基于组件合并的手写体汉字串分割*

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)

E-mail: pfshi@ippr.sjtu.edu.cn

http://www.sjtu.edu.cn

摘要: 人们对孤立的手写体汉字字符的离线识别做了大量的研究工作,而走向实用化的进展并不快.除了单字识别率不理想以外,从文本中正确分割出单个汉字字符也是一个主要难题,因为字符的识别离不开正确分割.利用汉字的基本结构特征,根据两个组件之间的上下、左右和包围关系,对组件进行合并形成完整的汉字图像.对整个汉字字符串中组件的宽度和相邻组件的间距进行分析,有助于左右关系组件的合并.实验结果表明,该方法对手写体汉字字符串具有理想的分割效果.

关键词: 手写体汉字串;结构特征;字符分割;组件;合并

中图法分类号: TP391 **文献标识码**: A

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[1~3].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,而离线手写体汉字识别与实用化要求之间还有一定的距离,这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存在着困难,另一方面是由于汉字字符的分割也是汉字识别技术中的一大障碍.实践表明,汉字识别系统的识别率与正确的汉字分割密切相关,错误的分割将导致错误的识别.许多手写汉字识别系统限定将汉字工整地书写在方格纸上,从而大大降低了汉字分割的难度,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分割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和票据自动阅读等)为突破口,将离线手写汉字识别技术推向实用大有前途.

西文字符和数字的分割可分为3类基本方法[4]:(1) 基于结构分析的分割,即从图像特征中寻找字符分割的规则;(2) 以识别为基础的分割;(3) 整体分割策略,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,对西文字符和数字分割的研究取得了较好的进展[4~7],对汉字字符的分割主要针对印刷体文本进行处理[8],而对手写体汉字的分割鲜有研究.

有一种分割手写体汉字的方法[9]是通过调整空间阈值和最小均方判据,由粗分割和细分割两步实现字符分割,从中挑选出最有可能的分割方案进行字符识别.该方法是对自由格式手写体汉字分割技术研究的有益尝试.但文中假定字符串的汉字书写比较匀称,难以满足实用要求;采用投影的方法分割手写体汉字,在字符间距较小和字符有重叠的情况下则显得无能为力.

还有一种分割手写体汉字的方法[10]是从汉字的笔画分析入手,在提取出汉字串的所有笔画之后,按笔画间相交和相离等不同位置关系作笔画的合并,最后得到完整的汉字.该方法能较好地解决因手写体汉字字符间距较小而造成的困难,但对汉字笔画的检测较费时,缺乏实用性.

---

* **收稿日期**: 1999-07-09; **修改日期**: 1999-09-21

基金项目: 国家自然科学基金资助项目(60075007)

作者简介: 吕岳(1968—),男,江苏泰兴人,博士生,工程师,主要研究领域为图像处理,文本分析,智能系统;施鹏飞(1939—),男,上海人,教授,博士生导师,主要研究领域为图像处理,模式识别,人工智能,数据挖掘;张克华(1975—),硕士生,上海人,主要研究领域为文本处理,智能控制.

汉字是表意文字,就结构而言,汉字字符与西文字符和数字有明显的区别.通常情况下,汉字字符的结构比较复杂,分割的难度大,只有从汉字的自身结构特点出发,结合整个字符串的书写特征,才能研究出适合于手写体汉字的字符分割方法.

从图像的角度看,像素组成连通元,而汉字图像包含一个或几个连通元.本文在连通元与汉字单字之间建立了"组件"的概念,以组件为基础,根据两个组件之间的上下、左右和包围关系对组件进行合并,从而实现手写体汉字字符的分割.在合并左右关系组件时,充分利用整个字符串的特征,判断左右相邻的两个组件是否进行合并.

我们以民用信函上手写体地址汉字串为研究对象,从中分割出汉字地址用于识别,为信函自动分拣机提供控制信息.实验结果表明,本文方法取得了理想的手写体汉字字符串分割效果.

1 根据汉字的结构特点的组件合并

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、偏旁部首和单字这3级组成[11].这一理论强调偏旁部首有特定的音或义,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(像素)、笔画、部件和单字这4个层次分析汉字[12].从图像处理的角度看,像素组成连通元,而汉字图像包含一个或几个连通元.由于汉字结构比较复杂,许多汉字的图像由两个以上的连通元组成.

汉字分割就是用一个矩形框对汉字图像定界,通过对组成汉字的几个连通元作适当的合并可以实现这一目的.为了讨论方便,本文在连通元和汉字单字之间建立了组件的概念.

定义1. 一个汉字字符 H 由 n 个连通元 $C^{(k)}(k=1,2,\ldots,n)$ 组成,记为 $H=\bigcup_{\forall k\in\{1,2,\ldots,n\}} C^{(k)}$,将包含一个或几个 $C^{(k)}$ 的集合称为组件 E,即 $E=\bigcup_{k\in\{1,2,\ldots,n\}} C^{(k)}$.

组件是从连通元到完整汉字图像的中间桥梁,组件大于或等于连通元,而小于或等于单字.一般而言,一个组件包括一个或几个连通元,一个汉字单字为包含一个或几个组件.合并开始时,一个连通元即为一个组件,合并结束时,组件就是一个单字.

经过分析发现,任意两个组件之间的位置关系有9种可能的情况,如图1所示,其中的字例是由两个连通元构成的单字,更复杂的汉字可由几个这样的关系复合而成.

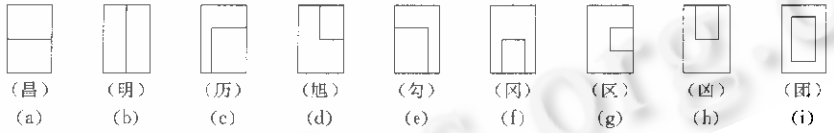


Fig. 1 Topological relation between two elements
图1 两个组件之间的位置关系

任意一个组件(m)的位置信息包括组件的宽度 $W^{(m)}$、高度 $H^{(m)}$、左上角坐标$(LT_x^{(m)},LT_y^{(m)})$、右下角坐标$(RB_x^{(m)},RB_y^{(m)})$以及中心位置$(C_x^{(m)},C_y^{(m)})$.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.将组件(m)和组件(n)合并为组件(k),作如下操作:

$$LT_x^{(k)}=\min(LT_x^{(m)},LT_x^{(n)}),\quad LT_y^{(k)}=\min(LT_y^{(m)},LT_y^{(n)}),$$
$$RB_x^{(k)}=\max(RB_x^{(m)},RB_x^{(n)}),\quad RB_y^{(k)}=\max(RB_y^{(m)},RB_y^{(n)}),$$
$$W^{(k)}=RB_x^{(k)}-LT_x^{(k)},\quad H^{(k)}=RB_y^{(k)}-LT_y^{(k)},$$
$$C_x^{(k)}=(RB_x^{(k)}+LT_x^{(k)})/2,\quad C_y^{(k)}=(RB_y^{(k)}+LT_y^{(k)})/2.$$

图1所示的9种位置关系可分为3类,即上下关系(如图1(a)所示)、左右关系(如图1(b)所示)和包围关系(如图1(c)~(i)所示).显然,对上下关系和包围关系的合并比较容易,但由于手写体汉字字符书写的随意性,对左右关系的组件的合并必须依据汉字的结构特征和整个字符串的书写特征等相关信息作出合理的合并.

2 左右关系组件的合并

当上下关系和包围关系组件合并后,字符串中只剩下左右关系的组件,对它们进行恰当的合并是字符分割

的关键和难点,也是整个合并算法的核心.一般而言,组件的宽度和相邻组件之间的间距是决定是否合并的因素,但是由于手写体汉字的书写千变万化,组件的宽度和间距很不稳定,难以确定可靠的合并规则.而从整个字符串的角度分析所有的组件间距和宽度,可以找到合适的合并规则.对整个字符串进行分析是基于这样的假设:

- 从文本中切取的一串字符由同一人在同一时间书写,其书写风格相对稳定.
- 汉字字符串中字间距与字内距有一定的区别,即字间距一般大于字内距.
- 一串字符中的汉字字符大小比较一致,字符的宽度变化在一定的范围内.

进行上下和包围关系的组件合并后,得到左右位置关系的组件,如图 2 所示.相邻两个组件之间的间距为 $G_e^{(k)}(k=1,2,\ldots,M-1)$,每个组件的宽度为 $W_e^{(k)}(k=1,2,\ldots,M)$,其中 M 为字符串中的组件数.组件(k)和(k+1)合并后形成的新组件的宽度为

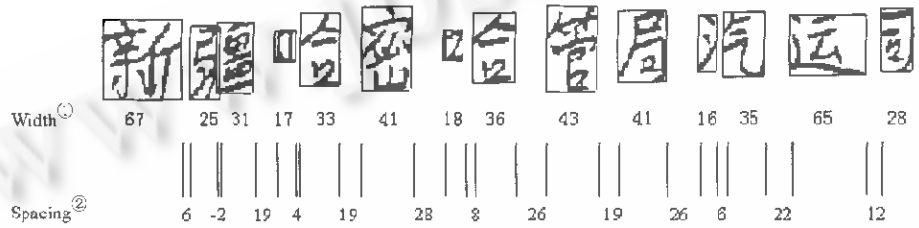
$$W_m^{(k)}=G_e^{(k)}+W_e^{(k)}+W_e^{(k-1)},\quad k=1,\ldots,M-1,$$

对 $G_e^{(k)}$ 和 $W_e^{(k)}$ 作归一化处理:

$$g_e^{(k)}=G_e^{(k)}/\max G_e,\qquad w_m^{(k)}=W_e(k)/\max W_m.$$

其中

$$\max G_e=\max_{1\leqslant k\leqslant M} G_e^{(k)},\qquad \max W_m=\max_{1\leqslant k\leqslant M} W_m^{(k)}.$$

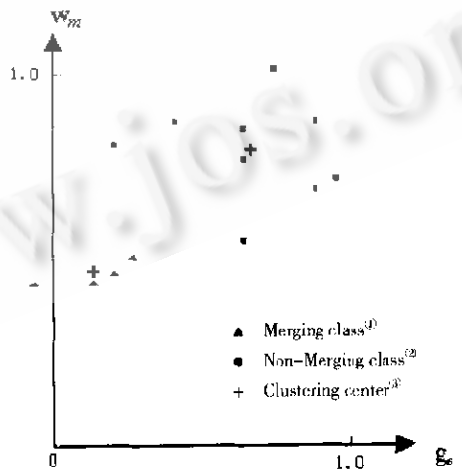


①宽度,②间距.

Fig. 2 The width of elements and spacing between neighbor elements

图 2 左右关系组件的宽度及其间距

以 $g_e^{(k)}$ 和 $w_m^{(k)}(k=1,\ldots,M-1)$作为描述字符串中左右相邻组件对[(k),(k+1)]的参数,通过自动聚类可以形成两个聚类中心(C_{g_1},C_{w_1})和(C_{g_2},C_{w_2}),它们分别对应于可合并类和不可合并类,图 3 是以图 2 中的组件为研究对象的聚类结果.这种分类虽然只是简单的初步估计,但可以为左右关系组件的合并提供直接的指导.



①可合并类,②不可合并类,③聚类中心.

Fig. 3 Clustering generated by writing style of left-right elements

图 3 由字符串左右关系组件的书写特征形成的聚类

在进行左右关系的组件合并前,字符串中有一些组件已经是完整的汉字字符.根据求得的两个聚类中心,如

果某一组件不能与其前后相邻的组件合并,则该组件可能是一个完整的汉字字符.以字符串中这类组件为基础,可以初步估计出汉字的字符宽度 W_a.

以上述对聚类中心和字符宽度的估计作为左右相邻组件的合并条件.首先将所有组件按中心位置的 x 轴坐标顺序排列,根据下面的合并算法对左右关系组件进行合并.

左右关系组件合并算法.

输入:按 x 轴方向顺序排列的组件位置信息.

输出:汉字字符的定界.

步骤:

Step 1. 初始化 flag=F

Step 2. 对所有组件 //合并字符串中狭小的组件

If $W^{(k)}<\delta W_a$

If $(RB_x^{(m+1)}-LT_x^{(m)})<\lambda W_a$

合并(m)和$(m+1)$,flag=T

If $(RB_x^{(m)}-LT_x^{(m-1)})<\lambda W_a$

合并(m)和$(m-1)$,flag=T

Step 3. 对所有组件

If $(RB_x^{(m+1)}-LT_x^{(m)})<\varepsilon W_a$. and. $[(m),(m+1)]$是可合并类

合并(m)和$(m+1)$,flag=T

If $(RB_x^{(m)}-LT_x^{(m-1)})<\varepsilon W_a$. and. $[(m-1),(m)]$是可合并类

合并(m)和$(m-1)$,flag=T

Step 4. If flag$\neq T$ go to Step 1, else stop.

其中 δ,ε 和 λ 是常数因子.

3 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,行分割获得收信人地址区的图像,再作字符串分割获取独立的汉字字符图像,供地址识别.图4给出了字符串分

(a) Original image

(a) 原图

(b) Elements by connected components labeling

(b) 连通元标记后的组件

(c) Elements by upper-bottom merging and inside-outside merging

(c) 上下和包围关系合并后的组件

(d) Elements by left-right merging

(d) 左右关系组件合并的结果

Fig. 4 Segmentation procedure of character string

图 4 字符串分割的过程

割的过程,先对字符串图像进行连通元标记,并将每个连通元作为一个组件看待,再根据组件之间的位置关系对

上下关系和包围关系的组件进行合并,然后估计整个字符串的书写特征,获得两个聚类中心和对字符宽度的估计等相关信息,作为合并左右关系组件的依据,最后分割出单个的汉字字符图像.图5是本文分割算法应用于部分信函地址汉字串分割的结果.

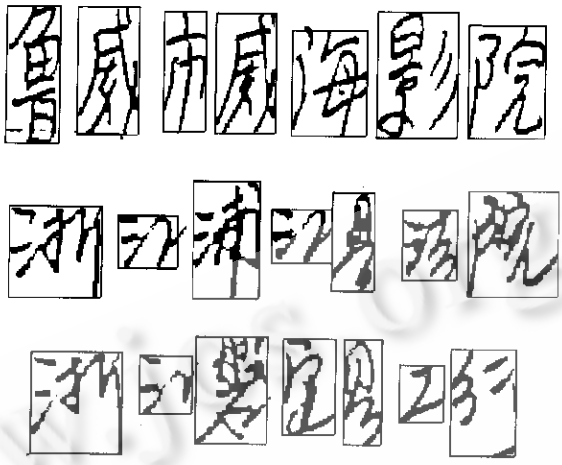


Fig. 5 Segmentation results for character strings of letter address

图5 信函地址的字符分割结果

4 结 论

手写体汉字字符分割是离线汉字识别技术实用化中的关键技术之一.本文利用汉字的结构特征,将上下、左右和包围关系的组件合并为汉字图像,并根据整个字符串的书写特征和对字符宽度的估计来指导左右关系组件的合并,实验结果表明效果较好.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分割,进一步的研究是,在汉字分割过程中融入部件识别、单字识别和语义理解,将会得到更好的分割性能.

References:

[1] Wu, You-shou, Ding, Xiao-qing. Chinese Character Recognition: Theory, Methods and Implementation. Beijing: Higher Education Press, 1992 (in Chinese).

[2] Zhou, Chang-le. Machine Recognition of Handwritten Chinese Character. Beijing: Science Press, 1997 (in Chinese).

[3] Zhang, Xin-zhong. Chinese Character Recognition Technology. Beijing: Tsinghua University Press, 1992 (in Chinese).

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

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

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

[7] Lu, Zhong-kang. Segmentation and recognition of connected handwritten characters [Ph. D. Thesis]. Shanghai Jiaotong University, 1999 (in Chinese).

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

[9] Hong, C., Loudon, G., Wu, Y., *et al*. Segmentation and recognition of continuous handwriting Chinese text. International Journal of Pattern Recognition and Artificial Intelligence, 1998,12(2):223~232.

[10] Tseng, L., Chen, R. A new method for segmenting handwritten Chinese characters. In: Bob, W ed. Proceedings of the 4th International Conference on Document Analysis and Recognition. Los Alamitos, CA: IEEE Computer Society, 1997. 568~571.

[11] Fu, Yong-he. Study on Structure and Components of Chinese Characters. Shanghai: Shanghai Education Press, 1993.

108~169 (in Chinese).

[12] Han, Bu-Xin. Combination of Chinese character constituentsi——a latent structural unit. Journal of Chinese Information Processing, 1995,9(3):27~32 (in Chinese).

附中文参考文献:

[1] 吴佑寿,丁晓青.汉字识别原理方法与实现.北京:高等教育出版社,1992.

[2] 周昌乐.手写汉字的机器识别.北京:科学出版社,1997.

[3] 张炘中.汉字识别技术.北京:清华大学出版社,1992.

[7] 陆钟慷.连接手写体字符的分割与识别[博士学位论文].上海交通大学,1999.

[11] 傅永和.汉字结构和构意成分的研究.上海:上海教育出版社,1993.

[12] 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.

Segmentation of Handwritten Chinese Character String Based on Merging of Elements

LÜ Yue, SHI Peng-fei, ZHANG Ke-hua

(*Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai* 200030, *China*)

E-mail: pfshi@ippr.sjtu.edu.cn

http://www.sjtu.edu.cn

Received June 9, 1999; accepted September 21, 1999

Abstract: A number of papers concerning the off-line recognition of handwritten Chinese characters have been published in the recent years, and almost all of them focus on the recognition of isolated characters. However, off-line recognition of handwritten Chinese characters is not satisfactory. One reason is that the recognition rate is low, the other is that the segmentation of handwritten Chinese characters is a difficult problem because recognition of characters relies on correct segmentation of characters. In this paper, according to structural features of Chinese characters, elements are merged based on their topological relations, viz., upper-bottom, left-right and inside-outside. The width of elements and the spacing between neighboring elements in the whole handwritten Chinese character string are analyzed to guide the merging of left-right elements. Experimental results show that the method has satisfactory performance for segmenting handwritten Chinese character string.

Key words: handwritten Chinese character string; structural feature; character segmentation; element; merging