ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, Supplement, December 2008, pp.52−58

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于多边形逼近和有限状态机的笔段提取-合并算法*

吕新桥+

(华中科技大学 计算机学院,湖北 武汉 430074)

A Segment Abstraction-Integrate Algorithm Based on Polygon Approximation and Finite State Machines

LÜ Xin-Qiao+

(College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

+ Corresponding author: E-mail: xqlv@hust.edu.cn

LÜ XQ. A segment abstraction-integrate algorithm based on polygon approximation and finite state machines. *Journal of Software*, 2008,19(Suppl.):52−58. http://www.jos.org.cn/1000-9825/19/s52.htm

Abstract: In this paper, a segment extraction-integrate algorithm based on polygon approximation and finite state machines for on-line Chinese characters recognition (OLCCR) is presented. With this method, the point with the smallest interior angle which is less than the given value is detected and the whole stroke is split into two adjacent curves by this point, which is called as a cut-off point or an inflexion. To each of the two curves, the same step is performed to detect the cut-off points respectively. The same operations are performed iteratively until the smallest interior angle in all the curves is larger than the given threshold value. All the cut-off points and the start-end points compose the stroke and every pair of adjacent points constructs a segment. After segments have been extracted, Finite State Machines is used to check whether the adjacent segments need combination thus redundant segments can be reduced. Experiments proved that this method has the advantages of less computing complexity and better approximating effect than other methods.

Key words: polygon approximation; finite state machines; segment abstraction-integrate; cut-off point; interior angle; stroke

摘 要: 汉字的基本特征表示是笔段,提出一种基于多边形逼近和有限状态机的笔段提取-合并算法.该算法首先找到笔画的拐点(最小内角值小于指定阈值),然后分别寻找拐点两侧曲线段上的拐点,反复执行,直到再也找不到拐点为止.依次连接一个笔画中所有曲线的起点和终点,就形成了该笔画的笔段系列.随后,运用有限状态机描述并判定笔段的状态,并以此判定笔段的合并要求,以最大限度地减少冗余笔段.实验表明,这种算法具有较低的计算复杂度和很好的逼近效果,能适应手写汉字的笔段提取合并要求.

关键词: 多边形逼近;有限状态机;笔段提取-合并;拐点;内角;笔画

联机汉字识别(OLCCR)系统中,汉字的特征提取是非常重要的环节.Liu[1]把汉字分成 5 个层次,笔段位于最底层,描述了整个汉字最基本的信息.以笔段描述汉字的特征是手写汉字系统的通用的做法.笔段提取与合并的的成功与否极大程度地影响着整个汉字识别系统的性能.上世纪 70 年代以来,人们提出了许多算法以有效地提

* Received 2008-05-01; Accepted 2008-11-25

取笔段.例如:基于边缘的方法[2],基于距离的方法[3],基于面积的方法[4],基于遗传算法的方法[5],基于时刻的方法[6,7],基于波浪线唯一性描述符方法[8]等等.

Ramer. U[9]提出了一种基于多边形逼近和使用距离公式的迭代逼近算法.算法抽取当前曲线上到近似直线的距离小于阈值ε的点作为特征点.该算法计算量小,算法稳定且能对任何二维数字曲线进行逼近.

郑胜林[10]提出一种逼近整合算法.这种算法首先求出近似多边形,接着对多边形上的边线进行整合并计算出笔段的方向码,最后根据方向码对笔段进行合并.这种算法能将初始曲线精确地趋近为一个多边形,而且还能处理带有圆圈的曲线.但是,对于弯曲程度稍高的笔画,它的提取结果就变得非常奇怪,并产生出许多冗余的笔段.此外,它要求的计算量也远远高于上一种算法.

常用的合并算法是合并具有相同方向码的两个相邻笔段,这样的算法简便快速.郑胜林[10]提出了一种考虑笔段的方向码,长度和角度的合并算法,用来合并连续的三个笔段.这种合并方法虽然一定程度上保证合并的有效性,但忽略了很多应该考虑合并的情况,因此可能会会产生错误的合并结果,如图 1 所示.

Fig.1 The abstraction result of handwritten character “人”

(it’s obvious that the first segment “捺”should be combined with the second segment “撇”)

图1 手写体汉字“人”的笔段抽取结果(“捺”笔段显然应该与“撇”笔段合并)

本文提出一种基于多边形逼近的手写体汉字笔段的提取方法.该方法求取笔画曲线上内角最小的点作为笔画曲线的特征点,依次连接特征点,就得到笔画曲线的近似多边形,多边形的每一条边就是一个笔段.提取了笔段之后,运用有限状态机检查笔段的状态以确定这些笔段是否需要合并以及该如何合并这些笔段.实验证明,笔段提取算法的计算复杂度低,逼近效果好,合并算法能正确地根据笔段的状态对笔段进行合并,减少了笔段的冗余信息,提高后续整字匹配的效率和准确率.

1 笔段提取算法

一个联机手写汉字由 m 个笔画组成,记为

$$w=\{s_i \mid i=1,2,3,...,m\} \tag{1}$$

其中,w 是汉字的轨迹,m 是笔画的个数.

s_i 是该汉字的第 i 个笔画(一条曲线段),由 n 个连续的点构成,记为

$$s_i=\{p_j \mid j=1,2,3,...,n\} \tag{2}$$

其中, p_j 为笔画 s_i 的第 j 个笔迹点,n 是笔画 s_i 的点的个数.

笔画 s_i 也可以表示为

$$\begin{aligned} s_i &= \{seg_k \mid k=1,2,3,...,r\} \\ seg_k &= \overline{p_{ks}p_{ke}} \end{aligned} \tag{3}$$

其中, seg_k 是笔画 s_i 的第 k 个笔段, p_{ks} 和 p_{ke} 分别是笔段 seg_k 的起点和终点.这意味着,一个笔画由一系列的笔段首尾相连而成.笔段描述了汉字的重要的结构特征信息,它具有方向、长度等属性.

从式(2)和式(3)可以看出,笔段的抽取实际上就是在每一个笔画里分别找到一系列的特征点(拐点),这些特征点把整个笔画分割成若干条首尾相连的曲线段,连接这些曲线段的起点和终点,就构成了该笔画的笔段系列.

为叙述方便,我们用曲线段上第 i 个点 p_i 的“内角” α_i 来描述曲线段在该点的弯曲度.点 p_i 的“内角” α_i 为该点与曲线段起点连线和该点与曲线段终点连线的夹角,即

$$\alpha_i = \text{ang}(\overline{p_s p_i}, \overline{p_i p_e}) \tag{4}$$

内角 α_i 描述了曲线段在该点的弯曲的程度,显然, α_i 越小,曲线段在该点弯曲得越厉害,意味着笔画在该点发生变化的可能性也越大,该点成为拐点的可能性也越大.一个笔画的书写因人而异,变化大,因此可能一个笔画可能会找出一个或者多个拐点,这就需要为拐点的内角规定一个阈值ε,只有内角小于阈值ε的点才有可能成为曲线上的拐点.

定义. 一个曲线段上的拐点 p_i 指的是满足条件(1)和(2)的点.

条件(1): $\alpha_i = \min(\alpha_m)$,m=1,2,…,n,n 为曲线段上点的个数

条件(2): $\alpha_i < \varepsilon$.

图 2 描述了一个曲线段分别被抽取为 1 个、1 个和 2 个笔段的笔画.图 2(a)中,曲线段书写时带有抖动.图 2(b)的笔画有一定的弯曲,但曲线上没有满足条件(2)的点,所以也被抽取为一个笔段.而图 2(c)存在满足条件(1)和条件(2)的拐点,因此杯抽取为 2 个笔段.

(a) (b) (c)

Fig.2 Three stroke curves with 1,2 segments abstracted.

The curve is the origin curve and the line is the segment abstracted

图 2 3 种具有 1 条或者 2 个笔段的曲线段实例.曲线段是笔段的初始痕迹,直线是抽取出来的笔段

算法.

对于给定的一个手写笔画 $s_i = \{p_j \mid j = 1,2,3,...,n\}$,定义栈 S_c 和 S_o 以及角度阈值 ε ,提取笔段的具体操作步骤如下所述:

(1) 初始化:将轨迹曲线的起始点 p_1 压入 S_c 栈,将轨迹曲线的结束点 p_n 压入 S_o 栈;

(2) 取得 S_c,S_o 的栈顶元素的值,分别记为 x,y,并设置一个临界点位置 pos=x+1,设置找到的标志位 Found=false,初始化 Max 为一个合适的值;

(3) 若 $pos \neq y$,计算由 p_x,p_{pos},p_y 这 3 个点所形成的内角,记为 Angle,否则执行(5);

(4) 若 Angle≥Max,则设置 pos=pos+1;转到(3),否则保存当前点 *TempPos*=pos,设置 pos=pos+1,Max=Angle,Found=true,转到(3);

(5) 若 Found=false,将栈 S_o 的栈顶元素取出并压入 S_c 栈,否则将 *TempPos* 压入 S_o 栈;

(6) 若栈 *So* 不为空,转到(2),否则转到(7);

(7) 依次取出栈 *Sc* 中的所有元素,每一对相邻的两个点确定一个笔段.

算法简单易行,对手写笔迹有较好的抽取效果,图 3 显示了一个手写笔画应用本算法提取后的结果.

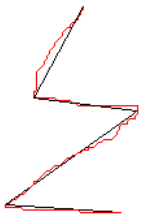


Fig.3 Abstraction result of one handwritten stroke

图 3 一个手写笔画的抽取结果

2 合并算法

书写的随意性使得需要对提取后的笔段进行合并,以减少特征的存储量和冗余信息.常用的算法以方向码来描述笔段(4 方向或者 8 方向)并把相邻的方向码是否相同作为相邻笔段是否需要合并的依据.

本文在充分考虑笔段的方向码,笔段长度的同时,还考虑日常书写习惯以及相邻方向笔段过渡的因素,利用确定的有限状态机(deterministic finite state machines,简称为 DFSM)判定相邻笔段合并的可能性.

有限状态机是一种具有离散的输入和输出系统的数学模型.确定的有限状态机 M 是一个五元组:$M=(K,\Sigma,f,S,Z)$,其中各元素的定义如下:

(1) K 是一个有穷集,它的每个元素称为一个状态;

(2) Σ是一个有穷字母表,它的每个元素称为一个输入字符;

(3) f 是转换函数,是在 $K\times\Sigma\to K$ 上的映射;

(4) $S\in K$ 是唯一的一个初态;

(5) $Z\subset K$,是一个终态集.

在本算法中,DFSM 定义为 $M=(\{S,A,B,C,D,Z\},\{a,b,c,d,e,f,g,h,i,j,k\},f,S,\{Z\})$.

一个笔画被抽取成笔段后,形成了笔段系列,判断相邻 3 个笔段的状态,以及它们之间的转换过程,就可以有效地进行笔段合并.

状态的定义.

S:笔画的起始笔段;

A:相连的两个笔段,满足合并条件;

B:相连的两个笔段,不满足合并条件;

C:相连的 3 个笔段,前两个笔段满足合并条件,后两个笔段也满足合并条件;

D:相连的 3 个笔段,前两个笔段满足合并条件,后两个笔段不满足合并条件;

E:相连的 3 个笔段,前两个笔段不满足合并条件,后两个笔段满足合并条件;

F:相连的 3 个笔段,前两个笔段不满足合并条件,后两个笔段也不满足合并条件;

Z:已无笔段需要判断是否合并,即终止状态.

其中,状态 E,F,因为前两个笔段不满足合并条件,所以无需处理,因此,用状态机描述时可以忽略不计的.

输入字符的定义.

a:输入一个笔段,该笔段与前一笔段满足合并条件;

b:输入一个笔段,该笔段与前一笔段不满足合并条件;

c:合并两个笔段,合并后的笔段与前一笔段或者后一笔段满足合并条件;

d:合并两个笔段,合并后的笔段与前一笔段或者后一笔段不满足合并条件;

e:笔段集非空;

f:笔段集为空;

g:合并笔段.

转换函数 f 的定义:

状态之间需要进行转换,而且是在输入条件的驱动下进行转换.显然,转换的过程如下:

(1) 有一个笔段,输入一个笔段,如果两笔段符合满足合并条件,则进入状态 A,否则进入状态 B;

(2) 对 A,输入一个笔段,如果该笔段和前一笔段满足合并条件 a,则进入状态 C,否则进入状态 D;

(3) 对 B,输入一个笔段,如果该弊端和前一笔段满足合并条件 a,则进入状态 A,否则进入状态 B;

(4) 对 C,根据前两和后两笔段的合并系数,决定合并前两和后两笔段,合并后的笔段和两一个比合并的笔段如果满足条件 c,则转换到状态 A,否则转换到状态 B;

(5) 对 D,转换过程如(4);

(6) 如果笔段集为空 f,则如果当前状态为 A,直接合并,终止,如果当前状态为 B,直接终止.

因此,一个笔画的笔段状态转换如图 4 所示.

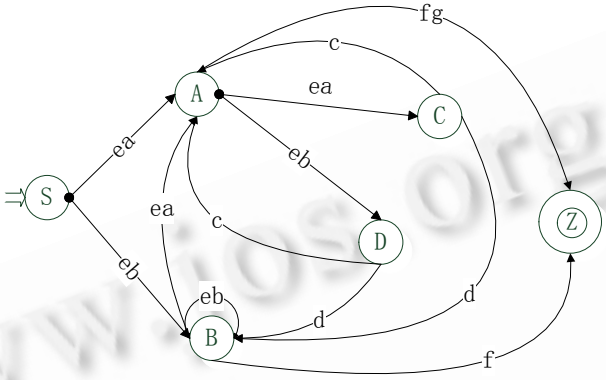


Fig.4 State-Transformation chart of one stroke

图 4 一个笔画的笔段状态转换图

其中,满足笔段合并的条件为如下中的一种:

(1) 两个相邻笔段的方向码相同;

(2) 两个相邻笔段的方向码不同,但是长度比太大或者太小;

(3) 两个相邻笔段的方向码不同,但是符合人工手写习惯的.

3 实 验

图 5 是用本文算法对一个曲线笔画的提取结果,图 5、图 6(a)是 Ramer. U[9]算法的提取结果,图 6(b)是郑胜林[10]算法的提取结果.这 3 种算法都是基于相同的原始数据.从中可以看出,采用本文中的新算法,能够精确地近似笔画曲线.相比其他两种算法,更少的笔段数量使得本算法在合并阶段花费更少的时间.

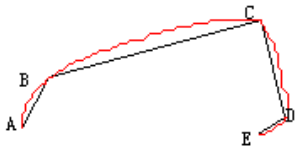
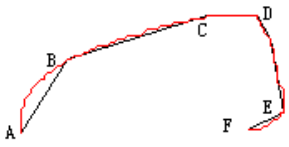


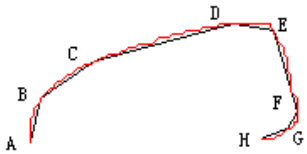
Fig.5 Result used our algorithm

图 5 本文算法的提取结果



(a) The result with Ramer. U

(a) Ramer. U 算法的提取结果



(b) The result with Shenlin Zheng

(b) 郑胜林算法的提取结果

Fig.6

图 6

图 7 显示了应用本文算法对一个手写字符进行笔段提取的结果,图 8 是采用郑胜林[10]算法的笔段提取结果.采用本文算法,从手写体汉字“百”中提取了 13 条笔段,而在郑胜林[10]算法下,相同的书写笔迹产生了 18 条笔段.

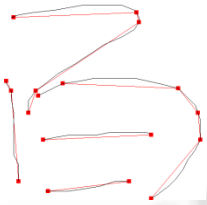


Fig.7 The abstraction result with the new algorithm

图 7 新算法的笔段提取结果

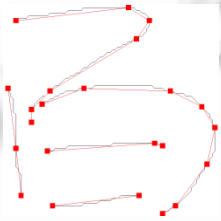
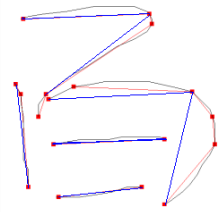


Fig.8 The abstraction result with Zheng[10] algorithm

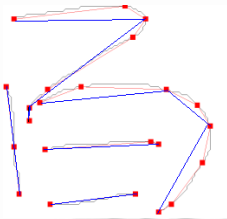
图 8 郑胜林算法[10]的笔段提取结果

图 9(a)是采用本文中新笔段合并算法的结果,图 9(b)是采用郑胜林方法的笔段合并结果.从中可以看出,图 9(a)中的笔段最终被合并为 7 条笔段,而在图 9(b)中被合并为 9 条笔段.可以得出这样的结论:采用文中新合并算法能更精确地描绘汉字的特征,在匹配阶段所要花费的计算量要少于采用郑胜林[10]方法的计算量.



(a) The combination result with our algorithm

(a) 本文笔段合并算法的合并结果



(b) The combination result with Zheng[10] algorithm

(b) 郑胜林[10]方法的笔段合并结果

Fig.9 The segments-combination result

图 9 笔段合并结果

4 总 结

本文提出一种适用于联机汉字识别(OLCCR)的基于多边形逼近的笔段提取算法.首先检测一个笔画上的拐点,以此拐点为界,将整条笔画划分为相邻的两条曲线,在每条曲线中分别执行相同的步骤,直到所有曲线上找不到相应的拐点为止.连接所有拐点,得到抽取后的笔段系列,然后根据相邻笔段的状态,运用有限状态机进行合并,取得了理想的结果.实验证明,这种方法具有很低点的计算复杂度和很好的逼近效果,能够很好地对手写汉字的笔段进行有效地合并.该方法应有于联机汉字识别系统中,达到了每秒 20 个字的识别速度和 97.2%的有效识别率.实验表明,本文提出的算法对于手写汉字的结构特征提取是非常有效的.

References:

[1] Liu YJ. An on-line Chinese character recognition system for handwritten in Chinese calligraphy. In: Int'l Workshop on Frontiers in Handwriting Recognition. 1991.

[2] Sklansky J, Chazin RL, Hansen BJ. Minimum perimeter polygons of digitized silhouettes. IEEE Trans. on Computer, 1972,21(3): 260−268.

[3] Sklansky J, Gonzalez V. Fast polygonal approximation of digitized curves. Pattern Recognition, 1980,12(5):327−331.

[4] Wu JS, Leou JJ. New polygonal approximation schemes for object shape representation. Pattern Recognition, 1993,26(4):471−484.

[5] Huang SC, Sun YN. Polygonal approximation using genetic algorithms. Pattern Recognition, 1999,32(8):1409−1420.

[6] Shu HZ, Luo LM, Zhou JD. Moment-Based methods for polygonal approximation of digitized curves. Pattern Recognition, 2002,35(2):421−434.

[7] Li YS, Zhou JD, Zhan H. Momen *t*-based methods for polygona l approximation of digitized curves. Chinese Journal of Computers, 2001,24(4):354−357 (in Chinese with English abstract).

[8] Zhou ZD, Zhang PZ, Shu HZ. Polygonal approximation of digitized curves based on uniqueness wavelet descriptor. Journal of Data Acquisition & Processing, 2005,20(1):40−43 (in Chinese with English abstract).

[9] Ramer U. An iterative procedure for polygonal approximation of plane curves. Computer Graphics and Images Processing, 1972,1(3):244−256.

[10] Zheng SL, Pan BC, Zhao XJ. Approximation-Integrate method of extracting stroke features on line. Computer Engineering and Design, 2006,27(7):1248−1250 (in Chinese with English abstract).

附中文参考文献

[7] 李松毅,周瑾丹,张惠.基于矩的数字图像多边形逼近方法.计算机学报,2001,24(4):354−357.

[8] 周正东,张品正,舒华忠.基于小波惟一描述子的多边形逼近方法.数据采集与处理,2005,20(01):40−43.

[10] 郑胜林,潘宝昌,赵学军等.联机手写笔画特征抽取的逼近-合并算法.计算机工程与设计,2006,27(7):1248−1250.

吕新桥(1973－),男,湖北武汉人,硕士,主要研究领域为计算机图形图像,模式识别,企业信息化.