

# 基于有向笔段甲骨文输入方法的设计与实现

吴琴霞<sup>1,2\*</sup>, 栗青生<sup>1,2</sup>

(1. 安阳师范学院 计算机与信息工程学院, 河南 安阳 455002; 2. 甲骨文数字化工程研究中心, 河南 安阳 455002)

(\* 通信作者电子邮箱 wqx0218@163.com)

**摘要:**提出一种利用有向笔段和笔元相结合的方法来描述甲骨文字,以期解决甲骨文字输入难、定量难、定形难的问题。该方法首先对甲骨文字笔元进行统计分析归类,针对每种笔元确定其有向笔段的组成,利用有向笔段来描述甲骨文字的算法体系;然后构建出基于有向笔段的甲骨文字输入平台,通过设计人机交互指令方便用户进行甲骨文字特别是未释字和异体字的输入。实验证明使用该方法进行甲骨文字输入输出,可以使字体更加规范,甲骨文字库可以自由增减,字形可以随意修改等。

**关键词:**有向笔段;笔元;甲骨文;字库

**中图分类号:** TP391.1 **文献标志码:** A

## Design and implementation of oracle-bone-script input system based on directed stroke

WU Qin-xia<sup>1,2\*</sup>, LI Qing-sheng<sup>1,2</sup>

(1. School of Computer and Information Engineering, Anyang Normal University, Anyang Henan 455002, China;

2. Institute of Digital Inscriptions on Oracle, Anyang Henan 455002, China)

**Abstract:** This paper proposed a new method to describe oracle-bone-script based on the combination of directed stroke and strokes, which aimed at solving the difficulties in input, quantification and jell of oracle-bone-script. Firstly, the oracle-bone-script strokes were classified and analyzed, then the composition of the directed stroke was determined according to the stroke, and an algorithm of the system to describe oracle-bone-script by directed stroke was formed. The input platform for oracle-bone-script was constructed. Users could input oracle-bone-script especially in variant forms or didnot identify oracle-bone-script through designing human-computer instruction. The experimental results prove that this input method can standardize the font, increase and decrease oracle-bone-script library, and freely modify glyphs.

**Key words:** directed stroke; stroke; oracle-bone-script; word library

## 0 引言

如同认读汉字,认读甲骨文就是从图形到抽象的字符的理解过程。书写甲骨文就是从抽象的字符到形象的图形的生成过程。甲骨文数字化首先要解决的是甲骨文的输入与输出,即建立甲骨文字库,然后对甲骨文字形进行编码,把字形从字库中调出来。在已有的甲骨文输入方法中主要是通过输入一串键盘字符(即甲骨文字对应的外码),它按照一定的规则将每一个甲骨文字形和一个符号串(或者加上数字串)对应起来,从而把甲骨文字输入到计算机中<sup>[1]</sup>。随着计算机信息技术的发展,研究学者对标准甲骨文字形产生巨大的需求,要求计算机能像现代文字一样处理甲骨文字。多年来,在甲骨文数字化过程中,一直存在着甲骨文难以输入的问题。

可以看到,在解决甲骨文的计算机输入的过程中,许多学者参照现代汉字的计算机输入方案,从形码、音码等多个方面研究出发,提出了各种各样的解决方案,解决了部分甲骨文字在输入方面的困难<sup>[2-3]</sup>。但到目前为止,仍然没有一个完整的方案能够解决全部甲骨文字的输入问题。甲骨文字与现代汉字不同,它是一种契刻文字,一字多形的异体字和未释字大量出现。这样甲骨文字库中存放的甲骨文字的个数是不确定的,从字形库中调出的甲骨文字并不一定是用户所需要的。

另外由于甲骨文字没有统一的 Unicode 编码,这给甲骨文字的输入带来了很大的麻烦,针对这一特点,本文提出了一种基于有向笔段的甲骨文输入方法,在该平台上用户可以根据需要自由地修改甲骨文字形,修改后的字体以坐标点的形式保存下来,这样甲骨文字库可以动态地增减。使用该平台可以解决甲骨文字的自由输入,特别是异体字和未释字的输入。

## 1 甲骨文字形描述

### 1.1 概念、规则和定义

书写甲骨文字与汉字不同,甲骨文字是契刻文字,它没有完整的笔画结构,给书写和识别带来了很大的困难。参照现代汉字的书写方法,引入有向笔段和笔元的概念<sup>[4]</sup>。

#### 1.1.1 有向笔段

有向笔段即有方向的线段,设 $(X_1, Y_1)$ 是起点, $(X_2, Y_2)$ 是终点,则一个完整的有向笔段的描述为

$$B_{12} = \{(X_1, Y_1) | (X_2, Y_2)\} \quad (1)$$

有向笔段的起点也叫始点(或势点),有向笔段的终点也叫驻点。

#### 1.1.2 笔元

笔元是由一个或多个有向笔段组成的一个完整的笔画结构,设一个笔元由 $n$ 个有向笔段来组成,则笔元的描述为

收稿日期:2012-01-16;修回日期:2012-03-07。

基金项目:国家自然科学基金资助项目(60973051);河南省科技厅重点科技攻关项目(112102210375)。

作者简介:吴琴霞(1980-),女,河南扶沟人,讲师,硕士,CCF会员,主要研究方向:语义 Web、中文信息处理;栗青生(1966-),男,河南安阳人,教授,博士,主要研究方向:多媒体智能信息处理、智能计算。

$$SS_n = \{BS_1, BS_2, \dots, BS_n\} \quad (2)$$

或者为

$$SS_n = \{(X_{i1}, Y_{i1}) | (X_{j1}, Y_{j1}), (X_{i2}, Y_{i2}) | (X_{j2}, Y_{j2}), \dots, (X_{in}, Y_{in}) | (X_{jn}, Y_{jn})\} \quad (3)$$

笔元的起始界点也称为始界点,笔元的终结界点也称为终界点。

### 1.1.3 甲骨文的基本笔元

笔元相当于现代汉字的笔画。一个甲骨文字笔元的多少与甲骨字的结构有关,由于笔元由多个有向笔段组成,因此同一笔元的描述方法有多种,如:横“一”的描述可以以左边作为起点描述,也可以右边为起点描述。同样,“竖”、“撇”和“捺”也同样如此,参照现代汉字的书写原则,甲骨字笔元的描述按照“由左至右,由上至下,由外到内”的顺序去描述。根据目前已发现的6755个甲骨文字,通过分析其字形,进行分析归类<sup>[5-6]</sup>。根据1.1.2节的笔元定义,可将甲骨文字的笔元分成两类基本笔元:一类是折线笔元,一类是弧线笔元,如图1、2<sup>[7]</sup>所示。

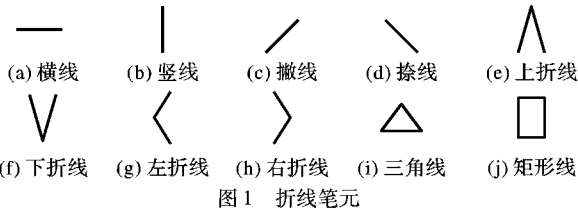


图1 折线笔元

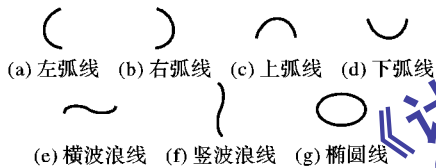


图2 弧线笔元

折线笔元的有向笔段:组成折线笔元的有向笔段比较简单,通常由1到4个有向笔段组成,如横线、竖线、撇线、捺线只有一个有向笔段,上折线、下折线、左折线和右折线有两个有向笔段,三角形线有三个有向笔段,矩形线有四个有向笔段。

弧线笔元的有向笔段与折线笔元相比,组成弧线笔元的有向笔段比较复杂。一个笔元中,有向笔段的数量越多,数据描述越精细,文字的识别越准确,但计算的复杂度会越高。经过多次实验,在该系统中弧线笔元通常设定6个有向笔段,如图3所示各个弧线笔元的有向笔段组成图。

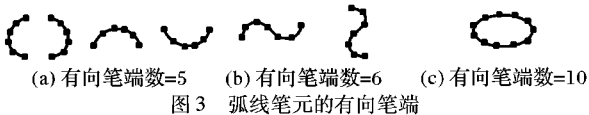


图3 弧线笔元的有向笔端

### 1.1.4 笔元的起始点和终节点的界定

甲骨文字由笔元组成,每一个笔元都有起点和终点,为了不使多个笔元之间的起始点和终节点发生错乱,必须对笔元的起始点和终节点进行界定,为此,本描述方法使用二维特征字符进行分割界定。设 $(S_i, S_i)$ 为笔元的起始界点, $(E_i, E_i)$ 为笔元的终结界点。这样,每个笔元就可以描述为

$$SS_n = \{(S_i, S_i), BS_1, BS_2, BS_3, \dots, BS_n, (E_i, E_i)\} \quad (4)$$

或者为

$$SS_n = \{(S_i, S_i), (X_{i1}, Y_{i1}) | (X_{j1}, Y_{j1}), (X_{i2}, Y_{i2}) | (X_{j2}, Y_{j2}), \dots, (X_{in}, Y_{in}) | (X_{jn}, Y_{jn}), (E_i, E_i)\} \quad (5)$$

甲骨文字形可以描述为多个笔元的组合,这个组合可以表示为排列(空间)位置上的组合或者书写顺序(时间)上的组合,由于甲骨文不像汉字那样规范,因此,按笔元排列位置上的组合不便于操作,因此本描述方法采用书写顺序上的组合,也就是按照书写的顺序将各个笔元进行排列,设一个甲骨文字有 $n$ 个笔元,则这个字的描述<sup>[7]</sup>可以表示为

$$ZX = \{(S_{i1}, S_{i1}), (X_{i1}, Y_{i1}) | (X_{j1}, Y_{j1}), (X_{i2}, Y_{i2}) | (X_{j2}, Y_{j2}), \dots, (X_{in}, Y_{in}) | (X_{jn}, Y_{jn}), (E_{i1}, E_{i1}), (S_{i2}, S_{i2}), (X_{i1}, Y_{i1}) | (X_{j1}, Y_{j1}), (X_{i2}, Y_{i2}) | (X_{j2}, Y_{j2}), \dots, (X_{in}, Y_{in}) | (X_{jn}, Y_{jn}), (E_{i2}, E_{i2}), (S_{i3}, S_{i3}), (X_{i1}, Y_{i1}) | (X_{j1}, Y_{j1}), (X_{i2}, Y_{i2}) | (X_{j2}, Y_{j2}), \dots, (X_{in}, Y_{in}) | (X_{jn}, Y_{jn}), (E_{i3}, E_{i3}), \dots, \dots, (E_{in}, E_{in})\} \quad (6)$$

在组合后的笔元中,分割界定的二维特征字符只是一个分界符号,与笔元没有直接的关系,另外,一个起点和一个终点的分界符可以合成为一个起点的分界符。因此将其作归一化处理,这样式(6)就可以表示为

$$ZX = \{(S, S), (X_{i1}, Y_{i1}) | (X_{j1}, Y_{j1}), (X_{i2}, Y_{i2}) | (X_{j2}, Y_{j2}), \dots, (X_{in}, Y_{in}) | (X_{jn}, Y_{jn}), (S, S), (X_{i1}, Y_{i1}) | (X_{j1}, Y_{j1}), (X_{i2}, Y_{i2}) | (X_{j2}, Y_{j2}), \dots, (X_{in}, Y_{in}) | (X_{jn}, Y_{jn}), (S, S), (X_{i1}, Y_{i1}) | (X_{j1}, Y_{j1}), (X_{i2}, Y_{i2}) | (X_{j2}, Y_{j2}), \dots, (X_{in}, Y_{in}) | (X_{jn}, Y_{jn}), (S, S), \dots, \dots, (E, E)\} \quad (7)$$

归一化处理之后的始界点和终节点统称界点。

## 2 利用有向笔段和笔元来描述甲骨文字形

由于甲骨文是契刻文字,不像现代汉字一样能拆分成若干个部件,所以进行描述时按以下步骤<sup>[7]</sup>进行。

1)确定笔元。根据1.1.3节定义的笔元,选择对应的笔元,确定笔元数量。

2)绘制笔元。确定笔元的数量后,就要按照“由左到右,由上到下,由外到内”的顺序去绘制。为了保证绘制的精确性,通常可以采用透明临摹的方式来进行。如图4所示。

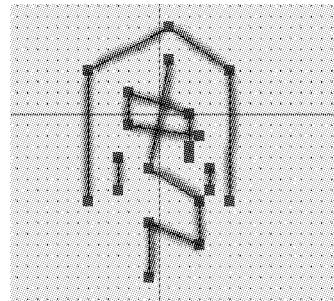


图4 利用透明临摹方式绘制笔元

3)笔元的编辑。经过绘制的文字往往不能满足对笔元进行识别的要求,因此要对笔元进行编辑处理,对笔元的编辑包括对有向笔段的绘制、移动、删除、添加等操作,编辑的最终目的是:既要保证甲骨文字形的正确性,还要保证折线和弧线笔元的规范性。

4)描述字形的存储。经过编辑处理后的甲骨文描述字形,要进行存储操作,才能加入到甲骨文字形描述数据库,以备在编辑系统中进行输入和识别<sup>[8]</sup>。如图2所示,本平台建

立一个区域为  $42 \times 42$  的正方形区域,在正方形区域中建立一个和计算机屏幕方向一致的坐标系  $xoy$ ,组成甲骨字的笔元以一组描述序列保存起来。如果设定  $(-64,0)$  为界点,  $(-64,-64)$  为终点,所有的甲骨文字都以这样的描述序列存储于甲骨文字库中如图6所示,用户动态地修改字形就形成了新的描述序列。

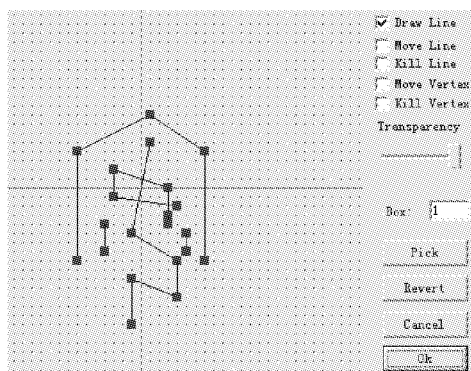


图5 甲骨字笔元自由编辑

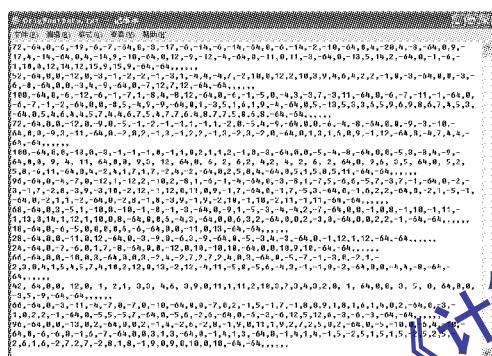


图6 甲骨文字对应的描述字符序列

### 3 甲骨文字形的描述算法

#### 步骤1 变量的初始化。

初始变量一方面包括对界点  $S$ 、始点  $B$ 、驻点  $Z$ 、点数  $P$  和字形库  $ZXDATA(i)$  进行初始化。另一方面还要对两个笔元之间的界定初始化操作。

//变量的初始化操作:

Point  $S$ ;

$P \leftarrow 2$ ;

Point  $B, Z$ ;

Int  $m$ ;

$ZXDATA(i) \leftarrow \{2 + 2i, m, 0, m, m\}$ ;

// 有向笔段之间的界定初始化操作

$BS \leftarrow T$

//  $BS$  是有向笔段,  $\sim BS$  是非有向笔段

$Z \leftarrow T$

//是驻点

#### 步骤2 增加笔元。

增加笔元是通过增加有向笔段来实现的。有向笔段是一个由点  $B$  和点  $Z$  组成的向量。

CASE 1

BEGIN;

If  $BS$

Then

$\{ ZXDATA(i) \leftarrow (2 + 2i, m, 0, B.X, B.Y, Z.X, Z.Y, m, m);$

// 如果  $BZ$  是有向笔段,则在字形描述库中

// 增加二个二元组  $(B.X, B.Y)$  和  $(Z.X, Z.Y)$

$B \text{ lineto } Z;$

// 在  $BZ$  之间连线

$i++;$

GOTO BEGIN;

}

Else

$ZXDATA(i) \leftarrow (2 + 2i, m, 0, B.X, B.Y, m, 0, Z.X, Z.Y, m, m);$   
// 如果  $BZ$  不是有向笔段,则在字形描述库中增加两个  
// 二元组  $(B.X, B.Y)$  和  $(Z.X, Z.Y)$  和一个驻点  $(m, 0,)$

$B \text{ moveto } Z;$

EXIT;

#### 步骤3 移动笔元。

移动笔元是通过修改组成笔元中各个点坐标来实现,设  $vertex[i]$  是移动后的笔元中的一个点,  $pPnt[i]$  是移动前的笔元中的一个点,  $VertexCount$  是点的个数,则:

CASE 2

BEGIN;

double  $dist$ ;

//定义移动距离

double  $maxdist$ ;

//设置最大移动距离

long  $dx, dy$ ;

//移动后点的水平和垂直增量

int  $index$ ;

int  $VertexCount$

$dist = \sqrt{\text{pow}(((double)(vertex[index].X -$

$pPnt[index].x)), 2) +$

$\text{pow}(((double)(vertex[index].Y -$

$pPnt[index].y)), 2))$ ;

if ( $dist < maxdist$ )

{  $dx = vertex[index].X - pPnt[index].x$ ;

$dy = vertex[index].Y - pPnt[index].y$ ;

while (( $index > 0$ ) && ( $vertex[index].X > -M$ ))

$index--$ ;

$index++$ ;

while ( $vertex[index].X > m$ )

{  $vertex[index].X = dx$ ;

$vertex[index].Y = dy$ ;

$index++$ ;

}

EXIT

#### 步骤4 删除笔元。

删除笔元是通过删除有向笔段来实现,也就是减少组成有向笔段中的点的数量。为节省篇幅,只给出主要的算法步骤。

int  $loop$ ;

while (( $index > 0$ ) && ( $vertex[index].X > -m$ ))  $index--$ ;

$VertexCount--$ ;

for ( $loop = index$ ;  $loop < VertexCount$ ;  $loop++$ )

{  $Vertex[loop] = Vertex[loop + 1]$ ;

#### 步骤5 移动势点和驻点。

和步骤3类似,只要用目标点替换移动点即可。

$vertex[index].X = pPnt[index].x$ ;

$vertex[index].Y = pPnt[index].y$ ;

#### 步骤6 保存或取消

### 4 输入平台的实现

根据上述算法思想,设计开发了基于有向笔段的甲骨文字输入平台<sup>[9-10]</sup>。该平台的基本架构如图7所示。

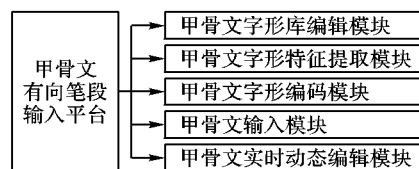


图7 甲骨文有向笔段实验系统基本结构



利用这一平台可以通过人机交互的方式对甲骨文字进行动态的描述,对描述的字形进行编码,动态地扩增字形库,方便了用户对字形的动态修改,解决了未释字的输入<sup>[11-12]</sup>。

实现的基本思路是通过人工方式根据甲骨文的特点输入指令,然后由机器根据本文设计的描述算法提取各个字符的特征,最后形成甲骨文的字形描述数据库。人机交互指令包括有向笔段的操作指令、势点操作指令和环境属性的修改指令。有向笔段的操作指令如表4所示,势点的操作包括移动势点和删除势点对应的指令为MS和CS。环境属性修改指令包括改变线条、临摹选择、取消操作和存储操作。

表1 有向笔端操作指令

操作指令	操作标志
始点(START)	S
画线(DRAW_TO)	D
移动(MOVE_TO)	M
拐点(ZHU)	Z
结束(TERMINATE)	E
删除(DELETE)	D

8个笔段方向和区域见图8和图9所示,8个方向中按照“从左向右,从上到下,从左上到右下,从右上到左下的”为正向,反之则为负向原则进行,正向标志可以省略。

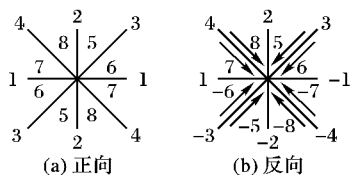


图8 笔段方向图

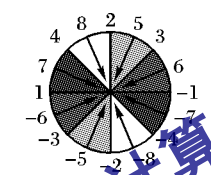


图9 笔段方向区域图

针对每一个笔元设计不同的操作指令,对于折线笔元,使用“近则上线”的指令设计原则来进行,即用1、2、3、4、-1、-2、-3、-4表示8个方向,其中,1表示由左向右的横线,-1表示由右向左的横线,2表示由上到下的竖线,-2表示由下到上的竖线,3表示由右上到左下的撇线,-3表示由左下到右上的撇线,4表示由左上到右下的捺线,-4表示由右下到左上的捺线。每个折线都归到这八个方向中,因此,折线笔元的命令都落在线上。对于弧线笔元则使用“区域逼近”的指令设计原则来进行,即用5、6、7、8、-5、-6、-7、-8等表示8个区域,以和弧线的弦平行的切线为逼近线,在副近线两侧分别用区域值表示方向,例如:弧线中有逼近水平横线1的区域为-6和7,弧线中有逼近水平横线-1的区域为6和-7,……,依此类推。

根据每一个甲骨文字设计人机交互指令,然后通过人机交互指令输入平台将指令输入计算机,根据人机交互指令的输出结果,可以将甲骨文字形动态显示出来,结果如图10所示。

## 5 结语

基于有向笔段的甲骨文输入系统利用有向笔段的描述方法去描述甲骨文字元,再由字元拼接为字形,较好地解决了甲骨文中弧线笔元的描述方法,和只利用笔段技术描述字形相比,不仅字形描述更准确,而且字形更美观,如图11所示,使甲骨文字各部件的拼合更加准确、完善,方便了甲骨文字的识别。

另一方面,由有向笔段描述甲骨文中的字元,使甲骨文字

各部件的拼合更加准确、完善,方便了以后甲骨文字的识别。如图12所示。

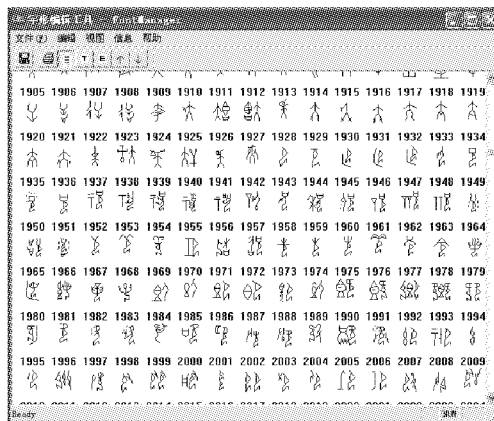
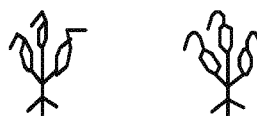
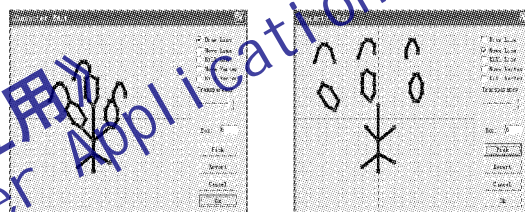


图10 指令输出结果的动态显示



(a) 笔段描述 (b) 有向笔段描述

图11 笔段和有向笔段描述字形对比



(a) 甲骨字整体结构

(b) 甲骨字的字元分解

图12 甲骨文的字元分解

## 参考文献:

- [1] 顾绍通. 甲骨文数字化处理研究进展[J]. 广西民族大学学报: 自然科学版, 2008(1): 80-82.
- [2] 胡金柱, 肖明. 关于甲骨文象形码输入法的编码原理研究[J]. 计算机科学, 2002, 29(8): 109-111.
- [3] 肖明, 赵慧, 甘仲惟. 甲骨文象形码编码方法研究[J]. 中文信息学报, 2003, 17(5): 60-65.
- [4] 林民, 宋柔. 一种笔段网格汉字字形描述方法[J]. 计算机研究与发展, 2010, 47(2): 318-327.
- [5] 沈建华, 曹锦炎. 新编甲骨文字形总表[M]. 香港: 香港中文大学出版社, 2001.
- [6] 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.
- [7] 栗青生, 杨玉星, 王爱民. 甲骨文识别的图同构方法[J]. 计算机工程与应用, 2011, 47(8): 112-114.
- [8] 江铭虎, 邓北星, 廖盼盼, 等. 甲骨文字库与智能知识库的建立[J]. 计算机工程与应用, 2004, 40(4): 45-47.
- [9] 栗青生, 王蕾. 甲骨文图文编辑系统的设计与实现[J]. 安阳师范学院学报, 2011(5): 69-72.
- [10] 聂艳召, 刘永革. 甲骨文自由笔画输入法[J]. 中文信息学报, 2010, 24(6): 103-107.
- [11] 吴琴霞, 刘永革. 基于XML/Schema 甲骨文字语料库语料标注的研究[J]. 科学技术与工程, 2009, 9(17): 5185-5188.
- [12] LI QINSHENG, LIU GUYING. Multi-resolution Markov random field model with variable potentials in wavelet domain for texture image segmentation [C]// 2010 International Conference on Computer Application and System Modeling. Piscataway: IEEE, 2010, 9: 342-346.