详细介绍了 Viterbi 算法在汉字分割中的使用,文中示例为竖向排列文字的分割,可以看出达到了很好的效果。由于 Viterbi 算法对于非粘连字的分割效果明显优于粘连字,所以 Liang[8]将其作为预分割的第二个步骤,配合垂直投影波形图分析,专门应用于重叠字符分割,分割过程如图 5 所示。

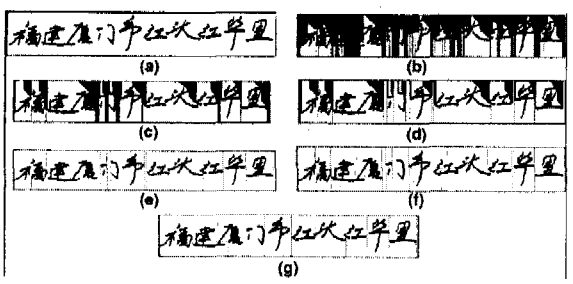


图 5 无粘连字的 Viterbi 算法切分

此外,Casey R G[9]还提到了一种基于神经网络识别器的切分方法。神经网络依次寻找图像中可识别的字符,并自动继承识别结果忽略已识别字符,继续搜索相邻区域的新的字符图像。基于隐马尔可夫模型(HMM)的切分方法[8]是另一种基于识别的切分方法。这两种识别方法都与后期识别密切相关,虽然对于各种复杂情况都有一定的适应性,且可以动态选择分割点,减少分割错误,但由于单字识别耗时且正确率有限,所以这种过分依赖识别的方法仍然有待进一步改善。

5 手写体汉字分割的进展和展望

十几年来,脱机手写体汉字分割取得了令人瞩目的进展和成绩,尤其是特定领域无限定汉字分割,向实际应用迈出了一大步。多种方法的结合使用可以适应多种汉字排列的复杂情况,每种方法具有针对性地解决复杂情况中的单个问题,可以有效地简化算法,提高效率。以文献[8]为例,针对信封封面手写地址分割,首先利用垂直投影和宽度递归算法预处理,再采用 Viterbi 算法分割无粘连字符,然后采用背景细化的方法分割粘连字符,最后应用模糊匹配规则合并不恰当分割。不同算法被最具针对性地应用在了不同场合,分割结果也是很好的,在对包含 7931 个汉字的 921 个自由体汉字地址串识别中,正确率可以达到 87.6%,每个汉字串的识别平均时间为 1.089 秒。

回顾近年来手写体汉字分割的发展,基于识别的分割是一条非常有价值的发展道路,尤其,基于垂直投影波形图分析预分割后,先识别笔画简单的汉字和偏旁部首,并以此为依据合并汉字及寻找切分路径,可以有效避免汉字错分和过分。过分是现阶段比较难解决的问题之一,各种分割方法对于过分汉字都无法达到很高的正确切分率,主要是因为合并部首时采用的模糊匹配算法还存在缺陷,合并法则主要依据切分块的高宽比、两切分块间的空隙宽度等条件判断,对于独体竖型汉字,如“乡”、“页”、“自”等来说,手写体的宽度与一般部首宽度相差不大,有些情况下两汉字间的距离比单个汉字内部间距还窄,这样就很容易产生错分。还有一种合并依据是考虑已切分字块的笔画密度,一般认为部首的笔画较少,相同领域内的笔画密度较小,但“了”、“几”等字明显比部首“身”笔画稀疏。对于将以上各种条件以不同重要性分配权重参数求取和值,与固定阈值比较,决定合并结果的合并算法,参数的选取相当重要,需要经过多次检测决定。不仅如此,判断条件是固定而僵硬的,而汉字的书写和排列却是灵活多变的,如何用死的规则划分活的汉字,是一个值得思考的问题。现阶段,这种规则似乎总是只能应用在某一种书写类型的汉字切分中,这说明规则中使用的汉字特征仍有待改变,还需要进一步寻找更基本的特征组合。

除此以外,曲线分割是汉字分割领域的必然发展趋势,背景细化和 Viterbi 算法的有效性就证明了这一点。由于相邻汉字间关系的复杂性,无论直线或斜线常常都无法正确分割,可以采用统计的方法协助寻找分割线。比如文献[5]利用快速傅里叶变换区别较宽汉字和粘连汉字,

(下转第 190 页)

元素个数

4.3 比较结果

针对抗原 x，比较传统识别方法和云识别方法识别它时的伪否定率，从而证明了云识别能够降低此类伪否定率。

5 小 结

将研究不确定性问题的云模型理论引入到人工免疫学的识别过程，借鉴云模型的思想和数学工具，提出了云识别的算法，用以识别不确定性抗原。相对于传统的识别方法对不确定性抗原的识别，云识别降低了伪肯定率和伪否定率。如果有多个识别器能够成功识别很多这样的不确定性抗原，则可将这些不确定性抗原聚类，形成一个多粒度的数据场，从而实现数据场的由微观到宏观的转变。

参考文献：

[1] Chao D L, Forrest S. Information Immune Systems[A]. International Conference on Artificial Immune Systems (ICARIS)[C]. Canterbury, England: University of Kent at Canterbury Press, 2002. 132 - 140.

[2] González F, Dasgupta D, Kozma R. Combining Negative Selection and Classification Techniques for Anomaly Detection [A]. In Proceedings of the Congress on Evolutionary Computation[C]. Honolulu, HI: IEEE Press, 2002. 705 - 710.

[3] 梁意文. 网络信息安全的免疫模型[D]. 武汉：武汉大学，2002.

[4] Forrest S, Hofmeyr S. Immunology as Information Processing [A]. in Design Principles for the Immune Systems and Other Distributed Autonomous System[C]. UK: Oxford University Press, 2001. 361 - 388.

[5] Hofmeyr S, Forrest S. Architecture for an Artificial Immune System[J]. Evolutionary Computation, 1999, 7(1): 1289 - 1296.

[6] 杜 鹢，李德毅. 基于云的概念划分及其在关联采掘上的应用[J]. 软件学报，2001，12(2)：196 - 203.

[7] Forrest S, Perelson A, Allen L, et al. Self - NonSelf Discrimination in a Computer[A]. Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy[C]. Los Alamos, CA: IEEE Computer Society Press, 1994.

[8] Dasgupta D, Ji Z, González F. Artificial Immune System (AIS) Research in the Last Five Years[A]. in the proceedings of the international conference on Evolutionary Computation Conference(CEC)[C]. Canberra, Australia: [s. n.], 2003. 8 - 12.

[9] Kim J, Bentley P J. Immune Memory in the Dynamic Clonal Selection Algorithm[A]. in the proceeding of the first International Conference on Artificial Immune Systems (ICARIS)[C]. Canterbury, UK: [s. n.], 2002. 9 - 11.

(上接第 186 页)

就是将统计和结构方法结合的很好范例。

当然，不同方法适用于不同领域的分割要求。信函分拣系统中对于汉字分割的要求很高，需要将几种方法结合分割；而文稿自动输入系统则对于输入速度要求更高，可以要求书写者选用方格纸书写，使汉字间不存在重叠或粘连现象，这种情况下采用垂直投影和模糊匹配的方法就完全可以胜任了。

参考文献：

[1] Chen Y. Research on hand - printed Chinese character recognition[D]. Beijing: Tsinghua University, 1997.

[2] 丁晓青. 汉字识别研究回顾[J]. 电子学报，2002，30(9)：1364 - 1368.

[3] 高彦宇，杨 扬. 无约束手写体汉字切分方法综述[J]. 计算机工程，2004，30(5)：144 - 146.

[4] Lu Y. Machine Printed Character Segmentation - An Overview[J]. Pattern Recognition, 1995, 28(1): 67 - 80.

[5] 朱 锴，赵宇明，吴 越. 一种离线手写体汉字切分的自适应算法[J]. 计算机工程与应用，2004(6)：47 - 50.

[6] 陈 强，娄 震，杨静宇. 非限定手写汉字的分割研究[J]. 南京理工大学学报，2004，28(1)：95 - 98.

[7] Zhao Shuyan, Chi Zheru, Shi Penfei, et al. Two - stage segmentation of unconstrained handwritten Chinese characters [J]. Pattern Recognition, 2003, 37: 145 - 156.

[8] Liang Z, Shi P. A metasynthetic approach for segmenting handwritten Chinese character strings[J]. Pattern Recognition Letters, 2005, 26: 1 - 14.

[9] Casey R G, Lecolinet E. A Survey of Method and Strategies in Character Segmentation[J]. IEEE Trans PAMI, 1996, 18(7): 690 - 706.

[10] 王琳琬，杨 扬，颉 斌，等. 基于连通域单元和穿越算法的汉字切分[J]. 信息技术，2004，28(4)：30 - 35.

[11] Tseng Lin Yu, Chen Rung Ching. Segmenting handwritten Chinese character based on heuristic merging of stroke bounding boxes and dynamic programming[J]. Pattern Recognition Letters, 1998, 19: 963 - 973.

[12] 王 嵘，丁晓青，刘长松. 基于笔画合并的手写体信函地址汉字切分识别[J]. 清华大学学报，2004，44(4)：498 - 502.

[13] Lu Zhongkang, Chi Zheru, Wan - Chi Siu, et al. A background - thinning - based approach for separating and recognizing connected handwritten digit strings[J]. Pattern Recognition, 1998, 32: 921 - 933.

[14] 魏湘辉，马少平. 基于凸包像素比特征的粘连汉字切分[J]. 中文信息学报，2005，19(1)：91 - 97.

[15] Gonzalez R C, Woods R E. Digital Image Processing[M]. Boston: Addison - Wesley, 1992.

[16] Tseng Yi - Hong, Lee His - Jian. Recognition - based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm[J]. Pattern Recognition Letters, 1999, 20: 791 - 806.