

一种女书曲线轮廓字形自动生成方法^{*}

严红娟, 王江晴, 李 波

(中南民族大学 计算机科学学院, 武汉 430074)

摘 要: 针对点阵字形存储量大、缩放字形质量难以保证的缺点, 对将女书点阵字形自动转换成曲线轮廓字形的方法进行了研究。首先提取女书轮廓数据; 然后提出了一种基于最大误差值的特征点提取方法; 最后用直线和二次 Bezier 曲线拟合字形轮廓段, 存储曲线轮廓数据, 从而得到用曲线描述的女书字形。实验结果表明, 该方法计算简单, 大大减少了存储空间, 拟合出来的曲线和原字形轮廓一致, 同时也消除了因手写而造成的毛刺现象, 为在计算机中处理女书奠定了基础。

关键词: 女书; 点阵字形; 曲线轮廓字形; 特征点; 二次 Bezier 曲线

中图分类号: TP391.1 **文献标志码:** A **文章编号:** 1001-3695(2012)09-3558-03

doi: 10.3969/j.issn.1001-3695.2012.09.095

Approach for automatic font outline generation of NüShu

YAN Hong-juan, WANG Jiang-qing, LI Bo

(College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China)

Abstract: For the disadvantages of bitmap font, such as large storage, the quality of zooming was hard to guarantee, this paper designed an approach for auto-generating curve contour character from bitmap. The main idea of the implementation was to extract the contour data of NüShu character, and then proposed an approach for detecting feature points based on maximum error, at last fitted the contour based on straight line and quadratic Bezier curves and stored the data of curves and generated the character of NüShu described by curves. The result shows that this method is very simple and reduces the storage space greatly. The generated curve contour is the same as the original font and it also can eliminate burrs caused by hand-writing. This method is also the basis of computer processing of NüShu characters.

Key words: NüShu; bitmap font; curve contour character; feature point; quadratic Bezier curve

0 引言

女书, 又叫做女字, 是由女性创造并只在女性中流传、使用的一种神秘符号, 是世界上唯一的女性文字^[1]。目前女书的传承主要是通过手写的方式进行, 然而随着女书传人的相继去世, 女书已经处于濒临灭绝的状态。数字化女书对女书的传承和发展具有重大的意义, 其中最关键的问题之一是解决女书在计算机中的存储与显示。

曲线轮廓字形具有较小的数据存储量和更好的逼近效果, 并且起到了真正的“无级变倍”的效果, 消除了因放大而出现的锯齿现象^[2]。然而目前女书在计算机中大多以点阵图片的形式存储, 这种方式存储量大, 且字形在缩放过程中易出现锯齿现象。虽然目前也有女书 TrueType 字库出现, 但女书点阵字形到 TrueType 格式的曲线轮廓字形的转换是由修字工人手工完成的, 此工作耗时费力且效率很低, 研究由女书的点阵字形到曲线轮廓字形的自动转换将大大减少工作量。Sarfaraz 等人^[3]提出了阿拉伯文字的曲线轮廓字形自动生成方法, 马小虎等人^[4]针对甲骨文提出了将点阵字形转换成轮廓字形的方法, 然而目前还没有自动生成女书曲线轮廓字形的方法。

与汉字字形相比, 女书呈菱形, 且包含圈和弧笔这两种汉

字所不具有的笔画类型, 弧度或大或小、变化多端。本文根据女书自身的字形特点, 结合传统字形生成技术, 利用图形学中的曲线拟合方法可不受字形规则限制的特点, 设计了一种女书曲线轮廓字形自动生成方法, 包括以下步骤: a) 对扫描进计算机的女书图片进行预处理; b) 提取其边界轮廓并跟踪出轮廓数据; c) 提取轮廓特征点; d) 得到描述字形轮廓的直线和二次 Bezier 曲线; e) 存储曲线字轮廓数据, 得到女书曲线轮廓字形。其中针对现有的特征点提取方法在点阵字形上应用的缺点, 提出了一种基于最大误差值的特征点提取方法。由于二次 Bezier 曲线中间控制点计算简单, 且为了与 TrueType 字库格式一致, 本文采用直线和二次 Bezier 曲线来描述轮廓段。

1 女书轮廓数据提取

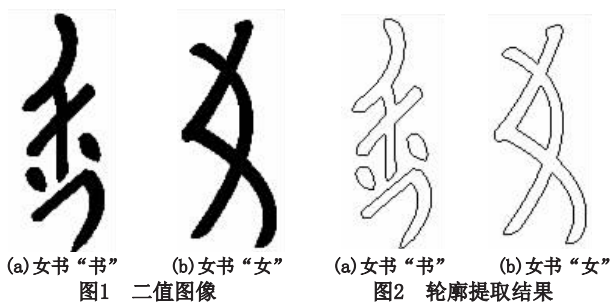
1.1 预处理和轮廓提取

传递、记载女书的物品主要有四种——书、纸、扇、巾^[1]。要获取女书, 必须扫描女书图片作为输入, 但由于女书物品数量庞杂、年代久远、保存不善等原因, 扫描出来的图片一般质量不高, 所以必须对它们进行预处理, 以满足字形处理的需要。预处理过程包括灰度化、图像增强、平滑滤波, 最后二值化得到二值女书图像, 如图 1 所示。

收稿日期: 2012-01-13; **修回日期:** 2012-02-24 **基金项目:** 国家自然科学基金资助项目(60975021); 中南民族大学基本科研业务费专项基金资助项目(CZQ11026)

作者简介: 严红娟(1988-), 女, 湖北公安人, 硕士, 主要研究方向为人工智能、数字图像处理(Lyan198810@126.com); 王江晴(1964-), 女, 教授, 博士, 主要研究方向为人工智能、数字图像处理; 李波(1975-), 男, 讲师, 博士, 主要研究方向为数字图像处理、模式识别。

轮廓提取的目的是获得图像的轮廓特征,本文采用“掏空内部点”的方法^[4]提取二值图像的边界轮廓。图2是对二值图像(图1)进行轮廓提取的结果,其中图2(a)只包含外轮廓,图2(b)包含内轮廓和外轮廓。



1.2 女书轮廓数据跟踪

轮廓跟踪的目的是为了获得连通轮廓的数据信息。在跟踪的过程中需要使轮廓点集有序化,使得女书字形的外轮廓点集按顺时针方向排列,内轮廓点集按逆时针方向排列。文献[4]中使用的模板(图3(c))能完整地跟踪出轮廓数据,但没有具体描述如何保证轮廓数据的有序性。为了既保证轮廓点集的有序性,又能完整地跟踪出轮廓数据,本文综合采用 3×3 顺时针模板(图3(a))、逆时针模板(图3(b))、无断点模板(图3(c))跟踪轮廓数据。

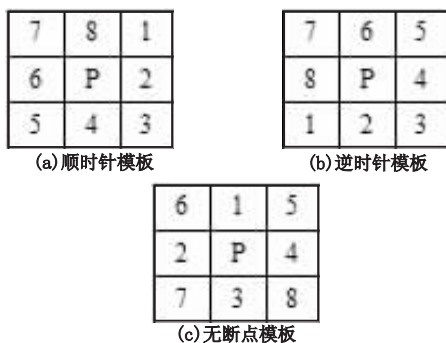


图3 3×3 跟踪模板

对于一条连通的轮廓,首先判断此轮廓线是内轮廓还是外轮廓,若是外轮廓则用顺时针模板跟踪起始的两个点,保证外轮廓的顺时针方向;若是内轮廓则用逆时针模板跟踪起始的两个点,保证内轮廓的逆时针方向。然后用无断点模板跟踪整个轮廓线数据,保证跟踪出的轮廓数据无断点。图4为跟踪出的轮廓数据,其中不同的连通轮廓用不同的颜色标记出来(见电子版)。

2 轮廓特征点提取

轮廓跟踪得到的大部分数据点对于字形数据的存储是冗余的,而且跟踪出的轮廓线曲率变化大,不可能用一条直线或者二次 Bezier 曲线直接拟合,所以必须通过提取特征点将轮廓线分段以减少存储量。

2.1 特征点提取方法分析

特征点能够代表边界曲线的特征,主要包括角点、切点和拐点等。目前具有代表性的特征点检测方法主要有角检测法和多边形逼近法^[5]两大类。

文献[6]使用基于最大距离迭代的方法提取特征点,该算法在采用文献[7]的方法提取尖点的过程中需要求出某个区间内的曲率最大值,而这个区间不好确定。文献[2]在文献

[6]的基础上进行了改进,针对向量字轮廓提出了三段式提取特征点的方法,由于向量字轮廓曲线已离散化,所以这种方法不适用于点阵字形。文献[8]描述了基于弦高度值的多边形逼近方法,但此方法需要多次重复计算各个像素点的弦高度值,计算量大。文献[9]采用求局部最小夹角的方法,在实现过程中 k 的值不好确定,而且此方法不能检测出一长段曲率缓慢变化的曲线段上的特征点,所以在曲线拟合的最后阶段需要设置分段点对原始轮廓再进行分段,以得到与原轮廓基本一致的曲线。

女书中含有大量的弧笔,且弧笔通常由一段较长的缓慢变化的曲线段来表示,所以用角检测的方法在这样的曲线上检验不出特征点,需要进行特殊的分段处理。本文提出一种新的特征点提取方法,此方法基于原始轮廓点到逼近线段的最大误差值,迭代删除轮廓上的冗余特征点,算法简单,计算量小,且由于在提取特征点阶段即考虑原始轮廓与拟合线段之间的误差值,所以在拟合阶段能更加准确地拟合特征点之间的轮廓段,不需要再对轮廓段进行分段处理。

2.2 算法实现

本算法依赖于最大误差值 $E_{\max}$,对于相邻特征点 P_i, P_k 之间的轮廓段,设 $e_j (j=1, 2, \dots, n)$ 表示原始轮廓上的点 $P_j (j=1, 2, \dots, n)$ 到线段 $P_i P_k$ 的距离,则 $E_{\max}$ 表示 $e_j (j=1, 2, \dots, n)$ 的最大值。对于一条封闭的轮廓,提取特征点的算法描述如下:

- 初始化阈值 d_k ,且将原始轮廓点集初始化为特征点集 $\{P_{\text{init}}\}$,并将 $\{P_{\text{init}}\}$ 中第一个点 P_{init1} 设为初始点 P_i 。
- 删除冗余的特征点,这是一个迭代的过程,其过程如下:

```
repeat
     $P_i \leftarrow P_{\text{init1}}, P_j \leftarrow P_{i+1}, P_k \leftarrow P_{i+2}$ 
repeat
    计算  $P_j$  到线段  $P_i P_k$  的距离  $d_j$ 
    if  $d_j \leq d_k$  then
        删除  $P_j$ ;
         $P_j \leftarrow P_k, P_k \leftarrow P_{k+1}$ 
    else
         $P_i \leftarrow P_k, P_j \leftarrow P_k, P_k \leftarrow P_{k+1}$ 
    end if
until  $P_j = P_{\text{init1}}$ 
 $P_{\text{init1}} \leftarrow$ 使  $d_j$  最大的特征点;
 $d_k = d_k + \Delta d$ ;
until  $E_{\max} > E_t$ 
```

对于不共线的相邻三个特征点 P_i, P_j, P_k ,点 P_j 到 $P_i P_k$ 的距离值最小为 $1/\sqrt{5} = 0.447$,因此本文阈值 d_k 取0.5,此时通过第一次迭代即可以删除共线点以及轮廓上的噪点。为减少迭代次数,且避免增量太大产生大量满足终止条件的冗余点,本文 d_k 的增量 Δd 取0.5。将终止条件设为 $E_{\max} > E_t$,随着迭代次数的增加, $E_{\max}$ 的值随之增大,当 $E_{\max}$ 的值大于 E_t 时,则认为原始轮廓线上的点到逼近线段的距离超过了容错范围,停止迭代删除过程,此时剩余的点即为特征点。图5展示了当 E_t 取2时特征点提取结果,其中小方块表示为提取出的特征点。

3 女书轮廓曲线拟合

提取出的特征点将每一条连通轮廓分成不同的轮廓段, m 个特征点 $cp_1, cp_2, \dots, cp_m$ 将连通轮廓分为 m 个轮廓段 $\text{seg}_1, \text{seg}_2, \dots, \text{seg}_m$,第 i 个轮廓段包含轮廓点集 $p_u, p_{u+1}, \dots, p_{u-1}$,

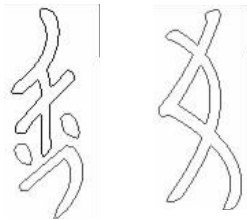


图4 轮廓数据跟踪得到的轮廓线

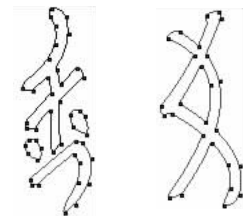


图5 特征点提取结果

p_w , 则 $p_u = cp_i$, 且

- 当 $1 \leq i < m$ 时, $p_w = cp_{i+1}$;
- 当 $i = m$ 时, $p_w = cp_1$ 。

3.1 确定拟合曲线类型

对于连通轮廓上不同的轮廓段,需要判断此轮廓段该用直线还是二次 Bezier 曲线来拟合。判断的依据为:对于特征点 cp_i, cp_{i+1} 之间的轮廓段,求得 cp_i, cp_{i+1} 之间轮廓线上的点到线段 cp_i, cp_{i+1} 的最大距离 d ;若 $d < 1$,则说明 cp_i, cp_{i+1} 之间的轮廓数据点基本在同一条直线上,可用直线拟合此轮廓段;否则说明此轮廓段是一条曲线,用二次 Bezier 曲线拟合。

3.2 反求二次 Bezier 曲线中间控制点

二次 Bezier 曲线可以用如下参数方程描述:

$$Q(t) = (1-t)^2 P_0 + 2t(1-t) P_1 + t^2 P_2 \quad (1)$$

其中: $0 \leq t \leq 1$ 。则基函数为

$$\begin{cases} \varphi_0(t) = (1-t)^2 \\ \varphi_1(t) = 2t(1-t) \\ \varphi_2(t) = t^2 \end{cases} \quad 0 \leq t \leq 1$$

其中,末端控制点 P_0, P_2 是曲线段上的特征点,即 $P_0 = p_u, P_2 = p_w$ 。

本文采用参数化弦长来估计参数 t :

$$t_k = \begin{cases} 0 & k = u \\ \frac{|p_u p_{u+1}| + |p_{u+1} p_{u+2}| + \dots + |p_{k-1} p_k|}{|p_u p_{u+1}| + |p_{u+1} p_{u+2}| + \dots + |p_{w-1} p_w|} & u+1 \leq k \leq w-1 \\ 1 & k = w \end{cases}$$

要拟合出理想的曲线必须找到二次 Bezier 曲线段的中间控制点 P_1 。本文采用参数型最小二乘法求解 P_1 ,用方差来衡量逼近程度,有

$$\phi(P_1) = \sum_{k=u}^w [Q_k(t_k) - p_k]^2$$

即:

$$\phi(P_1) = \sum_{k=u}^w | \sum_{j=0}^2 P_j \varphi_j(t_k) - p_k |^2 \quad (2)$$

要使式(2)最小,必须使 $\partial \phi / \partial p_1 = 0$, 即:

$$\frac{\partial \phi}{\partial p_1} = 2 \sum_{k=u}^w \sum_{j=0}^2 [P_j \varphi_j(t_k) - p_k] \varphi_1(t_k) = 0$$

即:

$$\sum_{j=0}^2 P_j \sum_{k=u}^w \varphi_j(t_k) \varphi_1(t_k) = \sum_{k=u}^w p_k \varphi_1(t_k) \quad (3)$$

解方程(3)即得 P_1 的坐标值。

图6展示了利用直线和二次 Bezier 曲线拟合字形轮廓的结果。其中彩色线条表示原轮廓,黑色线条表示拟合得到的曲线轮廓(见电子版)。从结果可以看出,提取出的特征点能较好地拟合原始轮廓,不需要再对原始轮廓进行分段。图7(b)为对手写女书的原始轮廓(图7(a))拟合的结果。从结果可以看出,拟合出来的曲线轮廓较原始轮廓更加光滑,能较好地消

除因手写而造成的毛刺现象。

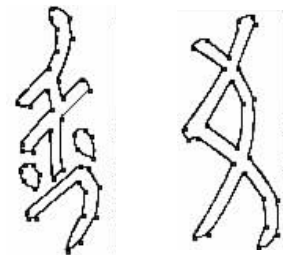


图6 曲线拟合结果

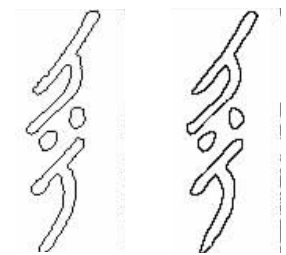


图7 曲线拟合结果对比

3.3 生成曲线轮廓字形

将前面处理得到的女书字形轮廓描述数据按照 TrueType 字库的存储格式存放,生成标准的 TTF 女书字库文件,至此完成了女书点阵字形到曲线轮廓字形的自动转换。

4 实验结果及分析

本文选取点阵大小为 70×170 的女书样本图片,在 Windows XP 操作系统、Microsoft Visual C++ 6.0 环境下编程实现。由于目前没有特征点提取算法应用于女书,将本文算法与文献[9]方法在女书字形上实现的结果进行比较。本文提取特征点算法($E_t = 2$)对比用文献[9]方法($T = 160^\circ$)实现结果的比较如表1所示,下面就特征点分布情况、数据压缩量及运行时间等方面进行分析。

表1 特征点提取结果比较

女书 字形	原始轮廓点个数	方法	特征点个数	运行时间 T/s
“书”	561	文献[9]细分前	37	1.081
		文献[9]细分后	58	
		本文方法	45	
“女”	533	文献[9]细分前	21	0.859
		文献[9]细分后	41	
		本文方法	26	
“山”	245	文献[9]细分前	14	0.532
		文献[9]细分后	28	
		本文方法	21	

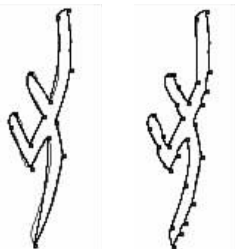
从表1可以看出,本文方法提取出的特征点个数介于文献[9]方法细分前和细分后提取的特征点个数之间,用文献[9]的方法对女书“山”字提取特征点并对轮廓段进行拟合的结果如图8所示,细分前特征点个数为14,从图8(a)可以看出虽然特征点的个数较少,但是对于具有一长段缓慢变化的轮廓段来说,会漏掉一些特征点,不足以准确地拟合出原始的字形轮廓,虽然通过添加分段点对轮廓进行细分能克服这一缺点,但是细分后特征点个数为28,出现了一些冗余的关键点(图8(b)),而且在特征点处相邻轮廓段的拟合曲线连接不够光滑。本文方法提取出的特征点个数为21,从图9可以看出本文方法提取的特征点相对于文献[9]方法细分前能更加准确地拟合原始字形轮廓,相对于文献[9]方法细分后特征点个数明显减少,从而减少了存储量,由于不需要再对轮廓进行分段,大大减少了运行时间。图10是生成的女书曲线轮廓字形,该字形与原字形基本一致。

5 结束语

本文实现了一种用直线和二次 Bezier 曲线自动生成女书曲线轮廓字形的方法。该方法能够把质量不高的女书图片自动转换为用曲线轮廓描述的女书字形,这种 (下转第3564页)

(上接第 3560 页)

字形存储量小,能较好地保持字形特征的稳定性和灵活性,使得女书在计算机中能任意缩放、变形而不会影响字形的质量。本文方法为在计算机中处理女书奠定了基础,同时也为其他少数民族文字在计算机中的使用提供了依据。



(a) 细分前 (b) 细分后
图8 文献[9]方法拟合结果



图9 本文方法拟合结果



图10 女书曲线轮廓字形

参考文献:

- [1] 易叶舟. 世界唯一的女性文字——江永女书漫谈[J]. 华夏文化, 2008(3): 39-40.
- [2] 陈登梅, 李学庆, 李德生. 一种新的曲线字库自动生成方法[J]. 系统仿真学报, 2006, 18(10): 2995-2997, 3002.
- [3] SARFRAZ M, KHAN M A. Automatic outline capture of Arabic fonts [J]. Information Sciences, 2002, 140(3): 1327-1336.
- [4] 马小虎, 黄文帆, 顾绍通, 等. 甲骨文点阵字形转换为甲骨文轮廓字形的方法[J]. 语言科学, 2004, 3(3): 3-11.
- [5] 刘永奎, 刘向东, 王春霞. 一个抽取边界曲线特征点的新算法[J]. 计算机应用研究, 2006, 23(6): 148-152.
- [6] 荆晶, 刘玉兰, 葛庆平. 一种基于迭代提取关键点的曲线拟合新算法[J]. 首都师范大学学报: 自然科学版, 2004, 25(12): 22-25.
- [7] 刘玉兰, 葛庆平. 一种间接提取轮廓特征点的算法[J]. 计算机工程与应用, 2004, 40(10): 51-52, 70.
- [8] 孙爽滋, 谷欣超, 杨勇, 等. 飞机图像的轮廓提取与多边形拟合研究[J]. 长春理工大学学报: 自然科学版, 2009, 32(3): 447-449.
- [9] 沈娟, 马小虎. 甲骨文的曲线轮廓字形自动生成系统[J]. 计算机应用与软件, 2009, 26(1): 67-68, 114.