第7卷第5期
1996年5月

软 件 学 报
JOURNAL OF SOFTWARE

Vol. 7, No. 5
May 1996

一个高质量汉字笔划字形到轮廓字形的转换系统*

董韫美　陈海明

(中国科学院软件研究所　北京　100080)

摘要　本文介绍一个基于三阶 Bezier 曲线的字形转换系统 SOCS(stroke to outline convertion system). 该系统作为汉字字形设计系统 CCDS(Chinese character design system)的后端,把 CCDS 产生的多笔划曲线轮廓字形数据加工成整字曲线轮廓字形数据. 其结果适于多种高精度输出,如 PostScript 印字机. SOCS 系统具有一定的通用性,它不只限于接受 CCDS 的加工结果,还可以作为独立的系统使用.

关键词　字形设计,字形表示,字形转换,轮廓字形,计算机图形学.

当代电子印刷技术,对汉字品种和字形质量不断提出新的和越来越高的要求. 然而,由于汉字结构的复杂和字符集的庞大,汉字字模的设计难度很大,从字稿得到适于高质量电子印刷的字形数据是艰苦的劳动,与对汉字的需求形成了巨大的反差. 在这种形势下,出现了各种以计算机为主要手段来进行汉字字模创作和字形数据获取的字形制作技术.

汉字字形设计系统是汉字字形制作的新型、先进的手段. 用汉字字形设计系统来产生汉字字形数据,计算机介入了从汉字字形的设计到字形数据产生的全过程,不仅可以设计新的汉字字形,而且生产效率大为提高,是适应汉字字形多品种、高质量需求的根本途径.

汉字字形设计系统一般采取由笔划来构造整字的方法. 产生的字形由互相交叉的笔划组成(一个笔划是一个封闭区域,由区域的轮廓线定义). 而在实际应用中,却往往希望用整个字形的轮廓线描述形式(相当于将各笔划描填出来得到整个字形之后,再提取轮廓线). 因为与笔划描述形式的字形相比,整字轮廓线描述的字形不仅生成速度较快,数据量亦较小(1.1 给出例子). 而且当用于某些特殊领域时,如文字刻绘、广告,只能用整字的轮廓线描述形式. 为了得到整字轮廓描述,需把前者转换为后者.

轮廓线又可分为向量近似表示和曲线表示. 曲线表示方式用光滑曲线如样条曲线表达了汉字字形的图形连续信息,表达能力强,精度高,与坐标系的选择无关,是描述高质量汉字字形理想的方式. 目前国际上先进的西文字形表示,如 PostScript 的 Type1,Apple 的 TrueType,都采用整字曲线轮廓表示,这无疑也是汉字字形的方向. Microsoft 公司于 1993－10

---

* 作者**董韫美**,1936年生,中国科学院院士,研究员,博士导师,主要研究领域为软件复用技术,字形设计技术等. **陈海明**,1966年生,助研,主要研究领域为软件复用技术,字形设计技术.

本文通讯联系人:董韫美,北京 100080,中国科学院软件研究所

本文 1995－03－01 收到修改稿

推出的中文版 Windows 3.1,就采用了 TrueType 中文字库.

高质量汉字字形设计系统CCDS(Chinese character design system)[1,2]是汉字字形设计系统中最具代表性的系统.为了使CCDS产生的笔划轮廓字形数据转换为整字轮廓字形数据,我们设计和实现了笔划字形到轮廓字形的字形转换系统SOCS(stroke to outline convertion system).

本文把基于笔划轮廓描述的字形称作笔划字形;基于整字轮廓线描述的字形称作轮廓字形;而把字形描述方式的转换称作字形转换.

1 字形转换系统SOCS的设计

1.1 CCDS字形简介

CCDS采用参量图形学方法,按笔划、子字、整字的途径来设计汉字字形.

1个用CCDS设计出来的汉字字形见图1(a). CCDS中字形的基本单位是元笔划,它是用三阶Bezier曲线和直线来刻画的1个平面封闭子图形.由于汉字具有很复杂的结构,为得到高质量的汉字,元笔划也可以有任意复杂的形状.

(a) (b)

图1

CCDS的着眼点是提供给用户满足自己设计需要与表现设计风格的能力,不强调字形发生速率,因而产生的是笔划字形数据.这种字形结构子图形较多,子图形间又是互相交叠,填色较为费时.以图1(a)为例,用本文所述系统转换所得的轮廓字形表示见图1(b).对这2个图形分别在激光印字机上用PostScript语言输出.填色时间分别是890和700(单位ms);另外,封闭区域的个数由前者的26个降为后者的3个;表示字形所需的曲线、直线总数前者为126,后者为86;字形描述数据量前者为498Bytes,后者为305Bytes.可见实际应用的字形数据以轮廓字形为好,它不仅生成速度较快,而且描述更简洁,数据量较小,数据压缩率更高.

1.2 SOCS系统设计

从使用上来说,字形转换系统应既能与字形设计系统CCDS联合使用,又能脱离它而单独使用.

SOCS系统要解决的问题包括以下几个基本方面:

①字形转换

把笔划字形经过计算转换为等价的轮廓字形.这是系统最主要的功能.它包括原字形数据的解释、预处理、字形转换计算、新字形数据的输出等环节.这一过程应做到自动化,但又必须具备人机交互的功能,以便在必要时能进行人工干预,如对字形编辑.

所接收的输入,可以是字形设计系统输出的零星字形,也可以是完整的字形数据库.

②字形数据组织

经字形转换生成的轮廓字形数据需组织成新的字形数据库.由于应用领域、用户需求的

不同,所需的轮廓字形数据库的格式可能不同.因此在字形转换后应产生内部标准格式的字形数据,以方便系统的使用者组织不同的字形数据库.

③字形数据维护和字形质量验证

不论是笔划字形数据还是新产生的轮廓字形数据,在存储、传送过程中均可能发生错误,因此需要有维护功能.轮廓字形的正确性及精度如何,也应能进行检验.

只有具备了上述功能,字形转换系统才是完整的.

考虑到 PostScript 输出设备日益而广泛的应用,对 PostScript 字形数据的支持也就成为一个很有价值的方面.

SOCS 系统的组成见图 2.整个系统由字形转换、字形数据组织、辅助工具和 ps 接口 4 个部分组成.SOCS 系统的数据可分为 3 类:

①输入数据——一般是CCDS的笔划字形数据,但也可以是其它种类的笔划字形数据,例如向量式的;②内部数据——字形转换后产生的轮廓字形数据,以标准格式存储,供 SOCS 系统自己使用,不对普通的轮廓字形用户提供;③输出数据——轮廓字形数据库,可以是特别指定的格式.

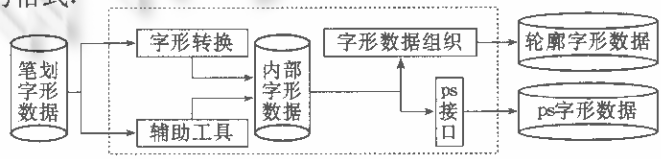


图2 SOCS系统

字形转换部分完成笔划字形到轮廓字形的转换,以标准格式输出字形数据.能够对笔划字形数据的非法情况报警,并进入交互状态;也可由用户设置为交互状态.有一个交互字形编辑器,可以在转换前对字形进行编辑,完成后退出交互状态,继续进行字形转换.

字形数据组织部分收集由字形转换程序加工得到的字形数据,按某种格式组织成字形数据库.

辅助工具部分包括进行字形数据文件维护、字形显示以及精度分析的一组基本工具.

ps 接口部分提供一个到 PostScript 语言的接口,包括字形数据格式的转换和字形数据解释程序(PostScript 程序).

2　字形转换

字形转换部分的流程如图 3.

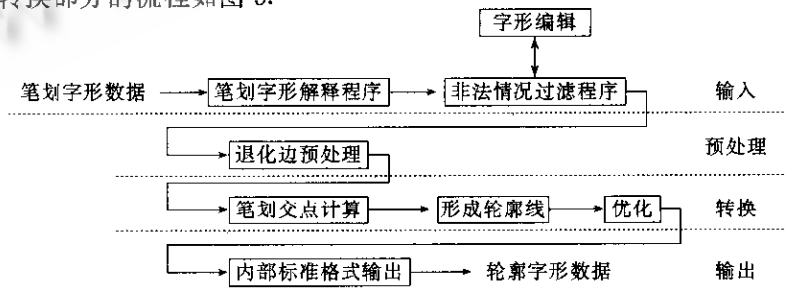


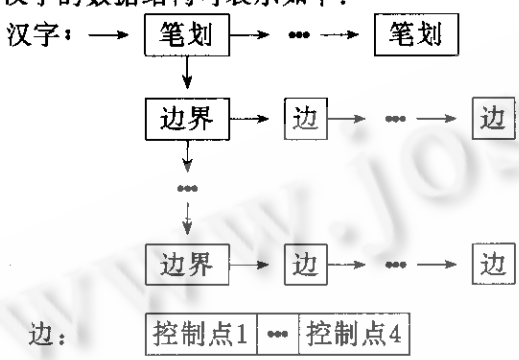
图3 字形转换流程

进行字形转换,1 种方法是把笔划的轮廓曲线变换为近似直线后,再进行转换.这种作法因为只要求直线计算,所以,加工的复杂程度低,但会损失字形质量.另 1 种方法是直接由各个笔划轮廓曲线进行转换,以得到整字轮廓曲线.这种方法加工复杂程度高,但字形质量好.

为保持字形质量,SOCS 系统的字形转换采取后 1 种方法.

汉字是元笔划的集合,元笔划边界由三阶 Bezier 曲线和直线组成.若元笔划的边全部为直线,则笔划退化为多边形,汉字为向量字形.

汉字的数据结构可表示如下:



若控制点2、3为空,则边为直线.

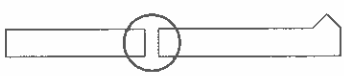


图4 笔划断接

2.1 输 入

输入部分读取笔划字形数据,转换为内部数据结构的表示形式,并处理非法情况."非法情况"即在输入的笔划字形数据中可能遇到的不合程序约定的情况,包括:(1)笔划字形数据与转换程序约定不符,如边界反向、自交等情况;(2)笔划字形错误,如笔划出现"断接"等,如图 4.在遇到上述情况后,系统给出提示,同时调用字形编辑器,进入交互状态,由使用者决定如何处理.

非法情况过滤程序还接受一个特殊的"非法情况",即由使用者指定的进入交互状态的参数,以便使用者进入交互状态进行一些特殊处理.

对于上述第 2 类非法情况,不同的字形库会有不同的情况,应在转换前有大概的了解.这就需要用系统提供的辅助工具进行浏览、显示.在此基础上,便可写出不同的非法情况处理程序,或在非法情况过滤程序中添加对相应情况的处理.

2.2 预处理

输入的笔划字形可能存在退化边.所谓退化边,指曲线边的控制点共线或 2 条相邻直线边共线.退化曲线边应成为直线边,2 条退化直线边应合并成为 1 条直线边.

2.3 转 换

设笔划汉字由 n 个笔划组成,n 个笔划分别记作 $R_1,R_2,\ldots,R_n$.在空白平面上依次加入 $R_1,R_2,\ldots,R_n$,并形成中间笔划.加入 R_i 时,若 R_i 与当前平面上的中间笔划相交,则 R_i 便与后者构成新的中间笔划;否则 R_i 加入平面上,仍为一个笔划.当加入 R_n 后,即得到轮廓字形.

以 $\{Q_1,\cdots,Q_t\}$ 表示两两无公共部分的笔划的集合,即 $Q_i\cap Q_j=\varnothing,i,j=1,\ldots,t,i\neq j$.

设加入 R_i 后,平面上形成的中间笔划集合为 $R_s^i=R_1\cup\cdots\cup R_i=\{R'_1,\ldots,R'_{i_0}\}$,$i_0\leqslant i$. 当加入 R_{i+1} 时,把 R_{i+1} 与 R_s^i 中的中间笔划 $R'_1,\ldots,R'_{i_0}$ 逐个比较,若 R'_j 与 $R_{i+1}(j=1,\ldots,i_0)$ 不相交,则把 R'_j 记入 R_s^{i+1},否则 R'_j 与 R_{i+1} 形成一个笔划,仍记为 R_{i+1}(R'_j 被 R_{i+1} 吸收),最后将 R_{i+1} 记入 R_s^{i+1} 中,即得 R_s^{i+1}. 上述过程可以表示如下:

$R_s^{i+1}=\varnothing$;

for ($j=1;j\leqslant|R_s^i|;j++$)

　if($R'_j\cap R_{i+1}=\varnothing$) $R_s^{i+1}=R_s^{i+1}\cup\{R'_j\}$ else $R_{i+1}=R_{i+1}\cup R'_j$;

$R_s^{i+1}=R_s^{i+1}\cup\{R_{i+1}\}$

其中 $|R_s^i|$ 表示集合 R_s^i 中的中间笔划的个数.

相交的笔划,要计算出它们的交点,再提取外轮廓段,即形成轮廓线. 如图 5.

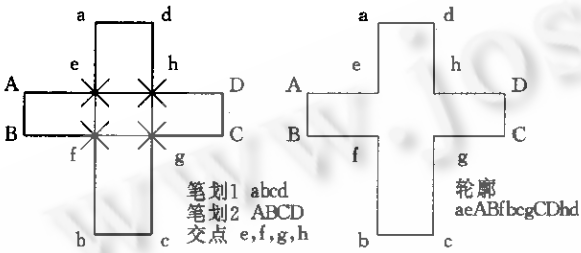


图5

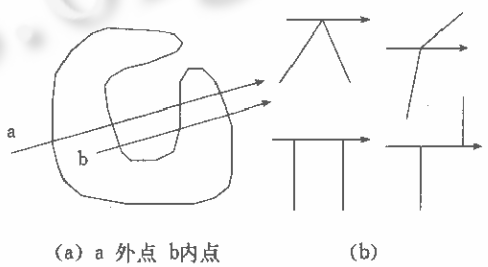


图6

Bezier 曲线求交,效率和健壮性是 2 个需要优先考虑的因素. 广为使用的且较为简单的方法是基于 Bezier 几何特性和 De Casteljau 算法的分段逐步逼近方法(Subdivision Method),具有健壮性,但速度较慢,分段收敛速度与曲线的几何分布有较大的关系. 近年来有人提出了基于 Bezier 剪切[4]的方法,这种方法仍是一种分段方法,具有健壮性,但迭代量较小,收敛速度较快. 我们就采用了这种方法. Bezier 剪切方法的要点在于迭代过程中只考虑曲线上可能包含交点的曲线段,利用线性运算求得这样的曲线段,裁去不必要部分,对 2 曲线交替进行上述过程,很快可得到交点. 它不象一般的分段方法不加区别地对待每一个曲线段,从而大大减少了迭代次数,提高了效率.

判断 1 个轮廓段是否外轮廓段,最简单的方法是射线法. 即任取该轮廓段上一点(如中点),从该点出发作一射线,若该射线与另一笔划相交偶次(包括 0 次),则该轮廓段为外轮廓段. 如图 6(a). 实现时,要注意射线通过端点的情况,如图 6(b).

在字形转换过程中,仍可能产生退化边,所以在形成轮廓线之后仍然要进行退化边的优化处理. 优化的另一个含义是如果 2 相邻边可用 1 个 Bezier 曲线来表示,则把 2 个边用 1 个长曲线边替换(现在尚未做).

2.4 输　出

输出部分把得到的轮廓字形数据以标准数据形式输出到文件中.

3 其它部分

3.1 基本辅助工具

本节简介几项主要的基本辅助工具.

3.1.1 字形数据文件的维护工具

字形数据文件是按某种格式存储多个字形数据的二进制文件,无法直接阅读;字形数据文件中包含字形数据的有用信息,有时需提取;字形数据文件在传送过程中可能有数据位出错,因而需要维护.字形数据文件的维护工具主要面对这些问题,包括字形数据文件阅读、浏览、统计和修改.

字形数据文件阅读是对字形数据的解释和表达,以一定形式表达控制信息和以十进制形式表达坐标值.浏览是对字形数据的有用信息的提取和表达.统计则是对字形数据的有用信息的提取和加工.如每字的平均曲线数、存储量.字形数据文件的修改是指允许用十进制或十六进制对数据进行修改.

3.1.2 屏幕图形设计与显示工具

该工具是在SUN-3/60工作站的cgi图形接口系统上实现,用于设计实验图形,显示转换结果与显示汉字.

设计包括以给出控制点的方式和点拟合的方式设计三阶Bezier曲线图形.

显示包括显示任意三阶Bezier曲线图形及字形,和自动连续显示多个字形.

利用这些工具可设计图形对转换方法进行检验并直观地看到转换效果,也可对转换过程的每一步显示转换结果,以动态观察转换行为,还可连续显示字形对字形数据文件正确性进行检测.

3.1.3 Bezier曲线区域的扫描转换与点阵的运算、显微工具

用于对转换算法进行精度分析.扫描转换可对凸的或凹的三阶Bezier曲线区域进行扫描转换,得到点阵.点阵显微即将点阵以可分辨的点显示出来.通过对转换前后图形的点阵的比较并显微,可精确地得出2个图形间的误差程度.

3.2 字形数据组织

这一部分收集由字形转换程序加工的字形数据,组织成字形数据库文件,可以进行字形数据格式的转换.系统的使用者可以根据需要,编写不同的字形数据组织程序.

3.3 ps接口

PostScript是当前页面描述语言的事实上的国际标准.ps接口专门针对PostScript输出设备,采用PostScript Type3方式即用户自定义方式对字形数据进行格式转换和字库组织,给出相应的PostScript解释程序.事实上,用SOCS系统转换的轮廓字形的字样,都是在PostScript印字机上输出的.

4 总 结

4.1 系统的使用和达到的指标

系统在SUN-3/60工作站上实现,用C语言写成,程序量约7×10^3行.并已移植到PC386的DOS环境下.系统实现后,已对CCDS产生的2套汉字(每套6 763个)进行了转换.经过对字形逐个核对,转换结果正确.可用于高质量的汉字输出.系统的一些统计数据如下:

平均数据量	CCDS笔划字形	399 bytes/字,轮廓字形311 bytes/字
平均线数	CCDS笔划字形	直线63.937/字,曲线47.687/字

轮廓字形　　直线 41.030/字,曲线 36.604/字

平均转换时间　约 6s/字

还对其它来源的向量笔划字库进行了转换,在 386 兼容机(主频 33MHz)上的平均转换时间约 10s/字,在 PC486 机(主频 33MHz)的转换时间则明显快得多.更多的实验有待于更多的字源.

4.2 系统的分析

SOCS 系统直接在曲线水平上进行字形转换,算法复杂度较高,好处是字形质量好和数据量较小.

SOCS 系统的转换方法并不限定字形轮廓线的曲线类型,在实现上,对不同的曲线只需换上相应的求交部分即可.系统的内部字形为标准数据格式,对于不同的笔划字形数据,只需修改输入部分;需要不同的轮廓字形库时,只需写相应的字形数据组织部分,即一个数据格式加工程序.事实上,SOCS 系统所配备的 ps 接口就是一个专门的字形数据组织程序.由于以三阶 Bezier 曲线描述高质量汉字字形是今后的趋势,及根据 SOCS 系统的上述特点,SOCS 系统及其方法对其它基于笔划的字形制作系统也有同样意义,有一定的通用性.

参考文献

1 董韫美,李开德.参量图形学方法与字形设计.IEEE Beijing Section First Annual Conf. Proc., Beijing, IEEE Beijing Section, 1987.130～133.

2 Dong Yunmei, Li Kaide. A parametric graphics approach to Chinese font design. In: R A Morris, J Andre eds., Raster Imaging and Digital Typography II, Cambridge University Press, 1991.156～165.

3 陈海明.笔划汉字到轮廓汉字的转换系统 SOCS[硕士论文].中国科学院软件研究所,1992.

4 Sederberg T W, Nishita T. Curve intersection using Bezier clipping. CAD, 1990,**22**(**9**):538～549.

CONVERSION SYSTEM OF HIGH QUALITY CHINESE CHARACTERS FROM STROKE FONT TO OUTLINE FONT

Dong Yunmei　Chen Haiming

(*Institute of Software　The Chinese Academy of Sciences　Beijing*　100080)

Abstract　This paper introduces the SOCS (stroke－outline conversion system) for Chinese font data processing. As a post processing system of the CCDS (Chinese character design system), SOCS converts the font data of multi－stroke curved outline, which is produced by CCDS, into the font data of whole－character curved outline. The later is especially suitable for high resolution outputting, for example the PostScript printer. SOCS is not only used for CCDS, but also used as an independent font conversion system.

Key words　Font design, font description, font conversion, outlined font, computer graphics.