一种基于词片识别的字符分割算法

岳思聪 王 庆 赵荣椿
(西北工业大学计算机学院，西安 710072)

摘 要 在字符识别领域，对粘连字符的识别是一个被广泛关注的技术难点，而且粘连字符的分割更是产生识别错误的主要原因之一。为了快速准确地进行字符分割，在总结已有方法的特点及不足的基础上，针对电子阅读笔系统的工作特点和实时性要求，提出并实现了一种面向电子阅读笔系统的基于词片识别的分割算法。该方法由于通过对字母组合的识别，降低了传统的基于孤立字符识别方法对于字符切分的要求，而且以中心生长法和改进的峰谷函数为切分工具来进行字符分割，简单实用，因而其在减少因粘连字符切分错误引起的识别错误的同时，不仅降低了运算复杂度，而且适合在阅读笔等嵌入式设备上应用。实验证明，该算法不仅效率高，而且实现简单，还能够降低分割错误带来的识别错误。

关键词 字符分割 词片识别 电子阅读笔
中图法分类号：TP391.4 **文献标识码**：A **文章编号**：1006-8961(2006)01-0008-05

An Optimal Character Segmentation Algorithm Based on Connected Component Recognition

YUE Si-cong, WANG Qing, ZHAO Rong-chun
(*School of Computer Science, Northwestern Polytechnical University, Xi'an* 710072)

Abstract Segmentation of merged characters is one of difficulties that have attracted a great deal of attention in optical character recognition(OCR). Nowadays, unsuitable segmentation is the primary cause of recognition errors. Based on the analysis of the shortcomings of some traditional algorithms for printed character segmentation, we notice that it is necessary to propose a fine method to meet with the requirement of real time processing and the characteristics of DSP module including relative low power and small memory comparing with PC. In this paper a new algorithm of segmentation and recognition based on connected component is proposed which can be used for electronic reading-pen. The proposed method reduces computation time by recognizing the connected component as a whole. It segments connected component by middle expansion method and peak-paddle function. As a result, recognition error arose by segmentation error can be reduced. Experiment results have proved that the algorithm is effective, easy to implement and it is reasonable and applicable for to ERpen.

Keywords character segmentation, connected component recognition, electronic reading-pen

1 引 言

当前，脱机字符识别(off-line character recognition, OCR)的研究热点虽然已经转向手写体字符的分割与识别[1,2]，但是在印刷体字符识别领域仍然存在一定问题，尤其是由分割错误引起的识别错误更为突出。鉴于如今即使性能很好的分割方法，也不能保证100%的切分正确，总会有一定的切分错误的风险，而且由于这些方法比较复杂，时间复杂度高，因此不适合直接应用在处理能力相对较低、实时性要求高的嵌入式设备上。

基金项目：国家自然科学基金项目(60403008)；陕西省自然科学基金项目(98K07-J2)
收稿日期：2004-11-09；**改回日期**：2005-03-20

第一作者简介：岳思聪(1979～)，男。2005年获西北工业大学硕士学位，现为西北工业大学计算机学院博士研究生。主要研究方向为文字识别、图像处理和计算机视觉等领域。E-mail：yuesicong@mail.nwpu.edu.cn

现有的字符分割方法主要分为以下几类:

(1)基于图像分析的分割[2-4]

这类方法的基本思想是通过图像分析来寻找字符之间合理的分割点,其中最具代表性的方法是静态的投影分析方法[2]、基于前景背景分析(background and foreground analysis, BFA)的切分方法[3]和基于轮廓的切分方法[4]等。

投影分析及相关的改进算法是通过统计字符串图像每一列的黑像素在水平方向上的投影个数来查找连续字符之间的空白区域和粘连区域,以确定分割点的位置。其优点是实现简单、速度快,但这种方法的不足也很明显,即它只能适用于垂直方向存在空白区域的字符,即使利用腐蚀运算或者其他的改进算法,也只能处理部分粘连字符。而 BFA 方法则由于用了前景和背景的分析,因此需要先对背景进行细化,再通过跟踪背景骨架来找到分割路径,由于这种方法的运算量巨大,因此不能满足实时性要求。基于轮廓的方法是先得到字符的外轮廓,然后用轮廓的预测模型或者傅里叶描述子来进行粘连字符的判断和分割,由于这种方法的计算很复杂,处理速度比较慢,故也不能满足实时性要求。

(2)基于识别的分割[5,6]

这类方法是首先通过图像分析来产生可能的分割位置,然后借助分类器的识别能力对各种可能存在的分割进行筛选,以便选择出其中合理的分割,这类方法中,Bayer 和 Dey 的递归分割算法[5,6]最具代表性。该算法是先通过粗略的图像分析寻找所有可能的切点,再采用矩形浮动窗口(窗口左边界的位置固定,右边界位置随不同的候选切点而变化)对窗口内的子图像依次进行识别;然后每识别出一个字符,就将窗口中的子图切掉,再对剩余图像继续进行识别。如果剩余图像无法识别,则可以向前回溯,直到每个窗口中的子图像都能找到匹配原型为止。

这些方法具有动态选择分割点的能力,其不足之处是效率不高,为了避免遗漏正确的切点,应将筛选分割点的条件放得较宽,但这样不仅增加了分类器识别的次数,而且可能发生多解现象,比如 mm 被识别成 rnm 等。同时由于算法的时间开销很大,故不适合实时应用。

(3) 整体识别[7]

以整个词作为待识别对象,即根据词的整体特征来识别,以避免分割对字符的损伤,这种方法一般在识别数量有限的关键性词汇时可以使用,但是如果是对大词汇量的词库进行识别,则这种方法将无法直接使用。其原因是词汇量的增大,使得分类器的输出种类数相应地增加,这势必会造成分类器性能的下降。

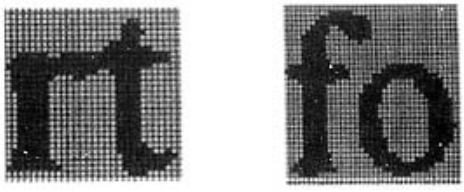
上述方法可以概括为以下两大类:第 1 类算法是先把单词分割成单个字符,然后通过识别孤立字符来达到对单词的识别;第 2 类不用分割任何字符,可直接把单词作为整体进行识别,由于两者都有不足之处,前者引起的分割错误是不可避免的,而后者则不适合大词汇量的识别。

本文设计的电子阅读笔(electronic reading-pen, ERP)是一种完全独立于计算机工作,以 DSP 为核心处理器,这是一个具有图像扫描、中英文识别以及中英文互译功能的小型嵌入式系统。笔者针对嵌入式系统的这些特点,在研究了现有多种方法的基础上,将上面两种方法结合,提出了一种基于词片识别的分割方法,有效地提高了分割结果的可靠性。

2　基于词片识别的分割算法

通过对基于投影分析方法得到的字符分割结果进行的分析发现,绝大多数的粘连块包含两个或者 3 个字符,多于 3 个字符的粘连块较少出现,而且多于 3 个字符的粘连块比较容易进一步切分成更小的字符块。通过统计发现,常出现的粘连多字符组合不超过 200 种,这个类别数目对于分类器并不是很难实现。根据这些实验结论,本文对分割算法的要求放宽,即不必将每个字符都分割出来,只要分割成词片就可达到分割的目的(词片就是包含一个字符或者两个粘连字符的分割块),而且这些词片可以用整体识别的方法来处理,这就是基于词片识别的分割算法的基本思想。

字符粘连的分类如图 1 所示,其中简单粘连是指字符相互接触,但闭包盒没有重叠的粘连,其粘连



(a) 简单粘连字符　　(b) 交错粘连字符

图 1　字母粘连分类

Fig. 1　Connected characters

位置如被准确探明，则能够采用垂直切割的方法将图像分成两个包含完整字符的子图像；交错粘连是一种逻辑上的粘连情况，其中两个字符的黑像素区并没有接触，属于不同的连通区，但由于垂直投影及闭包盒存在重叠现象，因此对此种粘连进行垂直切割无法找到合适的位置。

2.1 分割的基本方法

分割的基本方法是首先用中心生长法将无粘连和交错粘连（图 1）的字符块切分开，因为此种分割方法的代价函数的损失要求为 0，所以没有分割风险；然后利用改进的峰谷投影比函数将宽高比超过限制的粘连块分割为更小的词片，并使得词片中只包含单个字符或两个粘连字符，实验中，由于将改进的峰谷投影比函数的阈值设置较高，因此只有粘连比较薄的地方才会被切开，这一步的可靠性也很高；最后对单个字符进行识别，并把粘连双字符块作为整体进行识别，同时对于不容易切分的粘连词片不做切分处理，而当作一个整体进行识别，这样就可以减少由于切分错误而造成的识别错误。

2.2 中心生长法

中心生长法是对有限动态规划算法（limited dynamic programming，LDP）的合理简化[8]，它是一种基于连通区域分析的方法。该方法首先确定候选分割点，为了不遗漏每个可能的分割点，它是以单词的水平中心线上的 0 值线段的中点作为候选切分点（如图 2 所示），并且以这些点为搜索起始点，之所以没有从图像的上边界或者下边界开始搜索，这主要是为了防止非法路径的出现。

图 2 候选切分点示意图

Fig. 2 Candidata for segmentation

该方法的搜索空间和搜索策略：搜索空间限制在以当前点为中心的上（或下）一行的 $1 \times m$ 的窗内，m 一般取 5。对搜索空间的限制主要是为了防止不正确的切分路径出现，以缩小搜索范围和降低搜索运算量。搜索先从候选分割点出发，在限定窗内搜索 0 值线段，并且要求这个线段和当前点所在的 0 值线段能够连通，然后将搜索到的 0 值线段的中点作为分割路径上的点，记搜索路径经过的 n 个像素点为 $p_1, p_2, \cdots, p_n$，第 i 个像素点 p_i 的坐标为 (x_i, y_i)。由于字符轨迹边缘具有光滑性，所以只要字符间存在空白间隔，那么就一定可以在窗口内找到 0 值线段，而且用 0 值线段的中点组成的路径，近似等价于空白间隔的中心线，这条分割路径具有最好的分割效果。

该方法的代价函数：传统的 LDP 算法是采用 3 种不同的代价函数（穿越黑像素数、穿越笔划数和路径曲率）的组合来作为选择路径的代价函数。在本文的算法中，虽没有使用路径曲率作为代价函数，但是在搜索路径时，则选取最接近字母间空白间隙中点的点作为分割路径。另外，本文也没有使用穿越笔划数作为代价函数，因为采用基于词片识别的方法，不需要用这种方法来分割粘连字符。本文将代价函数 C 定义为搜索路径上的前景像素个数，即

$$C = \sum_{i=1}^{n} p_i \tag{1}$$

其中，p_i 为第 i 个像素点的值。

若存在一条路径，使得候选分割点能够沿着空白间隙穿越图像空间，从分割点到达单词的上下边界，则这条路径就是所要找的满足代价函数为 0（即满足式（2））的路径

$$\begin{cases} l = y_{\text{down}} - y_{\text{up}} \\ C = 0 \end{cases} \tag{2}$$

其中，l 表示路径长度，y_{up} 和 y_{down} 为单词的上下边界的纵坐标。

因为只有代价函数为 0 的路径才是绝对不会错误的路径。若不满足式（2），则停止搜索，且该候选分割点被排除。该方法不同于动态规划，即不会出现多种可能的路径，由于路径上每一个点的扩展都不是任意的，而是确定的，这样就大大减少了计算量，从而使得嵌入式应用变得可行。

2.3 改进的峰谷投影比值函数

上述的方法对于简单粘连（图 1）的情况仍是无效的。由于词片识别允许包含两个粘连字符的分割块，因此包含更多字符的分割块还需要用改进的峰谷投影比值函数 $\hat{f}(m)$ 做进一步的切分。传统的峰谷投影比值函数[9] $f(m)$ 定义为

$$f(m) = \frac{P(m-1) - 2P(m) + P(m+1)}{P(m)} \tag{3}$$

改进的峰谷投影比值函数 $\hat{f}(m)$ 定义为

$$\hat{f}(m)=\left(\frac{(P(m+1)-P(m))-(P(m)-P(m-1))}{P(m)}\right)^{a}$$
$$=\left(\frac{P(m+1)-P(m-1)}{P(m)}\right)^{a} \quad (4)$$

其中，$P(m)$是投影直方图中 m 列的投影值，$P(m-1)$和 $P(m+1)$是其相邻列的投影值。a 是指数常数(通常取 2)，它可以使得分割点的峰值更突出，相比而言，式(4)更能体现出直方图中峰谷之间的陡峭变化。

2.4　基于词片的识别

本文算法提出了一个新的策略，即在模板字典中增加双字符和多字符模板，使得无法切分的字符块可以被整体识别，称为词片识别。由于这种模板只是针对出现概率较大的粘连字符组合而设计的，如果组合太多，则会影响单个字母模板的识别效果，因此该方法进一步提高了系统的分割有效性，并使得分割算法得到简化。部分模板如图 3 所示。

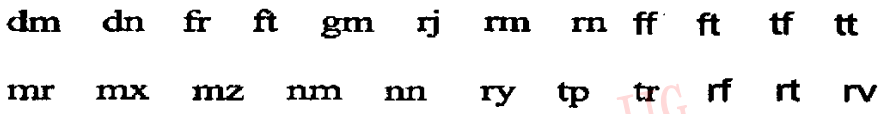


图 3　部分双字符模板

Fig. 3　Some templates of two connected characters

3　实验结果与分析

为验证本文分割方法的效果，选择几种不同字体字符，对基于孤立字符和基于词片两种识别方法进行了识别实验，结果见表 1，表 2。实验结果表明，本文提出的切分方法，与前面提到的几种算法相比，计算复杂度低、内存开支小，且实现方便易行，切分结果(如图 4 所示)表明，该方法能够快速高效的完成字符的切分，并能满足实时处理的要求，这就为后续的识别奠定了基础。实验时，系统先采用基于模板匹配的多级分类的识别方法[10]进行识别，最后利用综合置信度评价函数来确定正选字。

表 1　PC 机仿真对比测试结果(随机抽取的 5000 单词，5 号英文)

Tab. 1　Recognition results by different method on PC

识别方法	识别结果	字体			
		新罗马体	Arial 体	哥特体	宋体
基于孤立字符识别	错误数/单词总数	2 613/5 000	3 438/5 000	3 060/5 000	452/5 000
	识别率(%)	47.7	31.2	38.8	90.9
基于词片识别	错误数/单词总数	271/5 000	87/5 000	73/5 000	116/5 000
	识别率(%)	94.6	98.3	98.5	97.7

表 2　电子阅读笔实际测试结果
(随机抽取的 500 单词，5 号英文)

Tab. 2　Recognition results on ER-Pen

识别结果	字体			
	新罗马体	Arial 体	哥特体	宋体
错误数/单词总数	57/500	36/500	33/500	41/500
识别率(%)	88.6	92.8	93.4	91.8

在 PC 机上分别对两种分割识别方法进行了仿真对比测试，在阅读笔系统上通过实际测试本文提出的方法来验证算法的有效性。仿真测试的实验样本为英文新罗马体，Arial 体，哥特体和宋体 4 种字体各5 000单词(从词典库中随机抽取)的 5 号字打印文本，质量较好，再用扫描仪以 300dpi 扫描成图片作为测试样本。实际测试的样本为日常使用的正式出版书籍，字号一般为 5 号，样本质量较好，随机取其中 500 个单词(包含于词典库)作为测试样本。

仿真测试的目的是为了验证算法的有效性，它是用相同的算法进行大规模测试，而不用考虑时间和空间开销；而在实际测试中，由于英文的平均识别

图 4 切分结果

Fig. 4 Results of segmentation

速度达 3 ~ 4 词/s,因此能够满足实时性要求。从表 1、表 2 对比发现,传统的基于孤立字符识别的方法必须配合使用较为复杂的粘连字符分割方法才会有效,而本文提出的基于词片识别的切分方法则能有效地降低切分的复杂程度,同时能获得较好的识别效果。实际测试比仿真测试的识别率低一些,这是因为实际测试的扫描过程受到人手抖动的影响,其扫描的图像会发生一定的变形,所以导致识别率下降。这和分类器也有一定关系,目前实验中使用的简单的多级分类器只是为了验证基于词片识别的分割算法是否可行,其实可以使用鲁棒性更好的分类器,以便可以提高实际的识别率。

5 结 论

本文提出了一种面向电子阅读笔系统的字符分割识别算法。这种分割方法把切分和识别辩证统一地看待,即通过保留不容易切分的词片,在减小分割错误风险的基础上,引入基于词片识别的分类器来降低分割错误引起的识别错误,由于该方法没有利用识别结果的反馈来指导分割位置的确定,因此不会出现多解现象。该方法已在实验系统中取得很好的结果,不仅提高了分割算法的实用性和速度,而且降低了切分错误带来的风险,从而为字符分割在嵌入式系统中的应用开辟了一条新的道路。

参 考 文 献(References)

1 Arica N, Yarman-Vural F T. Optical character recognition for cursive handwriting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, **24**(6): 801 ~ 813.

2 Casey R G, Lecolinet E. A survey of methods and strategies in character segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, **18**(7): 690 ~ 706.

3 CHEN Yi-kai, WANG Jhing-fu. Segmentation of single-or multiple-touching handwritten numeral string using background and foreground analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, **22**(11): 1304 ~ 1317.

4 JUNG Min-chul, SHIN Yong-chul, Srihari S N. Machine printed character segmentation method using side profiles[A]. In: IEEE International Conference on Systems, Man and Cybernetics[C], Tokyo, Japan, 1999: 863 ~ 867.

5 Bayer T, KreBel U, Hammelsbeck M. Segmenting merged characters [J]. Pattern Recognition, 1992, **25**(2): 346 ~ 349.

6 Dey S. Adding feedback to improve segmentation and recognition of handwritten numerals[D]. Cambridge, Massachusettes, USA: MIT, 1999.

7 Akagi T, Hamamura T, Mizutani H, *et al*. Word-matching method based on the projection of the voting matrix[A]. In: Proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition[C], Amsterdam, Holland, 2000: 559 ~ 564.

8 LIU Gang, WEI Feng. A segmentation method of cursive handwritten digit string based on limited dynamic programming[J]. Journal of Beijing University of Posts and Telecommunications, 2003, **26**(1): 14 ~ 18. [刘刚, 魏锋. 基于 LDP 算法的手写数字串切分[J]. 北京邮电大学学报, 2003, **26**(1): 14 ~ 18.]

9 LU Yi, Haist B, Harmon L, *et al*. An accurate and efficient system for segmenting machine-printed text[A]. 5th Advanced Technology Conference U. S. Postal Service[C], Washington, DC, USA, 1992, **3**: A93 ~ A105.

10 REN Jin-chang, ZHAO Rong-chun, ZHANG Wei. A fast and effective algorithm for printed chinese character recognition[J]. Journal of Image and Graphics, 2001, **6**(10): 1011 ~ 1015. [任金昌, 赵荣椿, 张炜. 一种快速有效的印刷体文字识别算法[J]. 中国图象图形学报, 2001, **6**(10): 1011 ~ 1015.]