

甲骨文字形动态描述库及其字形生成技术研究

栗青生^{1,2,3,†} 吴琴霞^{2,3} 杨玉星^{2,3}

1. 武汉理工大学信息工程学院, 武汉 430070; 2. 安阳师范学院计算机与信息工程学院, 安阳 455002;
3. 甲骨文数字化工程技术研究中心, 安阳 455002; † E-mail: aylqs@163.com

摘要 基于甲骨文字形多变, 异体字多等特点, 提出一种甲骨文字形描述方法, 将甲骨文字形进行矢量描述。建立了甲骨文字形描述库, 通过字形描述库自动生成多种甲骨文字形, 有效地解决了由于使用轮廓字形描述甲骨文字形而存在的字形动态编辑和字形变换的困难, 为古文字的数字化编辑提供了一个新思路。

关键词 甲骨文; 字形描述库; 有向笔段; 字体

中图分类号 TP391

Dynamic Description Library for Jiaguwen Characters and the Reserch of the Characters Processing

LI Qingsheng^{1,2,3,†}, WU Qinxia^{2,3}, YANG Yuxing^{2,3}

1. School of Information Engineering, Wuhan University of Technology, Wuhan 430070; 2. School of Computer and Information Engineering, Anyang Normal University, Anyang 455002; 3. Institute of Digital Inscriptions on Bones/Tortoise Shells, Anyang 455002; † E-mail: aylqs@163.com

Abstract The Jiaguwen character glyph is changeable, and some of them have many variant forms. Considering these properties of Jiaguwen characters, a human computer interactive dynamic description method for Jiaguwen character is proposed. The method describes Jiaguwen characters by stroke-segments-vector and stroke elements, and it can solve the problem that the internal code space is too small to encode all Jiaguwen characters. The dynamic description of Jiaguwen character glyph provides a solution to the problem of the identification of Jiaguwen character glyph which is caused by the lack of standardization for Jiaguwen character glyph. A human computer interactive dynamic experiment platform is developed based on the dynamic description method. The platform can perform many functions such as inputting, searching and editing of Jiaguwen character glyph.

Key words Jiaguwen; glyph description; stroke-segments-vector; stroke elements

随着计算机信息技术的发展, 古文字的数字化研究、印刷和出版等领域对甲骨文字形的需求越来越高, 如何让计算机像处理现代文字一样处理甲骨文字, 成为计算机学者和文字与语言学者共同关注的课题。多年来, 在甲骨文数字化过程中, 一直存在两个难以解决的问题: 一是甲骨文难以定形, 即甲骨文不像现代汉字, 很难确定每一个甲骨文字的字形结构和类型; 二是甲骨文字难以输入计算机。为此, 许多学者试图通过研究和设计字形库来

解决计算机中甲骨文字形的显示和打印问题^[1-4], 通过设计各种各样的输入编码方案来解决甲骨文的输入问题^[5-7]。多年来的实践证明, 对于规范的现代汉字, 可以使用这样的办法去解决; 但对于甲骨文字, 这一方法并不完全适用。首先, 甲骨文字形复杂, 存在较多的异体字和合体字, 并且有近三分之二的字还没有完全考释出来。其次, 甲骨文字形的规范性不像现代汉字一样有一个统一的标准, 从近几年出现的不同的甲骨文字形库来看, 许多学者在

国家自然科学基金(60973051)和河南省科技厅重点科技攻关项目(112102210375)资助

收稿日期: 2012-05-30; 修回日期: 2012-08-15; 网络出版时间: 2012-10-26 17:04

网络出版地址: <http://www.cnki.net/kcms/detail/11.2442.N.20121026.1704.006.html>

甲骨文字形的认同方面存在着很大的差距。香港中文大学的沈建华等^[8]确定了六千多个甲骨文字形(包括异体字),给甲骨文研究带来了极大的方便。但是,由于过度“规范化”,其中很多字形已经脱离了原始字形所表达的意义,有的还存在着部件结构和位置的错误^[9]。在古文字数字化新环境中,它所代表的为适应传统文本形式检索要求而形成的甲骨文检索系统,不但原有缺陷被放大,而且显现出若干新的问题^[10]。

由于甲骨文字的特殊性,在建立甲骨文的输入和编码方案之前必须先建立一个合适的解空间,使得不论是可识的(或可隶定的)还是不可识的,是同字异形的还是异字同形的甲骨文字都能在这个解空间中进行定义和计算^[11]。显然,从目前标准的 Unicode 编码系统的一字一码的编码特点和所能提供的有限的码位空间来看,这个解空间很难确定。

1 传统的甲骨文字形描述方法及存在的问题

汉字的计算机表示方式由以信息交换为目的的汉字内码、以汉字输入为目的的汉字输入码(外码)和以汉字输出为目的的汉字字形(字体)等几部分组成,甲骨文作为中国最早的汉字,当然也不能例外。

要在计算机中编辑和输入甲骨文,首要的任务是建立甲骨文字形库。目前通常的做法是:通过对甲骨文原始数据——甲骨文拓片的字形采集,取得甲骨文字的轮廓,然后使用字库制作系统做进一步的编辑处理,生成甲骨文字模,最后编码形成甲骨文字库。图1是传统的甲骨文字库制作过程示意图。

建立甲骨文字库的核心问题是编码和字形。近年来,围绕甲骨文字的编码问题,中外学者对甲骨文字形进行了长期深入的研究,总结起来有两类编码方法:一是使用现代汉字对应的编码方式对其进行描述,即现代汉字和它所对应的甲骨文字采用相同的编码方式;二是使用 Unicode 空间的 Private Use Area 区间进行重新编码,即现代汉字和它所对应的甲骨文字采用不同的编码方式。第一种方法的优点是保证了每一个甲骨文字和现代对应的文字编码的统一性,不足的是对于甲骨文而言,几乎每一个字都有异体字,而且大多数不只一个,如表1中是几个常用字的甲骨文异体字表。另外,由于目前还有近三分之二的甲骨文字没能被完全考释,因此从甲骨文研究和应用的角度来分析,将甲骨文视为一种和现代文字编码一一对应的字体编码方法不可取。第二种方法将甲骨文编码限定在 E000 到 F8FF 的 Private Use Area 区间内进行编码,这固然可行,但同样存在对异体字编码的困难。目前,对于甲骨文中的异体字的认识还没有统一的标准。在文献[8]中,

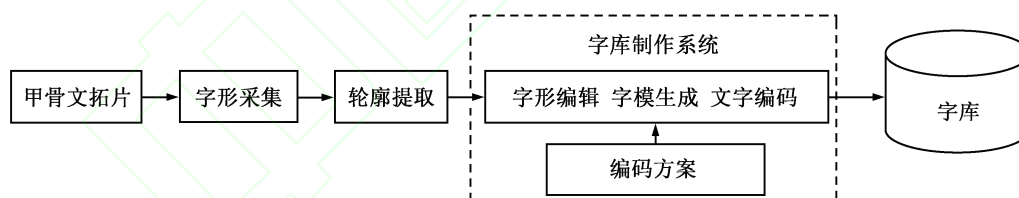


图1 传统的甲骨文字库制作过程

Fig. 1 Traditional Jiaguwen font library manufacture process

表1 部分甲骨文异体字表

Table 1 Variant forms part of the Jiaguwen

汉字	甲骨文	异体字1	异体字2	异体字3	异体字4
元					
天					
吏					
祥					
中					

异体字被限定为有限多个, 因此, 整个文字的规模和数量完全可以放在 Private Use Area 区间。但随着对甲骨文字研究的深入, 对字形结构不同的认识和处理, 都会导致文字和异体字的数量增加, 例如有学者提出形状和位置略有不同的甲骨文字都应视为异体字。这样, 有限的 Private Use Area 区间就不能容纳多达几万个甲骨文字了。

由于在有限的编码空间内不能实现所有的甲骨文字的有效编码, 甲骨文的编辑和处理就变得异常复杂。以上两种方法最大的不足是将甲骨文视为一种特殊字体, 就像现代汉字拥有“黑体”、“宋体”和“楷体”一样, 将甲骨文作为一种不变的字体, 这是不科学的。为此, 需要找到一种字形编辑方案, 使得甲骨文字形在使用过程中更容易编辑和修改。

计算机中通用字形库是以字体文件为单位, 每一个字体文件包含一类不同编码的汉字, 每一个汉字以字形的轮廓来描述(如图 2 所示)。甲骨文字形结构复杂, 使用字形的轮廓来描述更加复杂(如图 3 所示)。轮廓字形保证了字形输出的质量, 但不利于字形的编辑和字形的动态生成。

基于上述原因, 我们提出一种基于人机交互的甲骨文字形动态描述方法, 通过建立甲骨文字形动态描述库(Dynamic Description Library for Jiaguwen Characters, 以下简称 DDLJC), 使用有向笔段和笔元对甲骨文字形进行动态的矢量化描述。DDLJC 的

本质特征是将甲骨文字形用字形骨架来表示(如图 4 所示), 然后在字形骨架上增加特征点, 最后通过存储特征点形成 DDLJC, 以后的字形的编辑和输出都是由特征点来动态的生成甲骨文字。因此, DDLJC 是对甲骨文字的最大程度的抽象。DDLJC 的应用, 解决了由于使用轮廓字形描述甲骨文字形而存在的字形动态编辑和字形变换的困难。同时, 字形描述的动态性, 也为由于字形不规范而导致的字形认同问题找到了一个解决方案。

2 甲骨文字形动态描述库

与汉字不同的是, 甲骨文字是契刻文字, 不像现在的汉字有完整的笔画结构(如图 5 所示甲骨文合集第 3028 片)。甲骨文字形结构的特殊性给计算机输入和编辑带来了很大的困难。因此参照现代汉字的书写方法, 引入笔段和笔元的概念。

2.1 甲骨文的有向笔段和笔元

2.1.1 有向笔段和笔元的定义

有向笔段是有方向的线段, 设 (X_i, Y_i) 是起始点, (X_j, Y_j) 是结束点, 则这个有向笔段可以用一维向量 S_{ij} 描述, $S_{ij} = (X_i, Y_i, X_j, Y_j)$ 。有向笔段的起始点叫势点, 结束点叫驻点。将笔段定义为有方向性的意义在于字形生成过程中的起笔、运笔和落笔识别。

笔元是由一个或者多个有向笔段组成的甲骨文中一个完整的笔画结构, 设一个笔元由 n 个笔段组



图 2 现代汉字字形的轮廓来描述
Fig. 2 Outline of modern Chinese character glyph description



图 3 甲骨文字形的轮廓来描述
Fig. 3 Outline of Jiaguwen glyph description



图 4 甲骨文字形的抽象
Fig. 4 Jiaguwen glyph of abstract



图 5 甲骨文合集第 3028 片拓片
Fig. 5 Collection of 3028 Jiaguwen rubbing

成, 则这个笔元可以用向量 E_n 描述, $E_n = (S_{i1j1}, S_{i2j2}, \dots, S_{ijnj})$ 。其中对于任意的 $K \in \{1, 2, \dots, n\}$, 记 S_{ikjk} 为 S_k , 则上述笔元可简记为 $E_n = (S_1, S_2, \dots, S_n)$ 。 E_n 的第一个笔段 S_1 的势点叫做 E_n 的始点, 最后一个笔段 S_n 的驻点叫做 E_n 的终点。

2.1.2 甲骨文的基本笔元及其笔段

笔元相当于现代汉字的笔画。一个甲骨文字笔元的多少与这一文字的结构有关, 由于笔元具有方向性, 因此同一笔元的描述方法有多种, 例如: 一横“一”的描述可以以左端作为始点描述, 也可以右端为始点描述, 一竖“丨”可以以上端作为始点描述, 也可以下端为始点描述, 同样, “撇”和“捺”也同样如此。由于目前还没有甲骨文书写的统一标准, 因此, 参照汉字的“由左到右, 由上到下”的书写顺序, 使用笔元来表示甲骨文书写过程中的落笔、运笔和起笔的过程。图 6 是甲骨文合集第 3028 片中“天”字对应的甲骨文运笔示意图。

根据对目前 5092 个可释甲骨文的笔元进行统计, 我们将甲骨文的笔元分成 3 类基本的笔元: 第一类是直线笔元, 第二类是折线笔元, 第三类是弧线笔元, 如表 2 所示。

组成直线笔元的笔段比较简单, 通常只有一个有向笔段组成。例如: 横线“一”笔段是由左向右的一个笔元组成, 竖线“丨”笔段是由上向下的一个笔元

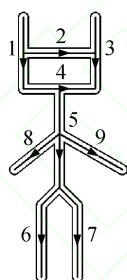


图 6 甲骨文运笔示意图
Fig. 6 Jiaguwen brush sketches

表 2 甲骨文的基本笔元
Table 2 Jiaguwen's basic strokes

笔元类型	基本笔元					
直线	—	丨	丿	㇏	丶	
	横线	竖线	撇线	捺线	点线	
折线	∧	∨	<	>	ㄥ	ㄣ
	上折线	下折线	左折线	右折线	撇折线	捺折线
弧线	(	)	—	—	~	〇
	左弧线	右弧线	上弧线	下弧线	横波浪弧线	竖波浪弧线 椭圆线

组成。对于从左上到右下 (“㇏”) 的笔元、从右上到左下 (“ㄣ”) 的笔元和点线 (“丶”) 也称为直线笔元。

折线笔元通常由两个或多个笔段组成。例如: 上折线、下折线、左折线和右折线有两个笔段, “撇”折线、“捺”折线笔元至少有两个笔段组成。

组成弧线笔元的笔段比较复杂, 通常最少设定 5~6 个笔段。弧线笔元的笔段组成如表 3 所示。

甲骨文字是刻绘文字, 弧线在甲骨文字中很多, 我们使用多笔段描述, 在字形生成过程中可以更方便地根据需要对弧线的弧度进行调整, 使描述字形更加接近原始字形。

一个笔元中, 笔段的数量越多, 数据描述越精细, 将来文字的识别越准确, 但计算的复杂度会越高。例如图 6 中的甲骨文字由 9 个笔元组成, 1, 2, 3, 4, 5, 8 和 9 是 7 个直线笔元, 6 和 7 是两个折线笔元。需要说明的是, 由于甲骨文还没有定形, 因此对甲骨文笔元的定义是动态的, 可以根据需要对已经描述的字形的笔元进行修改, 例如, 图 6 中 8 和 9 两个笔元也可以调整为折线笔元, 这样在后期字形生成时, 这两个笔元的变形效果会更好。

2.1.3 笔元的界定

甲骨文字由表 2 所示的基本笔元组成, 每一个笔元都有起点和终点, 为了使多个笔元之间的起点

表 3 甲骨文的弧线笔元的笔段
Table 3 Jiaguwen's directed strokes of arc strokes

名称	左弧线	右弧线	上弧线	下弧线	横波浪弧线	竖波浪弧线	椭圆线
笔段图							
笔段数量	5	5	5	5	6	6	10

和终点不发生错乱,必须对笔元进行界定。本文使用在每个笔元的始点之前添加界定符号的方式来界定笔元,界定符号叫笔元的界点。若界点为 $D = (D_1, D_2)$, 则笔元 E_n 的描述向量为

$$E_n = (D_1, D_2, S_1, S_2, \dots, S_n)。$$

甲骨文字是其笔元的集合,为了便于机器识别,我们将这个集合表示为笔元的排列。排列顺序可以按笔元的空间位置规定,也可以按笔元的书写顺序规定。由于甲骨文不像现代汉字那样规范,按笔元空间位置排列笔元不便于操作,因此采用书写顺序将各个笔元进行排列。设一个甲骨文字有 m 个笔元 $E_{n_1}, E_{n_2}, \dots, E_{n_m}$, 则这个字描述向量为

$$ZX = (E_{n_1}, E_{n_2}, \dots, E_{n_m})。$$

甲骨文字的描述向量(相当于甲骨文字库中的编码)存放于一个文本文件中,为了界定不同甲骨文字的编码,在每个甲骨文字的编码的第一个笔元之前加上一个界定符号 $H = (H_1, H_2)$, 最后一个笔元之后加上一个界定符号 $H = (T_1, T_2)$ 。上述甲骨文字 ZX 的描述向量为

$$ZX = (H_1, H_2, H_{n_1}, E_{n_2}, \dots, E_{n_m}, T_1, T_2)。$$

2.2 字形描述算法

动态描述算法的主要作用是将绘制或调整后的笔元信息进行规范和存储,算法如下。

步骤 1: 建立或打开字形描述库,并初始化变量。主要是对界点 D 、始点 H 、终点 T 、笔元数 $ele num$ 和字形描述库 DDLJC 进行初始化操作。

步骤 2: 选择操作类型。若操作类型为“Ins”,则插入笔元;若操作类型为“Mov”,则整体移动笔元;若操作类型为“Del”,则删除笔元;若操作类型为“Edit”,则编辑笔元;若操作类型为“NoOper”,则转向步骤 3。

步骤 3: 保存修改,关闭字形描述库。

插入笔元是通过逐个插入该笔元的笔段来实

现,整体移动笔元是通过修改笔元中各个点的坐标(界点除外)来实现,删除笔元是通过删除笔元的各个点来实现。编辑笔元若是指移动笔元中的某个笔段,主要通过移动相应笔段的势点或驻点来实现。综上所述,DDLJC 建立过程示意图如图 7 所示。

3 甲骨文字形的生成

甲骨文字形动态描述算法,为目前还没有完全定形的甲骨文提供了一种动态的描述方法,使用这一方法,我们开发了人机交互的甲骨文字形描述平台,建立了目前可释文字(5096 个,包括异体字)的 DDLJC,如图 8 所示。

在 DDLJC 中,用户不但可以根据需要任意增加或删除甲骨文字,而且还可以根据需要对笔元进行动态编辑。由于甲骨文字形描述库是对甲骨文字形最大程度的抽象,因此,在生成字形方面有较大的字形变换空间。使用 DDLJC 生成甲骨文字形,首先要对 DDLJC 中每个文字和笔元进行基本的坐标变换,然后再根据笔元的特性进行变换。

3.1 基本坐标变换

为了保证字形变换的方便性,在 DDLJC 中使用了和计算机屏幕类似的屏幕坐标系。DDLJC 中各坐标点之所以是动态的,是因为同一文字的不同字形大小描述数据不是唯一的(图 9)。把原点坐标 $O(0,0)$ 平移到新坐标 $O(0,0)$,DDLJC 中的坐标点可以通过坐标变换实现库坐标到屏幕从标的转换。

3.2 直线笔元的变换

直线是最基本的笔元,它由一个始点 $S(x, y)$ 和一个终点 $E(x, y)$ 组成,只要改变这两个点的属性和变换方向就可得到丰富的笔元效果。例如:在现代汉字笔画字形变换中,对横线末端的笔画变形效果如图 10 所示。同样的变形效果,在 DDLJC 中就可以使用直接笔元变换来完成,如图 11 所示。和轮廓

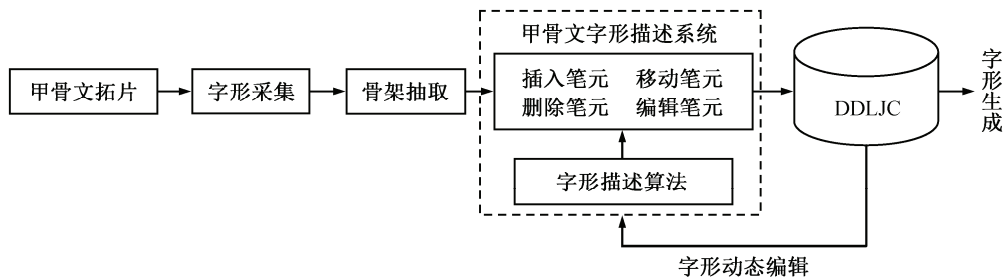
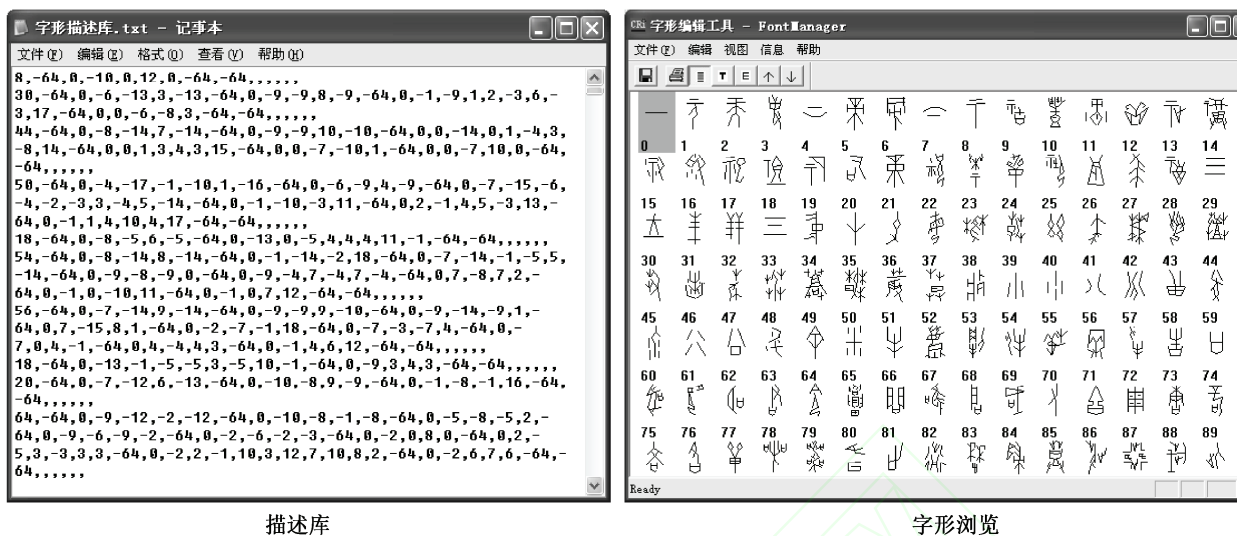


图 7 DDLJC 的创建

Fig 7 Creation process of DDLJC



描述库

字形浏览

图 8 甲骨文字形描述库

Fig. 8 Description library of Jiaguwen

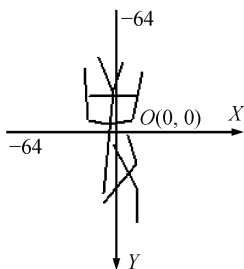


图 9 DDLJC 文字坐标示意图

Fig. 9 DDLJC text coordinates sketches

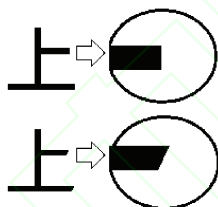


图 10 汉字笔画变形

Fig. 10 Characters stroke deformation

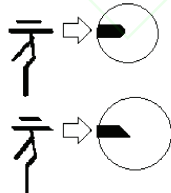


图 11 DDLJC 笔元变换生成字形

Fig. 11 DDLJC strokes transform glyph

线描述字形相比,利用笔元来描述文字,还可以实现更多的笔元变换,如图 12 所示。表 4 是通过直线笔元的不同变换动态生成的字形效果。

3.3 折线笔元的变换

从表面上看,折线笔元是由两个直线笔元组成,

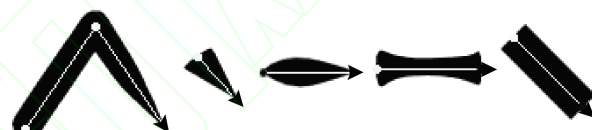


图 12 直线笔元的几种变换

Fig. 12 Several transformation of straight line strokes

表 4 直线笔元的变换

Table 4 Straight strokes transformation

变换类型	生成字形示例
对点的粗细进行变换	s e
对s和e进行方向和位移变换	s e

但由于折线笔元比直线笔元多一个驻点,因此可以通过变换产生更平滑的变化曲线。

给定空间 $n+1$ 个点的位置矢量 $P_i (i=0, 1, 2, \dots, n)$, 则 Bezier 参数曲线上各点坐标的插值公式为

$$P(t) = \sum_{i=0}^n P_i B_{i,n}(t), t \in [0, 1].$$

根据上式,可在折线笔元上插入控制点,产生平滑的变化曲线。表 5 是几种折线笔元的变换效果。

3.4 弧线笔元的变换

甲骨文字形结构与现代汉字字形结构最大的区别就是甲骨文中的笔元弧线很多,为了确保字形生成过程中的最大灵活性,我们在 DDLJC 中定义了弧线笔元。使用弧线笔元,即使不做 Bezier 插值计

表 5 折线字元的不同变换
Table 5 Fold Line of different transformation

变换类型	变换前	变换后	生成字形示例
对上折线进行 Bezier 变换			
对右折线进行 Bezier 变换			
对撇折线进行 Bezier 变换			

算,输出的字形也非常优美。表 6 是使用 DDLJC 库直接输出的字形和使用 Bezier 插值计算后生成的字形的比较。

不同字元的变换可以产生不同的字形效果。在实际应用中,还可以使用笔元变换的不同组合来生成不同的甲骨文字形,如图 13 所示是使用笔元变换的不同组合变换,由 DDLJC 自动生成的细刀体、中刀体、粗刀体、刀变体、细圆体、中圆体、粗圆体和圆变体等 8 种风格的甲骨文字形。

表 6 直接输出字形和使用 Bezier 插值计算后生成字
Table 6 Output glyphs using DDLJC and the generated glyph with Bezier interpolation calculation

生成方式和类型	生成字形示例
用DDLJC直接输出的字形	
用Bezier插值计算后生成的字形	



图 13 使用 DDLJC 开发的 8 种甲骨文字形

Fig. 13 Eight Jiaguwen glyphs using DDLJC development

4 结论

实验证明,DDLJC 不仅可以实现字形的动态编辑,而且在甲骨文字形的生成方面具有速度快、效率高等优点。以甲骨文字形开发为例,三年前,我们开发一套字数为 5096 个甲骨文字的字库,从字形采集到完成基本字模的设计,每人每天只能完成约 80 个,加上使用字库开发工具进行编辑和编码,完

成一套可释甲骨文的字库需要近半年的时间,并且由于是人工设计,字形的规范性较差,每个错误字形的更正都要从字模的设计开始。现在由于使用了 DDLJC,字模的生成变成了指令操作,设定好字形变换指令后,5096 个甲骨文字模在一小时内便可生成,还可以动态地编辑修改错误字形,大大提高了甲骨文字形的开发效率。如果对本算法进一步改进,DDLJC 同样可应用于现代汉字字形的生成。

作为未来的目标,我们将从甲骨文字形的生成质量和扩大 DDLJC 的应用范围两个方面去研究。例如,可以对字形生成算法做进一步的改进,对折线或弧线笔元进行更高阶的 B 样条曲线变换来提高字形的生成质量。可以对金文、小篆等古文字的字形描述进行类似的实验,进一步扩大 DDLJC 的应用范围,为我国古文字数字化研究和应用找到更合适的解决方案。

参考文献

- [1] 林民,宋柔.一种笔段网格汉字字形描述方法.计算机研究与发展,2010,47(2): 318-327
- [2] Li F, Woo P Y. The coding principle and method for automatic recognition of Jiaguwen characters. International Journal of Human-Computer Studies, 2000, 53(2): 289-299
- [3] 沈娟,马小虎.甲骨文的曲线轮廓字形自动生成系统.计算机应用与软件,2009,26(1): 67-68, 114
- [4] 江铭虎,邓北星,廖盼盼,等.甲骨文字库与智能知识库的建立.计算机工程与应用,2004,40(4): 45-47
- [5] 胡金柱,肖明.关于甲骨文象形码输入法的编码原理研究.计算机科学,2002,29(8): 109-111
- [6] 刘永革,栗青生.可视化甲骨文输入法的编码与实现.计算机工程与应用,2004,40(17): 139-140
- [7] 顾绍通,马小虎,杨亦鸣.基于字形拓扑结构的甲骨文输入编码研究.中文信息学报,2008,22(4): 123-128
- [8] 沈建华,曹锦炎.新编甲骨文字形总表.香港:香港中文大学出版社,2001
- [9] 刘志基.读《新编甲骨文字形总表》兼论甲骨文字形检索系统的完善.辞书研究,2006,28(2): 85-96
- [10] 陈婷珠,李新城.《新编甲骨文字形总表》中的异部误增.汉字文化,2010,22(2): 64-68
- [11] 栗青生,杨玉星.甲骨文识别的图同构方法.计算机工程与应用,2011,47(11): 112-114
- [12] 于省吾.甲骨文字诂林.北京:中华书局,1999