

# 基于汉字骨架手写识别算法分析

郭 星<sup>1,2</sup>, 吴建国<sup>1,2</sup>, 张义超<sup>1</sup>, 李 炜<sup>1</sup>

(1. 安徽大学 计算机科学与技术学院, 安徽 合肥 230039;

2. 安徽大学 软件学院, 安徽 合肥 230039)

**摘 要:** 汉字识别是目前识别领域中的难题, 而细化是其预处理第一步。研究了基于主曲线骨架提取算法和基于模板的 Hilditch 算法两种手写汉字识别算法。通过程序实现两个算法分析比较了两种算法对于不同质量的汉字图像骨架提取效果。实验表明前者的鲁棒性能好, 能适合各种质量差的图像, 但是算法还有待于优化, 而后者在图像质量差的情况下细化效果很差, 但是算法速度很快。

**关键词:** 主曲线; Hilditch; 骨架; 分析

**中图分类号:** TP301.6

**文献标识码:** A

**文章编号:** 1673-629X(2009)07-0114-03

## Analysis and Research Based on Skeleton of Handwritten Chinese Character Algorithms

GUO Xing<sup>1,2</sup>, WU Jian-guo<sup>1,2</sup>, ZHANG Yi-chao<sup>1</sup>, LI Wei<sup>1</sup>

(1. Institute of Computer Science and Technology, Anhui University, Hefei 230039, China;

2. School of Software, Anhui University, Hefei 230039, China)

**Abstract:** Chinese characters discernment is the hard problem in discernment field, and the thinning is the first step to deal with. Studies two of the skeleton algorithm, one is based on principal curve, and the other is Hilditch algorithm which is based on a template. Analysis of two different algorithms for the quality of the image of Chinese characters skeleton extraction. Experiments show that the former's performance well, and be able to fit a variety of poor quality images, but still to be optimization. Although the thinning effect of the latter is poor, but the speed is fast.

**Key words:** principal curve; Hilditch; skeleton; analysis

## 0 引 言

手写汉字识别是模式识别的经典问题之一, 而识别的关键在于特征值的提取, 根据抽取特征的方法一般将其分为两类: 统计特征(局部特征和全局特征)和结构特征<sup>[1]</sup>。结构特征在于汉字的外在表现, 如横、撇、竖等。但是汉字的字形变化很大, 粗细不一, 给特征提取带来一定的困难。因此采用提取骨架的方法使汉字图像粗细归一化是特征提取之前一个十分重要的步骤。

一般说来, 骨架主要有三个主要特征<sup>[2]</sup>: 连续性、最小宽度为 1 和中心性。目前骨架算法主要还是细化为主, 具有代表性的有 Rosenfeld 细化算法, Hilditch 细

化算法, OPTA 细化算法等。但是以上算法均是基于模板操作的。而有些为非模板型算法, 如基于主曲线的骨架提取算法, 文中着重研究比较 Hilditch 细化算法和主曲线算法。

## 1 主曲线算法

主曲线概念是 Hastie 和 Stuetzle 于 1984 年提出的<sup>[1]</sup>。主曲线是通过数据分布“中间”并满足“自相合”的光滑曲线, 其目的是根据给定的数据集合求出一个曲线, 使得这个曲线对给定的数据集合是某种意义下的对偶, 形象地说, 希望能寻找通过数据分布“中间”的曲线和曲面, 使它能真实地反映数据的形态, 即曲线是数据集合的“骨架”, 数据集合是这个曲线的“云”<sup>[3]</sup>。由此可见, 主曲线对数据的信息保持性好。主曲线的理论基础是寻找嵌入高维空间的非欧氏低维流形, 也是线性主成分的非线性推广。

主曲线算法步骤<sup>[4]</sup>:

收稿日期: 2008-11-11; 修回日期: 2009-03-03

基金项目: 安徽省自然科学基金资助计划项目(2006KJ013A)

作者简介: 郭 星(1983-), 男, 安徽庐江人, 硕士研究生, 研究方向为中文信息处理; 吴建国, 教授, 博导, 研究方向为中文信息处理; 李炜, 副教授, 硕导, 研究方向为计算机集成制造。

文中采用推广的多边形(PL)主曲线算法来提取手写汉字的骨架结构<sup>[3]</sup>。多边形线算法的基本运算方法是首先确定一条直线段,然后在循环算法中通过不断加入新的顶点来增加线段的数量。在加入一个新的顶点以后,所有的顶点位置在一个内部的环中被更新。算法如下:

#### (1) 初始化。

首先采用传统的细化算法来获取原始字迹模板的近似骨架,并用图  $G$  来存储得到初始曲线的顶点信息。 $G$  由两个集合  $V, S$  构成,其中  $V = \{V_1, V_2, \dots, V_n\} \subset R^d$  是顶点的集合,  $S$  是边的集合。

#### (2) 投影。

根据  $t_f(x) = f^{-1}(x) = \max_t \{t: \|x - f(t)\| = \min_{\omega} \|x - f(\omega)\|\}$  把数据集  $X_n$  划分到属于骨架图顶点和边的最近邻区。

#### (3) 拟合 - 光滑。

在这一步的目的是调整骨架图  $G$  的光滑性,使之更好地拟合字形。对给定数据集  $X_n = \{X_1, X_2, \dots, X_n\}$ , 用距离惩罚函数  $E(G) = \Delta(G) + \lambda W(G)$  取最小来优化骨架图,其中  $\Delta G = E \|X - f(\lambda_f(X))\|^2$  表示集中点到图形  $G$  的距离平方的平均值,  $W(G)$  是关于图形曲率的惩罚函数。 $\Delta G$  值较小表示骨架图可以较好地拟合数据,  $W(G)$  值较小能保证骨架图的光滑性较好。其次做顶点优化步,即调整骨架图  $G$  中顶点和边的位置,使得距离惩罚函数  $E(G)$  取得局部最小值。

#### (4) 重构。

利用骨架图的几何性质对顶点和边的结构进行修改,消除或校正初始图形的瑕疵,例如删除短分支、删除小圈等。

## 2 Hilditch 细化算法

Hilditch 算法是现有的一种细化算法,它适用于输入图像为 0 和 1 的二值图像<sup>[5]</sup>。像素为 1 的区域是需要细化的部分,像素为 0 的区域是背景。

Hilditch 算法可描述如下:设  $P$  为被检测的像素,  $f(P)$  为像素  $P$  的灰度值,  $n_i$  ( $n = 1, 2, \dots, 8$ ),  $n_i$  的位置如图 1 所示。

|       |       |       |
|-------|-------|-------|
| $n_4$ | $n_3$ | $n_2$ |
| $n_5$ | $P$   | $n_1$ |
| $n_6$ | $n_7$ | $n_8$ |

图 1 二值图像灰度值

设集合  $I = \{1\}$  表示需要细化的像素子集,集合

$N = \{g | g - m \leq 0\}$  表示背景像素子集,集合  $R = \{-m\}$  表示在第  $m$  次减薄时,  $I$  中被减掉的像素。图像细化的减薄条件为:

(1)  $P$  为黑点,在二值图像中为 1,即:  $f(P) \in I$

(2) 环绕在  $P$  周围的点  $n_1, n_3, n_5, n_7$  为黑点的总数大于等于 1,即:  $U(P) \geq 1$

其中  $U(I) = a_1 + a_3 + a_5 + a_7$ , 这里  $a_i$  为:

$$a_i = \begin{cases} 1, f(n_i) \in I \\ 0, \text{其他} \end{cases}$$

(3) 环绕在  $P$  周围的点,背景白点(值为 0)比黑点(值为 1)的总数多 2 个,即:  $V(P) \geq 2$ , 其中

$$V(P) = \sum_{i=1}^8 (1 - a_i)$$

(4) 环绕在  $P$  周围黑点的总数大于等于 1,即:  $W(P) \geq 1$ , 其中

$$W(P) = \sum_{i=1}^8 c_i$$

这里  $c_i$  为:

$$c_i = \begin{cases} 1, f(n_i) \in I \\ 0, \text{其他} \end{cases}$$

(5) 当满足条件  $n_{2i} - 1$  为背景像素  $n_{2i}$ , 并且为在减薄过程中减掉的点或是图像黑点, 或者  $n_{2i} + 1$  为减薄过程中减掉的点或是图像黑点时, 记为  $b_i$  ( $i = 1, 2, 3, 4$ ), 如果  $b_i$  累加和等于 1, 则该条件成立, 即:  $x(P) = 1$ , 其中

$$x(P) = \sum_{i=1}^4 b_i$$

这里,  $b_i$  为:

$$b_i = \begin{cases} 1, f(n_{2i-1}) \in N \text{ and } f(n_{2i}) \in I \cup R \\ 0, \text{其他} \end{cases}$$

(6)  $f(n_i) \in R$  或  $x_i(P) = 1$  ( $i = 3, 5$ ), 其中  $x_i(P)$  表示对  $P$  的第  $i$  个领域像素的  $x(P)$ 。

当对图像进行细化处理时,根据上述的条件,对扫描的像素进行判断,如果所扫描的像素同时满足上述的 6 个条件,就可以将该像素去除。直到经过  $m$  次后,图像上的像素均不能同时满足以上 6 个条件,则细化处理结束。

## 3 实验结果与分析

实验用画图的画笔和喷刷写了两个字,前者字迹清晰,后者字迹边界毛刺比较多(如图 2 所示),对这两字分别进行主曲线提取骨架和 Hilditch 算法细化,实验结果如图 3、图 4 所示:左边为主曲线算法提取结果,右边为 Hilditch 算法细化结果。

由以上实验可知,在图像清晰的情况下,两者的细化程度差不多,都能体现文字的主要结构特征。而在

图像有毛刺的情况下,两者表现就大不一样了,前者还是可以良好地体现文字的结构特征,而后者则多了太多的小圈,这非常不利于识别特征的提取。在手写识别的系统中,图像是扫描的,图像质量肯定会层次不齐,因此基于主曲线算法的适应鲁棒性较基于模板的好得多,这点对于手写识别系统很重要<sup>[6]</sup>。

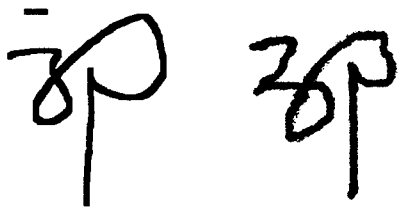


图 2 平整和毛刺原图

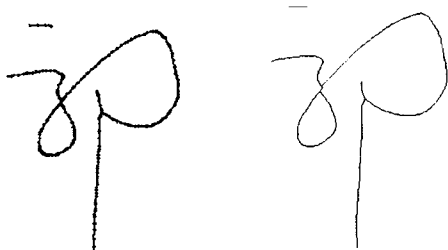


图 3 平整图主曲线和 Hilditch 提取

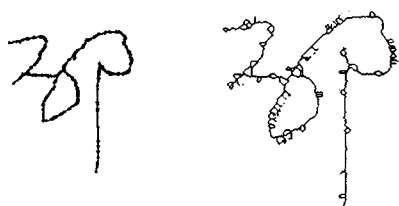


图 4 毛刺图主曲线和 Hilditch 提取  
两者之间的比较:

(1)传统的基于模板算法提取后的骨架,如果要提取其特征,由于图像是以位图点阵的格式存储的,在提取过程中,必须要对整幅图像进行扫描,这样会带来时间复杂度的提升。而使用基于主曲线提取的骨架图像,它是以以矢量形式存储的,这样在特征提取时,不仅带来了性能上的提升,而且还大大方便了特征提取。

(2)传统基于模板的细化算法操作对象是像素,如

果图像质量差、毛刺多,则细化后的图像效果极差。而主曲线算法则反之,它提取的是数据集分布,所以鲁棒性更好,在手写识别过程中,对于质量差的图片识别率将提高不少。

(3)由于主曲线提取的是汉字的骨架,而非脊梁骨,所以更能准确地反映字符的整体拓扑结构,在其基础上提取的特征比基于模板提取得到的特征不但数量少而且要更加准确。这将对提高字符识别率有很大帮助。

(4)由于在提取主曲线时,需要对大量数据进行投影运算,所以其时间效率比基于模板细化差,这方面还有待进一步研究。

#### 4 结束语

提取骨架是特征提取的重要课题之一。分析两类细化算法,并比较,可以看出基于主曲线提取算法质量较好,但是目前对于其时间效率还是不太满意,对于实时性较高的系统还是无法应用,这方面还有待于进一步提高。

#### 参考文献:

- [1] 张军平,王 珏.主曲线综述[J].计算机学报,2003,26(2):129-146.
- [2] Kegl B, Krzyzak A, Tomes R. Piecewise linear skeletonization using principal curves[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2002, 24(1): 59-74.
- [3] 史绍强.一种改进型的汉字字符图像细化算法[J].计算机技术与发展,2007,17(9):88-91.
- [4] Kegl B, Krzyzak A, Cursil T. A polygonal Line algorithm for constructing principal curves[C]// Proceedings of Neural Information Processing System. Denver Colorado, USA: [s. n.], 1999: 501-507.
- [5] 史绍强,王英健,唐贤瑛.基于整形特征和模糊识别的手写体汉字识别[J].微机发展(现更名:计算机技术与发展), 2004, 14(1): 114-116.
- [6] 何 斌,马天予,王运坚,等. Visual C++ 数字图像处理[M].北京:人民邮电出版社, 2001.

(上接第 113 页)

- [4] 王兆安,杨 君.谐波抑制和无功功率补偿[M].第 2 版.北京:机械工业出版社,2006.
- [5] Zhang Haibo, Chen J Y, Xiong Xianping, et al. Computer-based monitoring system of static var compensator[C]// International Conference on Power System Technology. [s. l.]: IEEE, 2002: 1904-1907.
- [6] 李庆庆,张燕平.基于模糊边缘检测算法的车牌定位[J].

计算机技术与发展,2006,16(12):7-8.

- [7] 李付鹏,汪继文.几种复合型数值方法的算法模拟与性能比较[J].微机发展(现更名:计算机技术与发展),2004,14(1):12-14.
- [8] Numberger G. Approximationly Spline Functions[M]. Berlin:Spring Verlag, 1989.