基于特征行必要-充分性匹配的字符识别方法*

李 佐, 王姝华, 蔡士杰

(南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)
E-mail: lizuo@263.net; sjcai@nju.edu.cn
http:// www.nju.edu.cn

摘要: 字符识别系统的性能在很大程度上依赖于所选取的字符特征.提出了一种基于特征行必要-充分性匹配的 OCR(optical character recognition)方法.该方法使用字符模板的特征行集,通过对待识字符位图进行必要性和充分性双向匹配来识别字符.并采用基线对齐归一化方法在特征匹配时准确定位,使识别率和识别速度都较为理想.同时,对字符位图特征行的交互选择、测试和调整等方法做了详细介绍.另外,还提出了字符骨架与位图重叠显示的方案,有效地提高了对识别结果校对的速度.最后,通过测试和比较,对识别效率进行了分析.

关 键 词: 计算机图形学;字符识别;特征抽取;特征行;必要-充分性匹配
中图法分类号: TP391 文献标识码: A

光学字符识别(optical character recognition,简称 OCR)技术的应用十分广泛,其发展有较长的历史,一些较为成熟的通用软件已经推出.当前,OCR 研究的热点涉及退化比较严重的文字、多字体文字和手写体的识别以及对包括印刷体识别在内的进一步提高识别效率的新方法的研究[1].由于 OCR 系统处理的字符数量一般都很大,尤其是在扫描仪与应用系统并行工作,大批量、流水线地处理文档时,寻找适合具体应用领域的识别方法以进一步提高识别率和速度是必要的.

目前,普遍采用 3 类方法来提高 OCR 系统的识别性能:一类方法是寻找更好的分类识别算法[2,3];另一类方法是将几种分类器结合起来,相互补充,根据不同方面的特征分类[4];第 3 类方法是抽取具有更强描述能力的特征,结合其他辅助特征来描述扫描位图[5~8].

从根本上说,OCR 系统的性能在很大程度上依赖于所选取的特征的高效性,上述的第 1 类和第 2 类方法是建立在第 3 类方法的基础上的.虽然多年以前人们就已经认识到这一点,并做了大量的工作,但近年来仍有不少文章在讨论特征抽取的问题.可见,针对各种应用领域,通过采用进一步优化的具有强描述能力的特征来提高识别性能的方法发展余地还很大.另外,只要存在错识或拒识,就需要为用户提供能够快速发现错误的手段.应改变用户以通读来识别结果并对照原文进行校对的现状,使系统更具实用性.

本文在总结全字符模板匹配的缺陷的基础上,提出一种基于特征行必要-充分性判定的 OCR 方法.该方法通过特征行匹配并采用基线对齐的归一化方法来减少计算量并提高识别率.文章还提出了一种字符骨架与位图重叠的识别结果显示方法,力图充分利用人眼的并行比较能力和提高用户校对速度.最后,通过实验数据的比较对识别效率进行了分析.

1 全字符模板匹配方法的缺陷

字符的标准位图模板包含字符的所有特征,但字符的实际扫描位图与标准位图相比,边缘部分会有一些差

* 收稿日期: 2000-03-21; 修改日期: 2000-08-12

作者简介: 李佐(1967－),男,浙江温州人,博士,主要研究领域为计算机图形学,模式识别;王姝华(1974－),女,江西南昌人,博士,主要研究领域为计算机图形学,模式识别,文档分析理解;蔡士杰(1944－),男,江苏太仓人,教授,博士生导师,主要研究领域为计算机图形学,CAD,人机交互,模式识别.

别.另外,在实际匹配中还可能出现待识别位图与模板错位的现象.因此,需要排除边缘差异给识别带来的困难.

定义 **1**. 匹配结果位图 N 中的非零像素点称为异点.设异点的集合为 P_N,则 $P_N=\{p\neq 0|p\in N\}$.

定义 **2**. 如果异点是由同一笔划退化或错位所造成的,则称为边缘异点;否则,即由不同笔划所造成的,称为笔划异点.

识别实际上是判断有无笔划异点,而有时边缘异点过多会使匹配算法失效(如图 1 所示),从而导致识别率不高.

区分边缘异点和笔划异点要花费较大的代价,无法满足高效率的要求.因此,需要有更理想的识别方法.

Fig.1 Quantity can not distinguish abnormal stroke spots from abnormal edge spots

图 1 从数量上难以区分笔划异点与边缘异点

2 基于特征行的必要-充分性匹配的 OCR 方法

在一个有限的字符集合中,确定一个字符的特征量不应太多,这样既可以减少匹配时间,又可以有目的地避免负作用.位图数据的横向组织决定了以行为单位的计算效率较高.

2.1 基于特征行的必要-充分性匹配算法

定义 **3**. 设$\{li\}$是某一标准字符位图中可用来区别其他字符的一组扫描行(i 为行标),X 为待识位图,Match 为匹配函数,若

$$Match(\{li\},X)=\begin{cases}1 & \text{当且仅当位图为该字符}\\ 0 & \text{位图为其他字符}\end{cases},$$

则称$\{li\}$为该字符的特征行集.

设 M_i 是某字符的标准字符位图的第 i 行,X_i 是待识别字符位图的第 i 行,则 X_i 与 M_i 相同的条件是

$$\sum_j (M_{ij}\oplus X_{ij})=0.$$

由于实际扫描的字符位图相对于标准位图总会有些差别,对应行难以完全相同,因此我们引入特征行的瘦位图 M_i^L 与胖位图 M_i^F 的概念.

定义 **4**. M_i^L 为 M_i 中每一连续显示点段的两端各去掉两点的结果(但必须保证每段最少有一个点), M_i^F 为 M_i 中每一连续显示点段的两端各增加两点的结果,则称 M_i^L 为 M_i 的瘦位图, M_i^F 为 M_i 的胖位图.

此时,X_i 与 M_i 匹配的必要条件和充分条件为

$$N_i=\sum_j[(M_{ij}^L\wedge X_{ij})\oplus M_{ij}^L]=0,$$

$$E_i=\sum_j[(X_{ij}\wedge M_{ij}^F)\oplus X_{ij}]=0.$$

于是,可将特征行的定义重新描述如下:

如果存在$\{M_i\}$,$i=k_1,k_2,\ldots,k_h$,满足:X 与 M 匹配,当且仅当

$$N=\sum_{i=k_1}^{k_h}N_i=\sum_{i=k_1}^{k_h}\sum_j[(M_{ij}^L\wedge X_{ij})\oplus M_{ij}^L]=0,$$

$$E=\sum_{i=k_1}^{k_h}E_i=\sum_{i=k_1}^{k_h}\sum_j[(X_{ij}\wedge M_{ij}^F)\oplus X_{ij}]=0,$$

则$\{M_i\}$为该字符的特征行集.

2.2 特征行的选择与测试

选择特征行的原则是,在应用它们进行匹配测试时,可以使对应字符的异点数量与非对应字符的异点数量差距拉大,即突出对应字符的特征.必须在操作员经验指导下,以交互方式选取特征行,并注意以下几点:

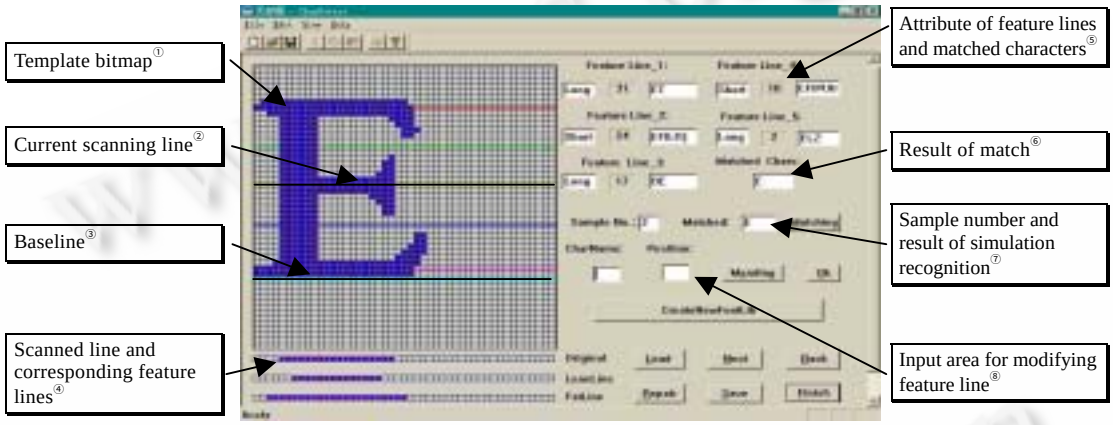
(1) 应反映本字符的典型笔划特征.水平笔划通过与其中心线一致的一条特征行(属性为 long)反映出来,垂直笔划通过能与其产生稳定的截断线的多条特征行(属性为 short)反映出来.

(2) 应反映本字符区别于其他相近字符细节的特征.

(3) 应反映字符骨架特点,避免选择笔划的修饰性部分.

(4) 在选择特征行的数量时,要考虑到字符图像特征的复杂度.由于实际文档常存在退化现象,特征行太少有可能造成误识,太多则会加大分类器的负担.

本文提供专门的工具来实现特征行的选择、抽取及测试工作(如图 2 所示),并在交互指定特征行时对实验样本进行模拟识别.该工具包含以下几个主要功能模块:



①模板位图,②交互指定的当前扫描行,③基线,④当前扫描行及生成的特征行,⑤特征行属性及匹配到的字符集,⑥所有特征行的匹配最终结果,⑦实验样本号和摸拟识别到的字符名,⑧对特征行进行修改时的输入区.

Fig.2 The tool used for selecting, extracting and testing feature lines

图 2 特征行的选择、抽取及测试工具

(1) 模板显示与特征行交互选择模块.该模块在左侧显示区显示模板位图(见图 2 中①),通过鼠标操作指定 3~5 条水平扫描线(见图 2 中②)来确定特征行位置,特征行的数量根据字符特征的复杂程度人为决定.

(2) 特征行抽取与显示模块.该模块按照识别算法的要求,根据当前扫描线位置抽取特征行位图,生成相应的瘦、胖位图并显示出来(见图 2 中④),供开发人员查看.

(3) 特征行属性与匹配结果显示模块.该模块显示每条特征行的属性(长/短行)、位置、该特征行匹配到的字符(见图 2 中⑤)及最终结果(见图 2 中⑥).最终结果是每条特征行的匹配结果的公共子集.

(4) 模拟识别测试模块.它是一个对一组单个字符位图逐个识别的程序.它使用当前选择的特征行对后台实验样本集进行模拟识别测试(每个样本集包含一套所有应能识别的字符),找出匹配成功的字符集.该字符集应该与当前字符模板一一对应(见图 2 中⑦).

(5) 修改特征行输入接口.在图 2 的⑧中输入要修改的字符以后,能够自动显示该字符位图、已有的特征行集以及该特征行集的有关数据,并进入交互修改状态.

(6) 特别编辑处理模块.在修正特征行后期,识别率已接近 99%,遗留问题主要是极相似字符的相互误识,如 1~l,I~l.通过简单地改变特征行条数和位置已经难以取得理想的效果,需要采用特殊的方法.Mending 按钮启动编辑功能,它在图 2 的④中显示指定要修改的特征行的位图、瘦位图和胖位图,开发人员可通过鼠标操作进行位编辑,改变特征行长度,从而改变匹配的约束条件,避免该字符被误识.较为苛刻的约束条件会增加该字符被

拒识的可能性,但在更大程度上减少了误识,因而是有效的.

用户在使用该工具指定特征行时,工具将该行位图加工成瘦位图和胖位图,并对标准位图库中所有字符位图的对应行进行测试,显示与之匹配的字符集,同时,计算并显示现有特征行匹配到的公共子集.当该公共子集中只有一个字符且就是该字符时,可使用模拟识别测试功能对实验样本进行模拟识别,以查看使用当前选定的特征行集的识别效果,如果满意,则将特征行保存起来,否则继续进行调整.

2.3 特征行模板与实际文档字符的匹配及修正

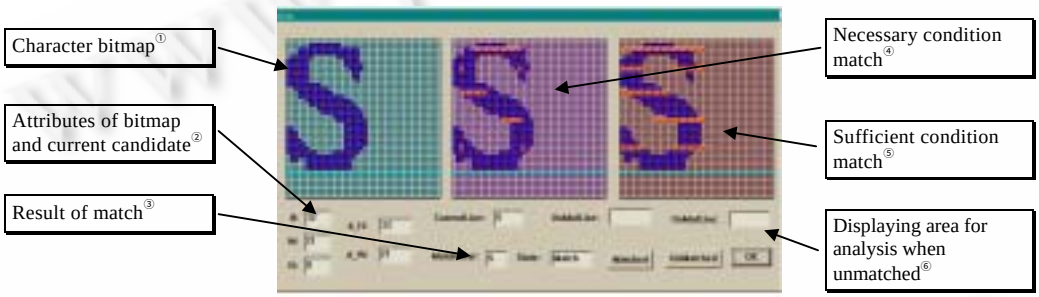
以上工具选定的特征行是根据理论计算和实验测定的结果,同时也包含了开发人员的经验.在对实际文档进行识别时还会遇到一些问题,主要有以下几类:

(1) 某些字符的特征比较相似,如 8～S,P～F,1～l 等,有时会造成相互误识.

(2) 对于容易产生粘连的字符,未考虑到字符分割对其产生的影响,导致特征行匹配对切割较敏感.

(3) 有些字符选取的特征行位置比较敏感,易造成误识或拒识.

为了达到更高的识别率,特征行应经进一步测试和修正.我们使用一个独立的跟踪模块对实际识别过程进行监视,实时跟踪候选字符特征行模板与待识字符图像的匹配过程,并将现场显示出来.同时,还将当前候选字符与待识图像的特征参数和识别结果报告给开发人员,并在匹配失败时显示失败原因(比如,具体哪一个特征行匹配失败,偏差多少等),协助开发人员对实际匹配情况进行分析,完善特征行的选择(如图 3 所示).



①字符位图,②待识图像与当前候选字符的属性参数,③匹配结果,④必要条件匹配,⑤充分条件匹配,⑥匹配失败自动分析显示区.

Fig.3 Display of match from the monitor

图 3 监视模块对匹配情况的显示

开发人员如果发现匹配失误,可以参考该模块的显示信息,使用如图 2 所示的专用工具来修正特征行.修正的主要方法有:增减特征行的条数、改变特征行的位置以及特别编辑处理.

2.4 候选字符的精选

以较少的计算量来优选候选字符集,减少匹配次数,是提高识别性能的必要手段.候选字符的确定主要是通过对字符位图的以下特征进行判别:

(a) 高宽比特征

对于多数字体来说,每一字符扫描位图的闭包框的高度和宽度有一定的规律.由于退化、毛刺以及其他环境因素的影响,可能会造成同一字符的扫描位图的高度和宽度有一定的差别,但是,这种差异应局限在一定的范围之内.

(b) 边界特征

从字符位图的左、右、上、下四个方向扫描,闭包框边界到字符笔划间的空白像素分布构成了字符的方向边界特征.有的字符有其独特的边界特征,有些字符则在某一方向拥有共同的边界特征.

(c) 基线位置特征

找出位图中字符串的基线位置,则在每一字符位图中,有效像素集的闭包框与基线的位置关系构成了字符的基线位置特征.

一旦测定了待识字符位图闭包框的高宽比,根据偏差域值,就可以大大减少候选字符的数量.从左往右扫描

字符的边界特征,可以快速地排除边界差异明显的字符.在确定了字符串基线位置的情况下,利用字符的基线位置特征,能够进一步排除不同基线类型的候选字符.

2.5 基线对齐的归一化方法

归一化是指匹配时待识位图向匹配空间的映射以及与模板的对位,是匹配前必须完成的一项重要工作.如果归一化的对位方式不可靠,则会在匹配时发生特征行偏移,降低识别率.归一化的方法通常有重心归一化和外框归一化两种.重心计算是整体性的,因此抗干扰性强,但计算量大.而边框计算是局部的,计算量小,但抗干扰性差.我们结合应用环境和特征行匹配的特点,提出了基线归一化方法,以解决抗干扰性与计算量之间的矛盾.

基线归一化方法的指导思想是,待识字符位图是整页位图的一部分,应充分利用其周围的信息,而不应将其视为孤立的个体.一个字符往往存在于一行文字中,因而可在文档分析阶段找到该行的基线位置作为待测字符的基线位置,在匹配时用于定位.这种方法可以避免字符局部退化对定位的影响,并减少了计算量.

3 字符骨架与位图的重迭显示

只要识别率达不到 100%,用户就必须对识别结果进行校对.现有的 OCR 系统都提供对照校对界面,但由于识别结果与原始位图显示在不同的区域,在校对过程中,人眼要在位图显示区和识别结果显示区来回移动,其效率较低,且容易疲劳.在大批量处理时,劳动强度非常大.

如果将识别结果字符的骨架以适当的颜色贴到原位图上,依靠人眼的并行观察能力,则可以很方便地发现不一致的地方,使“一目十行”成为可能(如图 4 所示).在修改识别结果时,只需在用鼠标点击误识字符后输入正确字符,即可自动搜索图像闭包盒并实现替换.这样就大幅度地提高了用户的工作效率,减轻了劳动强度.

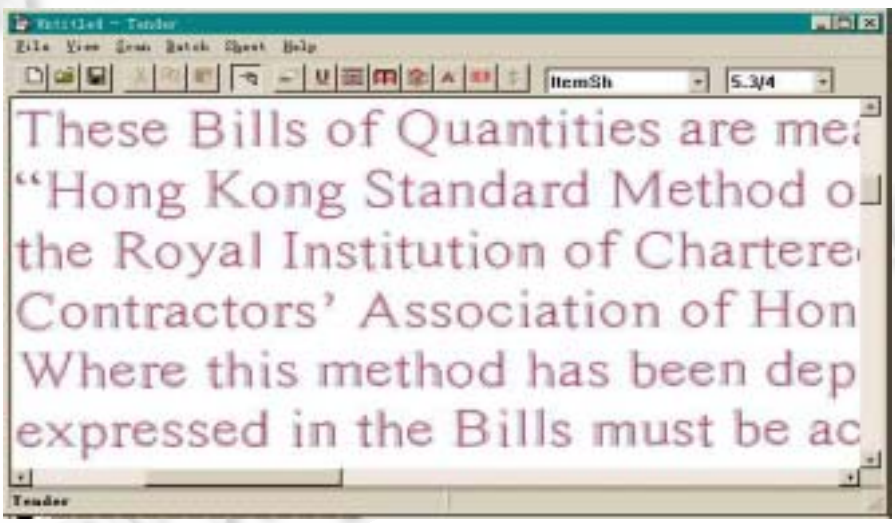


Fig.4 Displaying skeletons of recognized characters together with the original bitmap

图 4 字符骨架与位图的重迭显示方法

4 实验结果与性能分析

我们把依靠上述方法开发的标书处理系统 VHTender 与通用 OCR 系统 OmniPage Pro 10(OPP10)演示版进行了相同样本的识别比较.使用机型为 PⅡ233,内存为 128M.

OPP10 是 ScanSoft 公司的最新产品.我们之所以选择该软件是出于以下两方面原因:

(1) OmniPage Pro 系列是著名的 OCR 软件,发展时间较长.ScanSoft 公司的网页上称 OPP10 是当今世界上最好的 OCR 系统之一.

(2) 我们使用的样本是工程标书,常常出现下划线.不少 OCR 系统的文档分析能力较弱,无法处理存在下划

线尤其是与字符粘连的下划线的文本.而在 OPP10 中,下划线对字符识别的干扰不大.

实验使用了 12 页标书页面作为样本,统计字符共 12 786 个,两个系统的识别情况见表 1.

Table 1 Performance comparison between VHTender and OPP 10(demo)

表 **1** VHTender 与 OPP10(demo)的识别性能比较

Item① / System②	Quantity of error③	Recognition rate④ (%)	Average processing speed (s/page)⑤	Average recognition speed (s/page)⑥	Average speed of match function (s/times)⑦	Average speed of collation (min./page)⑧
OPP 10 (Demo)	88	99.31	1.95	N/A⑨	N/A	2.2
VHTender	41	99.68	2.49	0.51	1.78×10^{-4}	0.7

①数据项,②系统名称,③识别错误数,④识别率,⑤平均处理速度,⑥平均字符识别速度(秒/页),
⑦匹配函数平均执行速度(秒/次),⑧平均校对及修改速度(分/页),⑨无法测量.

对两个系统识别错误的字符,我们参照原图像逐个进行了原因分析,结果见表 2.

Table 2 Analysis about the reason of error

表 **2** 识别错误原因分析

Reason of error② / System name①	Similar shape③	Merged④	Degraded⑤	Unknown⑥	Total⑦
OPP 10 (Demo)	57	12	7	12	88
VHTender	5	25	11	0	41

①系统名称,②错误原因,③外形相似,④字线粘连和字字粘连,⑤字符退化,⑥不详,⑦合计.

通过对以上数据的分析可以得出以下结论:

(1) OPP10 在字线粘连和字字粘连的分割技术上性能卓越,而对于独立字符的分辨力则不如 VHTender,主要是极相似字符之间(如 l~l~I)的误识多一些,最终 VHTender 在总体识别率上比 OPP10 略高一些.

(2) 两个系统的页面处理速度存在一些差距,但由于 VHTender 是面向工程标书的文档处理系统,包含了文档分析和文档理解的功能,用于匹配识别的时间占五分之一.虽然我们无法准确地得到 OPP10 的字符识别时间,但从功能上全面分析,两系统的识别速度应该是比较接近的.

(3) 由于采用了字符骨架与位图重迭的显示方法,VHTender 的校对时间较短.

5 结束语

本文提出的 OCR 方法通过基于特征行的必要-充分性测试快速、有效地识别字符.通过采用骨架-位图重叠显示的方法改善校对环境、提高系统的实用性.该方法已成功地应用于工程标书处理系统 VHTender 中.

References:

[1] Trier, O.D., Jain, A.K., Taxt, T. Feature extraction methods for character recognition——a survey. Pattern Recognition, 1996,29(4): 641~662.

[2] Cho, S.-B. Recognition of unconstrained handwritten numerals by double self-organizing neural network. In: Proceedings of the 13th International Conference on Pattern Recognition. Vienna: IEEE Press, 1996. 426~430.

[3] Chuang, Chen-Tsun, Tseng, Lin-Yu. A heuristic algorithm for the recognition of printed Chinese characters. IEEE Transactions on Systems, Man, and Cybernetics, 1995,25(4):710~717.

[4] Ho, T.K., Hull, J.J., Srihari, S.N. Decision combination in multiple classifier system. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994,16(1):66~75.

[5] Oh, II-Seok, Lee, Jin-Seon, Suen, Ching Y. Analysis of class separation and combination of class-dependent features for handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999,21(10):1089~1094.

[6] Oh, Il Seok, Suen, Ching Y. A feature for character recognition base on directional distance distribution. In: Werner, B., ed. Proceedings of the 4th International Conference on Document Analysis and Recognition. Ulm: IEEE Computer Society, 1997. 288~292.

[7] Yamada, Keiji. Non-Uniformly sampled feature extraction method for Kanji character recognition. In: Werner, B., ed. Proceedings of the 4th International Conference on Document Analysis and Recognition. Ulm: IEEE Computer Society, 1997. 200~205.

[8] Wang, Shu-hua, Li, Zuo, Yang, Ruo-yu, *et al*. A document image understanding system for tender. In: Lü, Jian, ed. Proceedings of the International Symposium on Future Software Technology'99. Tokyo: Software Engineers Association, 1999. 360~362.

A Character Recognition Approach Based on Feature Line Necessary-Sufficient Condition Detection*

LI Zuo, WANG Shu-hua, CAI Shi-jie

(*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093, *China*)

E-mail: lizuo@263.net; sjcai@nju.edu.cn

http://www.nju.edu.cn

Abstract: The performance of a character recognition system depends heavily on what features are being used. In this paper, an optical character recognition method based on match of feature lines is presented. By extracting the feature lines from bitmap of character and detecting the necessary-sufficient condition with templates by baseline superposed, this method carries out the recognition with a very high efficiency. The process of selecting and adjusting the feature lines from template is also described. Then, a new way for collating is proposed. By the hand of displaying skeletons of recognized characters that overlap the original bitmap, this method makes finding errors easier. Finally, the performance evaluation based on comparison is given.

Key words: computer graphics; character recognition; feature extraction; feature line; necessary-sufficient condition match

* Received March 21, 2000; accepted August 12, 2000