

文章编号: 1003-0077(2018)10-0138-05

## 基于分形几何的甲骨文字形识别方法

顾绍通

(江苏师范大学 语言科学与艺术学院, 江苏 徐州 221009)

**摘要:** 甲骨文是流行于我国古代商朝的成熟文字系统,本质上是一种平面图形,笔画和结构不是非常稳定。很多字形具有图画性质,难以区分明显的结构,难写难记。已有的编码输入方法受众面小,效率很低,使用受限。该文分析了甲骨文字形的分形性质,在此基础上,通过字形的重心建立二维平面直角坐标系,将甲骨文字形的平面图形划分为四个象限。利用分形几何的原理,通过计算字形以及各个象限的分形维数,将甲骨文字形形式化为一组分形描述码。再通过与甲骨文字形的分形特征库进行配准,从而识别甲骨文字形。实验结果显示,利用分形几何可以较好地识别甲骨文字形。

**关键词:** 甲骨文;分形几何;分形维数;识别

**中图分类号:** TP391 **文献标识码:** A

## Identification of Oracle-bone Script Fonts Based on Fractal Geometry

GU Shaotong

(School of Linguistic Science and Art, Jiangsu Normal University, Xuzhou, Jiangsu 221009, China)

**Abstract:** Oracle-bone script is an mature writing system used in Shang dynasty, which is engraved on tortoise shells and animal bones. Oracle-bone script is essentially a plane figure, in which the strokes and structures aren't stable, and many characters look like a picture. So it's hard to distinguish obvious structures, hard to write and remember. The existing coding input methods have fewer audiences, low efficiency and limited use. This paper analyzes the fractal property of oracle-bone script according to the theory of fractal geometry. On this basis, the 2D plane rectangular coordinate system is established through the center of gravity of glyph, and the planar graph of oracle-bone glyph is divided into four quadrants. By using fractal geometry principle, the oracle-bone glyph is formed into a component description code by calculating the glyph and fractal dimensions of each quadrant. The oracle-bone script is identified by registration with a fractal feature library of the oracle-bone script. Experimental results show that the scheme of fractal geometry is effective to recognize the oracle-bone script.

**Keywords:** oracle-bone script; fractal geometry; fractal dimension; identification

## 0 引言

甲骨文是书写在龟甲和兽骨上的文字,是我国迄今发现的最早的一种成熟文字系统。

甲骨文字形的输入可以采用编码输入或者识别输入的方法。目前对甲骨文字形采用编码输入的方案已有多种,如基于甲骨文字形动态描述库的输入方法<sup>[1]</sup>、可视化甲骨文输入法<sup>[2]</sup>、基于拓扑结构的输

入方法<sup>[3]</sup>、甲骨文自由笔画输入法<sup>[4]</sup>和象形码输入方法<sup>[5]</sup>。以上方案或多或少需要记住某些规则,这对它的推广使用是不利的。以上方案从字形某一方面的特点出发进行编码,在一定程度上解决了甲骨文字形的输入问题,但也存在不足之处。出土甲骨拓片上的甲骨文字形中,大部分字形无法正确辨识其读音和意义,使得甲骨文编码输入方法存在规则繁琐、重码多和效率低的局限。要让一般用户掌握其复杂的规则并不现实,只有少数从事甲骨文研究

收稿日期: 2017-10-16 定稿日期: 2017-11-23

基金项目: 国家社会科学基金(13CYY039)

方面的专家学者才能掌握复杂的编码规则,这使得编码输入方法的实用性受到限制。随着信息技术的发展,甲骨文的识别输入受到越来越多的重视。目前,已出现多种甲骨文字形识别方案,如顾绍通提出的基于拓扑配准的识别方法<sup>[6]</sup>;周新伦等<sup>[7]</sup>提出利用图论和笔划特点来识别甲骨文字形的方法;李锋等<sup>[8]</sup>提出利用图特征的原理来识别甲骨文字形的方法,并且取得了不错的效果;栗青生等<sup>[9]</sup>提出利用图同构的方法来识别甲骨文字形,这种方法对于那些甲骨文中不同构但仍为同一字形的异写字的识别没有进行处理,且虽然同构但是却不是同一个字形的情况大量存在,这种算法的鲁棒性很低,因而实用性受到限制。

本文将首先分析甲骨文字形的分形性质,通过计算字形的分形维数并与通用甲骨文字库中字形的特征库进行匹配,实现甲骨文字形的识别。文章其余部分的组织结构如下:第一节介绍了分形几何的一般理论;第二节分析了甲骨文字形的分形性质;第三节是基于分形几何的识别算法;第四节是实验结果和分析;第五节对全文进行总结。

## 1 分形几何理论

普通几何学研究对象,一般都具有整数的维数。比如,零维的点、一维的线、二维的面、三维的立体、乃至四维的时空。分形几何研究的是客观事物具有自相似的层次结构,局部与整体在形态、功能、信息、时间、空间等方面具有统计意义上的相似性,成为自相似性。分形是关于自相似性的一般概念,由 Mandelbrot<sup>[10]</sup>提出,用于描述具有相似结构的几何形状。分形理论认为维数也可以是分数,数学家从测度的角度引入了维数概念,将维数从整数扩大到分数,从而突破了一般拓扑集维数为整数的界限。

严格的分形维数的定义如下:如果一个集  $X$  的 Hausdorff 维数  $h(X)$  不是整数,则称集全  $X$  是分形集。直观地说,当  $X \subset R^m$ , 令  $n(\epsilon)$  为覆盖  $X$  所需要的直径为  $\epsilon$  的  $m$  维球的数量,如果当  $\epsilon \rightarrow 0$  时  $n(\epsilon)$  的增加与  $\epsilon$  之间关系满足,如式(1)所示。

$$n(\epsilon) \propto \epsilon^{-D}, \quad \text{当 } \epsilon \rightarrow 0, \quad (1)$$

这时,称  $X$  的 Hausdorff 维数为  $D$ ,例如 Cantor 集的维数为  $D = \frac{\log 2}{\log 3} = 0.6309 \dots$ 。 $h(X)$  的严格定义为:令  $X$  为度量空间的一个子集,设  $d > 0$ ,由下列得到  $d$  维的外测度  $md(X)$ ,如式(2)所示。

$$md(X) = \lim_{\epsilon \rightarrow 0} \inf_{s_i \rightarrow 0} \left( \sum (\text{dim} s_i)^d \right) \quad (2)$$

其中,  $\inf$  是指用直径小于  $\epsilon > 0$  的集合  $s_i$  组成对  $X$  的所有的有限覆盖。 $md(X)$  可以为无,也可以为有限,其值取决于  $d$  的选取。F. Hausdorff 曾经证明存在唯一的非负实数  $d^*(X)$ ,其满足如下性质:若  $0 \leq d \leq d^*(X)$ , 则  $md(X) = \infty$ ,这说明测量的尺度太细小;若  $d^* < d$ , 则  $md(X) = 0$ ,这说明测量的尺度太粗。 $d^*$  称为  $X$  的 Hausdorff 维数,如式(3)所示。

$$h(X) = \sup \{d \in R_+; md(X) = \infty\} \quad (3)$$

Hausdorff 维数的基本思想是,对于任何一个有确定维数的几何体,如果用与它相同维数的“尺”去量度,则可得到一确定的数值  $N$ ;如果用低于它维数的“尺”去量它,结果为无穷大;如果用高于它维数的“尺”去量它,结果为零。其数字表达式为  $N(r) \sim r^{-D_H}$ ,对其两边取自然对数,再进行简单运算后,可得式(4)。

$$D_H = \ln N(r) / \ln(1/r) \quad (4)$$

式中  $D_H$  就称为 Hausdorff 维数。它可以是整数,也可以是分数。

一般来说,如果要严格地计算 Hausdorff 维数是很困难的。自然界存在大量统计意义下的自相似体,通常并不知道其分形维数。为了解决这类自相似体的维数计算,产生了多种计算相似维数的方法,如结构函数法、自仿射法以及盒子覆盖法,这些计算方法性能各异。在实际应用中,盒子覆盖法因计算简单、性能较好、快速准确,应用比较广泛。

设  $F$  是  $R^n$  的非空有界子集,  $N_r(F)$  是覆盖  $F$  的长度至多为  $r$  的集合的个数。 $F$  的上、下盒计数维数分别定义为式(5)~式(7)。

$$\underline{\dim}_B F = \lim_{r \rightarrow 0} \frac{\ln N_r(F)}{-\ln r} \quad (5)$$

$$\overline{\dim}_B F = \lim_{r \rightarrow 0} \frac{\ln N_r(F)}{-\ln r} \quad (6)$$

$$\text{若 } \underline{\dim}_B F = \overline{\dim}_B F$$

则称其公共值  $F$  的盒计数维或盒子维数

$$\dim_B F = \lim_{r \rightarrow 0} \frac{\ln N_r(F)}{-\ln r} \quad (7)$$

由于盒子维数的计算简便,在实际中应用广泛。例如为了得到平面集合  $F$  的盒维数,可以画出每个小格长为  $r$  的正方形或盒网,对各个充分小的  $r$  计数覆盖  $F$  的个数  $N_r(F)$ ,维数是当  $r \rightarrow 0$  时  $N_r(F)$  递增的对数比率,可以用  $\ln N_r(F)$  与  $-\ln r$  图像的斜率来估计其值。

## 2 甲骨文字形的分形性质

分形的基本特点是自相似性。所有的分形都具有一个重要的特征：可以通过一个特征数，即分形维数来测定其不平度、复杂性或卷积度。由于书写材料的质地坚硬、甲骨文创制人员的复杂，使得甲骨文字形的形态变化多样。具体表现在不同的人对一个甲骨文字形有多种不同的刻写方法，不同的契刻方法造就了不同的甲骨文形体，不同字形之间形体差别很大。体现在分形特点上，每个字形的分形特性也不一样。具体表现是每个字形由于刻画形状不一样，分形维数也会存在细微差别。不仅不同字形之间在分形特性上存在差异，就每一个具体字形来讲，同一个字形由于每一部分笔画形状不一样，每一部分字形的笔画在分形特征上也存在差异。具体来说，如果将一个字形的重心为原点，建立平面直角坐标系，则字形在平面上可以划分为如下四个象限：第一象限、第二象限、第三象限和第四象限，如图1和图2所示。

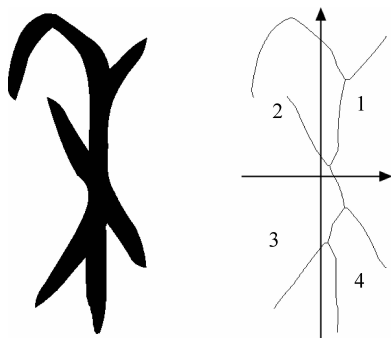


图1 甲骨文字形“𠄎”及细化处理后的四个象限

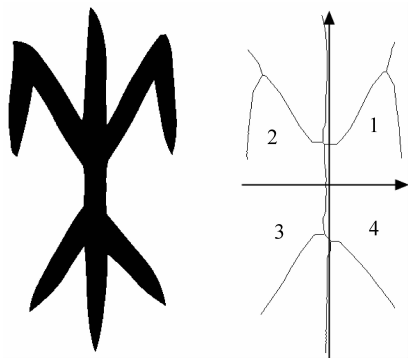


图2 甲骨文字形“𠄎”及细化处理后的四个象限

对于一个字形，四个象限之间的分形维数，也未必相同。由分形维数的定义可知，四个象限的分形维数均值应等于整个字形的分形维数的算术平均

数，即  $d = (d_1 + d_2 + d_3 + d_4) / 4$ ，其中， $d$  表示整个字形的分形维数， $d_1, d_2, d_3, d_4$  分别表示第一、二、三、四象限内的分形维数。如图1中甲骨文字形“𠄎”经细化处理后的分形维数是1.170513，其第一、二、三、四象限内字形部分的分形维数分别是1.260856、1.104295、1.022054、1.295044， $(1.260856 + 1.104295 + 1.022054 + 1.295044) / 4 = 1.170562 \approx 1.170513$ 。图2中甲骨文字形“𠄎”经细化处理后的分形维数是1.148092，其第一、二、三、四象限内字形部分的分形维数分别是1.213270、1.210993、1.052002、1.115948， $(1.6543 + 1.7211 + 1.7365 + 1.7833) / 4 = 1.148053 \approx 1.148092$ 。因此，甲骨文字形用分形维数可以描述如下： $F(C) = (d, d_1, d_2, d_3, d_4)$ 。这样，一个甲骨文字形就可以用一个特征向量来描述。甲骨文中一些字形经细化处理后的分形维数如表1所示。

表1 甲骨文字形分形维数

字	$d$	$d_1$	$d_2$	$d_3$	$d_4$
𠄎	1.177154	1.180801	1.090156	1.278401	1.159390
𠄎	1.170470	1.114956	1.159534	1.252284	1.154974
𠄎	1.106536	1.018731	1.126350	1.157547	1.123678
𠄎	1.130239	1.106671	1.139124	1.179118	1.095924
𠄎	1.192724	1.158605	1.183785	1.217451	1.211241
𠄎	1.121446	1.108521	1.108511	1.142344	1.126540

从以上分析可见，甲骨文字形可由字形的分形维数以及第一、二、三、四象限的分形维数，即  $(d, d_1, d_2, d_3, d_4)$  描述。显而易见，仅仅利用四个象限分形维数的有限组合，如1和4象限  $(d_1, d_4)$  或2和3象限  $(d_2, d_3)$ ，或1和2象限  $(d_1, d_2)$  或3和4象限  $(d_3, d_4)$ ，在识别字形的有效性上并不如  $(d, d_1, d_2, d_3, d_4)$  五元组向量识别甲骨文字形。实验数据如表2所示。

表2 分形维数组合识别有效性对比

	$d, d_1, d_2, d_3, d_4$	$d_1, d_2$	$d_3, d_4$	$d_1, d_4$	$d_2, d_3$
识别率 / %	94	24.3	25.5	23.9	22.6

甲骨文字形每个象限的分形维数在甲骨文字库中的分布如图3所示。

## 3 基于分形几何的识别算法

从以上对甲骨文字形分形特点的描述可以看

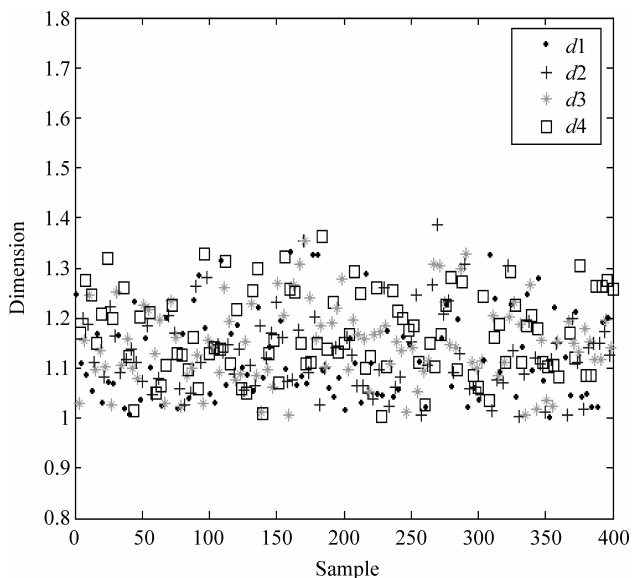


图3 分形维数分布

出,甲骨文字形可以利用其本身的分形维数来描述。分形配准是将不同图形的分形描述特征进行匹配的过程。分形配准可以定义如下:

给定两幅待配准的图形的分形描述如下  $F_1(x, y)$  和  $F_2(x, y)$ , 称其中之一  $F_1(x, y)$  为基准分形, 另一个  $F_2(x, y)$  为待配准分形, 则称分形配准为两分形关系的映射变换, 如式(8)所示。

$$F_1(x, y) = g[F_2(x, y)] \quad (8)$$

在这里,  $g$  为一个二维坐标变换。

分形配准的一般步骤是: 首先对两幅图像进行分形维数计算得到分形描述; 通过进行相似性度量找到匹配的分形描述。

分形特征提取是分形配准的重要环节。准确的分形特征提取为特征匹配的成功进行提供了保障。因此, 寻求具有良好不变性和准确性的特征提取方法, 对于匹配精度至关重要。如果能够精确描述两幅不同字形图像的分形特征, 就可以实现字形在分形关系上的配准。

综上所述, 甲骨文字分形配准算法如下:

**Step1** 对输入字形图像进行细化处理, 得到细化后的字形图像;

**Step2** 通过计算甲骨文字形的分形维数, 得到字形的分形描述;

**Step3** 计算待识字形分形描述码与甲骨文字形分形特征库中分形码( $d$ )的距离;

**Step4** 大于给定阈值的两个分形描述码的字形被识别为分形不等价, 否则被识别为分形等价。如果两个分形码等价, 并且识别结果出现重码, 则执

行 Step5;

**Step5** 计算待识字形分形描述码与甲骨文字形分形特征库中分形码( $d_1, d_4$ )、( $d_2, d_3$ )、( $d_1, d_2$ )、( $d_3, d_4$ )的距离;

**Step6** 大于给定阈值的两个分形描述码的字形被识别为分形不等价, 否则被识别为分形等价。如果两个分形码等价, 且识别结果出现重码, 则执行 Step7;

**Step7** 计算待识字形分形描述码与甲骨文字形分形特征库中分形码( $d_1, d_2, d_3, d_4$ )的距离;

**Step8** 大于给定阈值的两个分形描述码的字形被识别为分形不等价, 否则被识别为分形等价。

甲骨文字形配准识别系统识别甲骨文字形的流程如图4所示。

在判定两个分形描述码是否等价的过程中, 两个字形的分形描述码之间距离阈值的选取对于识别结果有着直接的影响。那么, 阈值如何确定呢? 一般来说, 如果两个甲骨文字形的分形描述码等价, 即属于同一甲骨文字形, 那么这两个分形描述码之间的距离要小于不同甲骨文字的分形描述码的距离。甲骨文中, 同一甲骨文字的异写字形有很多, 这些异写字形之间的分形描述码距离要小于其与另一甲骨文字分形描述码的距离。因此, 确定阈值的一个合理的解决办法是, 对每一个甲骨文字, 计算此甲骨文字异写字形之间分形描述码的距离, 在所有的甲骨文字中, 找出两个异写字形的最大的分形描述码的距离, 此距离作为阈值。

用数学语言描述如下: 令  $T$  表示阈值, 如式(9)所示。

$$T = \max\{\max C_1, \max C_2, \dots, \max C_n\} \quad (9)$$

其中  $\max C_n$  表示甲骨文字  $C_n$  的异写字形之间的分形描述码距离的最大值。

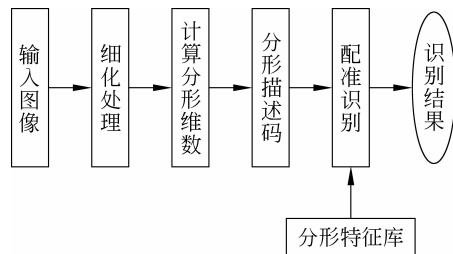


图4 甲骨文字形分形配准识别流程图

## 4 实验结果与分析

我们在 Windows 环境下主频 3.30GHz 的双处理器计算机上, 利用 Visual C++ 2010 和 OpenCV

3.0 实现了以上算法,设计并实现了基于分形几何的甲骨文字形自动识别系统。系统的字库平台是 Windows 环境下自主开发的通用甲骨文字库,字库中的甲骨文字形采用基于二次 Bezier 曲线的轮廓描述技术。系统识别的步骤如下:对输入的图形进行细化处理后,由识别系统计算字形的分形维数,对字形进行分形描述,得到字形的分形描述码。通过计算待配准字形的分形描述码与分形特征库中分形描述码的距离,实现甲骨文字形的配准识别。识别的结果在计算机屏幕上用曲线轮廓将甲骨文字形及对应的汉字显示出来。实验显示,本文提出基于分形几何的甲骨文字形识别算法,既可以识别目前已识读的甲骨文字形,也能识别目前尚无法识读的甲骨文字形,实验数据如表 3 所示。

表 3 实验数据表

识别原理	实验结果	
	总识别率/%	时间花费/s
图论和笔划特点	94	—
图特征	92.27	17
拓扑配准	97	4
分形几何	94	0.83

## 5 结论

甲骨文作为我国最古老的成熟文字系统,在科学研究、文化传承方面具有极其重要的价值。作为最古老的文字系统,甲骨文只为少数专家学者所认识,对于普通大众来说甲骨文字形难写难记,一般用户对传统的甲骨文字形编码输入方法很难掌握,使得编码输入方法的实用性受到很大限制。甲骨文字

形作为一种平面图形,由于书写形体不同,在一定程度上具有分形性质。本文利用分形几何的原理,把甲骨文字形视为分形图形,通过计算字形的分形维数以及平面图形上四个象限内部分的分形维数,利用一组分形描述码将甲骨文字形表示出来,实现甲骨文字形描述的形式化。再通过将甲骨文字形的分形描述码与分形特征库进行配准,从而识别甲骨文字形。利用本文提出的算法设计了甲骨文字形识别系统,实验结果显示,文章提出的算法是有效的。

## 参考文献

- [1] 栗青生,吴琴霞,王蕾.基于甲骨文字形动态描述库的甲骨文输入方法[J].中文信息学报,2012,26(04): 28-33.
- [2] 刘永革,栗青生.可视化甲骨文输入法的编码与实现[J].计算机工程与应用,2004,7: 139-140.
- [3] 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[J].中文信息学报,2008,22(04): 123-128.
- [4] 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(06): 104-107.
- [5] 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(05): 60-65.
- [6] 顾绍通.基于拓扑配准的甲骨文字形识别方法[J].计算机与数字工程,2016,10: 2001-2006.
- [7] 周新伦,李锋,华星城,等.甲骨文计算机识别方法研究[J].复旦学报(自然科学版),1996,05: 481-486.
- [8] 李锋,周新伦.甲骨文自动识别的图论方法[J].电子科学学刊,1996,S1: 41-47.
- [9] 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,2011,08: 112-114.
- [10] Benoît B, Mandelbrot. Fractals: Form, Chance and Dimension[M]. San Francisco, America: W H Freeman, 1977.



顾绍通(1978—),讲师,主要研究领域为工程语言学、甲骨文信息化处理。  
E-mail: gushaotong2000@163.com