

# Oracle-Bone Inscriptions Recognition Based on Topological Features

Hengyi Qu, Jianzhong Liu, Ji Wu

School of Software, Yunnan University, Kunming Yunnan  
Email: hengyiqu@qq.com

Received: Jun. 3<sup>rd</sup>, 2019; accepted: Jun. 17<sup>th</sup>, 2019; published: Jun. 24<sup>th</sup>, 2019

## Abstract

Oracle-bone inscription is an ancient hieroglyph in China. It is an ancient mature text of the Shang Dynasty in China. The study of Oracle-bone inscription is of far-reaching significance for the promotion of Chinese culture. Oracle-bone inscription is a hieroglyph with traces of original pictorial text; Oracle is divided into multiple periods, and the oracle bone shape of each period will change, so it is not possible to use the commonly used Chinese character recognition method to identify Oracle-bone inscription, which caused the stagnation of Oracle-bone inscription recognition research. In this paper, by analyzing the topological features of Oracle-bone inscription, the relationship between topological features such as topological feature points, connected domains and genus is constructed, and a recognition method for high-accuracy recognition of Oracle is proposed. The Oracle-bone inscription on the Oracle dictionary was identified and tested. The experiment showed that the Oracle recognition method accurately and quickly recognized the oracle bone text.

## Keywords

Oracle-Bone Inscription, Topological Feature, Recognition Method

# 基于拓扑特征的甲骨文识别

曲恒熠, 刘建忠, 吴 极

云南大学软件学院, 云南 昆明  
Email: hengyiqu@qq.com

收稿日期: 2019年6月3日; 录用日期: 2019年6月17日; 发布日期: 2019年6月24日

## 摘 要

甲骨文是中国的一种古老的象形文字, 是中国商朝的古老成熟文字, 甲骨文的研究对弘扬中华文化具有

深远的意义。甲骨文的许多字形具有原始图画文字的特点，并且甲骨文分为多个时期，每个时期的甲骨文字形都会有所改变，所以不能使用目前常用的汉字识别方法来识别甲骨文，这造成了甲骨文识别研究的停滞不前。本文通过分析甲骨文的拓扑特征，构建拓扑特征点、连通域、亏格等拓扑特征之间的关系，提出了一种能够高准确识别甲骨文的方法。对甲骨文字典上的甲骨文字进行识别测试，实验表明该甲骨文识别方法能准确并且快速地识别甲骨文字。

## 关键词

甲骨文，拓扑特征，识别方法

Copyright © 2019 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

甲骨文最先发现于河南安阳的殷墟，属于上古汉语。甲骨文因在商周时期刻在龟甲兽骨上而被称为“甲骨文”或“龟甲兽骨文”[1]。甲骨文在中国文字悠久发展历史上具有重要的地位，它是中华民族的智慧结晶，是我们不能丢失的文化瑰宝[2][3]。甲骨文的研究是对中国上古时期的文明研究的一个非常重要的研究方向，而甲骨文的正确识别将会直接影响到后续的研究，所以一个识别速率快且识别准确率高的甲骨文识别系统显得尤为重要[4]。目前针对甲骨文识别的研究十分的稀少，现代文字识别的方法大致可以分为：基于模板匹配的识别方法[5]、基于机器学习的识别方法。对于基于模板匹配的识别方法，由于在三千年前甲骨文需要人为地将甲骨文刻在质地坚硬的材料上(如骨头、青铜器等)，这导致了每一类甲骨文字形的形态变化多样，从而找不到合适的模板进行匹配，所以使用基于模板匹配的识别方法时并不能得到较好的识别结果。对于基于机器学习的识别方法，由于机器学习需要大量的数据集，而目前出土并且可作为甲骨文数据集的数据特别的稀少，从而无法生成识别准确率很高的模型，所以使用基于机器学习的识别方法也不能得到较好的识别结果。

目前国内已经有少部分的学者提出了甲骨文识别的研究方法，如 1996 年 12 月李峰提出的图论方法[6]，以及 1996 年周新伦提出的图特征方法来识别甲骨文。但是这些方法对甲骨文的统一字体不同版本的处理并不是很好，以及受年代技术的限制使得甲骨文的识别结果并不理想。2016 年 10 月提出通过拓扑配准来对甲骨文字形识别[7]，通过对甲骨文字形进行数字编码，并计算甲骨文的拓扑距离来识别甲骨文，取得了很高的识别准确率，但是这种方法的识别速率较慢。刘永革和刘国英于 2017 年 4 月提出的支持向量机(SVM)分类方法来识别甲骨文[8]，但是由于甲骨文字的异形体过于频繁，导致最终的识别准确率并不理想。

本文通过分析甲骨文的已提出的拓扑特征点之间的关系，在此基础上再增加分析甲骨文的连通域、拐角点、亏格等拓扑特征信息，在保证高准确率的同时，提高甲骨文的识别速率。

## 2. 甲骨文图像预处理

二值化：图像二值化是指将图像信息只保留黑白即图像像素中只有 0 和 255 两个值。根据阈值的选取[9]，将彩色图或者灰度图小于阈值的部分转为 0，大于阈值的部分统一转为 255，这里阈值的选取为 80。在进行甲骨文识别之前我们需要将甲骨文原始图像转为二值图，以便于后面的操作。

提取骨架：提取骨架即对甲骨文字体进行细化，得到甲骨文字体的中轴线，甲骨文提取成为骨架可以更清晰地分析其特征点。提取骨架算法当中，有些算法提取效果比较好的细化算法如 A-W 算法[10]，可以在提取骨架过程中避免毛刺和叉点分裂的产生，如图 1 所示，这是对甲骨文字体中的“马”进行提取的骨架图。



Figure 1. Extract feature skeleton for Oracle “horse”  
图 1. 对甲骨文“马”提取骨架

背景滤波：本文采用背景变化规则滤波算法，依据背景属性来对甲骨文图片进行滤波，使用噪声点周边的环境来规范噪声点，从而消除图片上的噪声点，并且与周边环境协调[11]，进而增加甲骨文识别的正确率。

去除毛刺：毛刺的去除是甲骨文识别图像预处理中的重要部分。有些甲骨文图片由于本身的问题，在提取骨架后会出现与原图不协调的类似于“刺”的短线即毛刺。针对毛刺的去除问题，这里有许多方法，文献[12]采用基于方向链码去除毛刺的方法，在方向链码标记跟踪图像的基础上，删除所有小于设定的长度阈值的毛刺；文献[13]是根据生成骨架的不同特点来选定阈值，并删除所有小于设定的长度阈值的毛刺。以上文献都是通过选取合适的阈值，去除小于选定阈值的毛刺，从而提高甲骨文识别的准确率。

骨架修补：有些图像本身存在损坏的问题，如图 2 所示，图 2(a)中线段出现断裂，从而提取骨架后图 2(b)也出现断裂，如果直接进行识别的话会生成不同的特征，这时需要对骨架进行修补。骨架修补时需要根据断裂处的距离以及线段的角度来进行判断是否需要骨架修补。图 2(c)为对图 2(b)的骨架修补图。

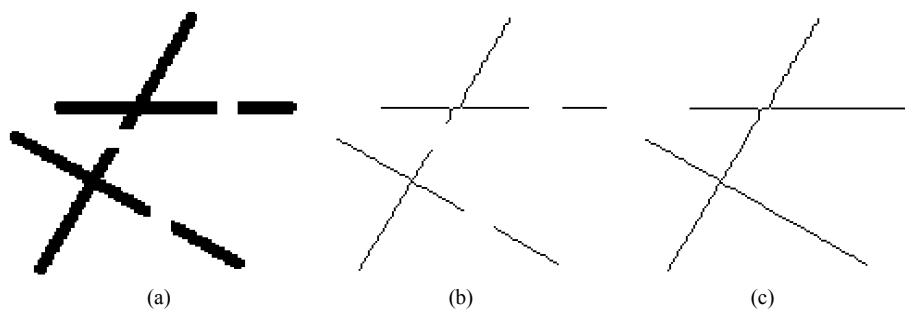


Figure 2. Skeleton repair  
图 2. 骨架修补

### 3. 甲骨文识别

#### 3.1. 甲骨文拓扑特征选取

拓扑学(topology)曾经叫作“形势几何学”[14], 拓扑学可以简单概述为将实体抽象成与其本身无关的点或其他特征, 通过分析点和其他特征之间的关系, 从而达到研究实体的目的。以下为甲骨文识别所要用的拓扑特征:

端点: 端点是线段的起点或终点, 它的连接数为 1, 如图 3 中的(a)所示, 图中有两个端点。

拐角点: 拐角点是指发生在边缘方向发生急剧改变的点[15], 它是特殊的连接点(该点的连接数为 2), 是甲骨文识别当中十分重要的拓扑特征点, 拐角点阈值的选定是根据甲骨文图像的特征, 这里将陡变度阈值设为 0.52 (30°)。

三叉点是指连接数为 3 的点, 如图 3(b)所示, 交点是指连接数为 4 的点, 如图 3(c)所示, 以此类推, 五叉点、六叉点分别为连接数等于 5、6 的点。

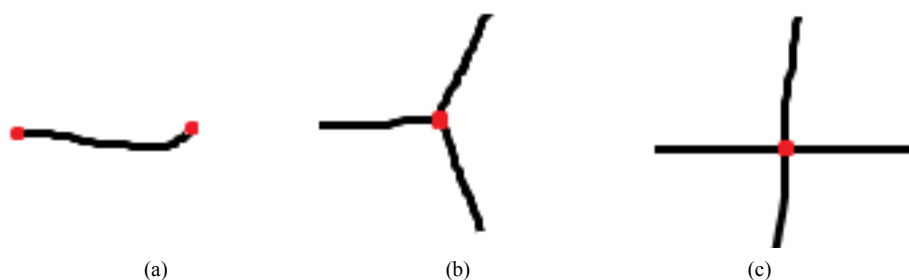


Figure 3. Topological feature point

图 3. 拓扑特征点

亏格: 亏格是甲骨文识别当中十分重要的拓扑特征。亏格是由一条或多条曲线相连而成, 形成的封闭的“孔”。作为甲骨文文字的重要特征之一, 亏格提取可以帮助我们识别甲骨文。如图 4 所示, 甲骨文“天”有一个亏格。



Figure 4. Genus

图 4. 亏格

连通域: 单独的一个点或者单独的一条线都可称为一个连通域。在二值图像中, 连通域是指在一个区域内的任意两个像素之间, 都能找到一条通路能把它们连接起来, 这条通路上的所有像素的灰度值均满足相似的准则。一个甲骨文字可能有一个或者多个连通域。

## 3.2. 甲骨文拓扑特征关系处理

### 3.2.1. 拓扑特征权重选取

由于甲骨文的每个拓扑特征的重要性不同,所以在进行拓扑特征关系处理时需要为每个甲骨文拓扑特征分配不同的权重。对于连接数越多的拓扑特征点,它的重要性也应当随之提高,所以这里  $W(\text{六叉点}) > W(\text{五叉点}) > W(\text{交点}) > W(\text{三叉点}) > W(\text{端点})$ 。而在提取骨架时,某些拓扑特征叉点可能会出现叉点分裂。导致交点以上的点分裂为多个三叉点,通过实验发现,一个交点会分裂为两个三叉点,一个五叉点会分裂为三个三叉点或者一个交点一个三叉点,一个六叉点会分裂为四个三叉点或者一个交点两个三叉点,所以对于六叉点、五叉点、交点、三叉点之间的权重要有相应的联系,权重比为  $W(\text{六叉点}):W(\text{五叉点}):W(\text{交点}):W(\text{三叉点}) = 4:3:2:1$ ,所以在这里三叉点、交点、五叉点、六叉点的权重分别设为 0.2、0.4、0.6、0.8。拐角点是特殊的连接点,它的重要性高于三叉点低于交点,所以这里拐角点的权重应设为 0.3。端点的重要性相较于其他拓扑特征是最低的,所以这里端点的权重设为 0.1。亏格是一些甲骨文特有的拓扑特征,通过亏格可以很容易的识别与之对应的甲骨文,这里亏格权重设为 1。连通域也是很重要的拓扑特征,这里连通域的权重设为 0.7。

### 3.2.2. 拓扑特征库的建立

根据甲骨文字典来建立拓扑特征库,在拓扑特征库中录入甲骨文字典中的每个甲骨文字的拓扑特征信息。同时为了能够提高甲骨文的识别速率,需要对拓扑特征库进行区域划分,从而提高遍历拓扑特征库的速度,这里根据拓扑特征库甲骨文特征信息的不同对拓扑特征库进行区域划分,采用二叉树的结构如图 5 所示,来对拓扑特征库的区域进行划分。

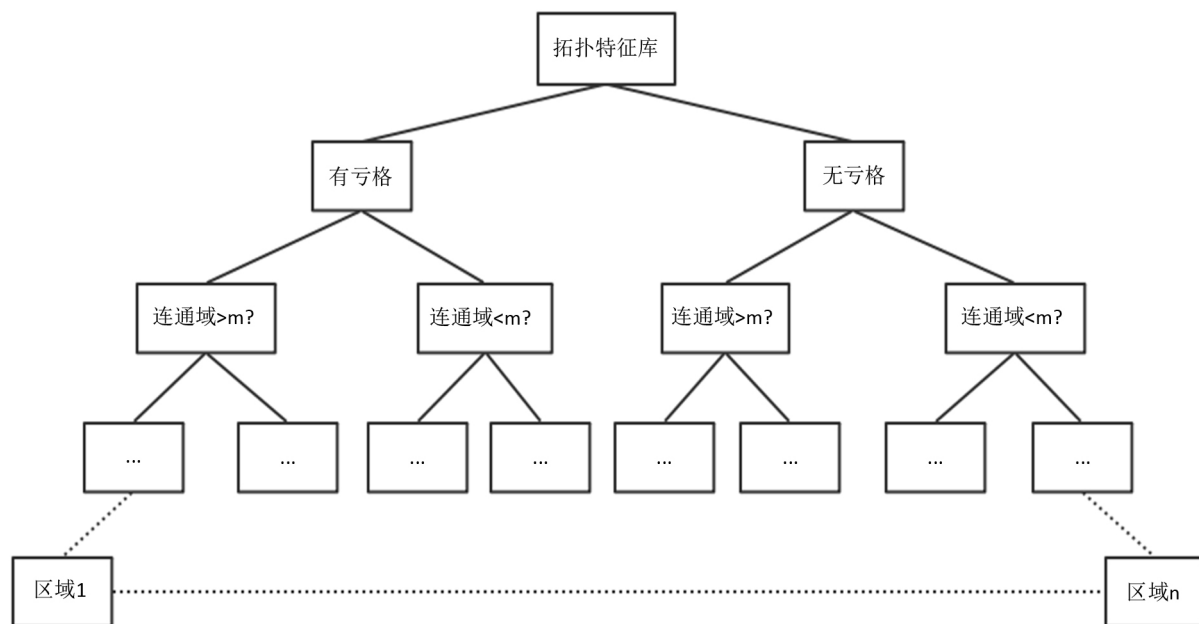


Figure 5. Binary tree

图 5. 二叉树

### 3.2.3. 甲骨文识别

对甲骨文的拓扑特征提取后,得到甲骨文的拓扑特征信息,然后再按照甲骨文的拓扑特征库的区域划分规则,找到该甲骨文所对应的区域,然后再与拓扑特征库中的拓扑特征信息进行相应的权重计算。

权重计算的规则如下所示：

$$R = |T_1 - T'_1| \times W_1 + |T_2 - T'_2| \times W_2 + \dots + |T_n - T'_n| \times W_n$$

这里  $T_1$  到  $T_n$  分别对应拓扑特征库中的每类拓扑特征信息， $T'_1$  到  $T'_n$  分别对应待识别甲骨文的每类拓扑特征信息。 $W_1$  到  $W_n$  分别每类拓扑特征信息的权重， $R$  表示相似度，相似度越低表示该甲骨文与拓扑特征中与之计算的甲骨文越相似。

甲骨文识别系统的识别步骤如下所示：

- 1) 对读取的甲骨文图片二值化操作；
- 2) 二值化进行后在对甲骨文进行提取骨架操作；
- 3) 提取骨架后对目前图像中进行背景变化规则滤波，去除噪声点；
- 4) 滤波后对甲骨文进行去除毛刺操作等操作；
- 5) 对甲骨文提取拓扑特征；
- 6) 将提取的拓扑特征与拓扑特征库中的数据进行对比，进行权重计算；
- 7) 将对比最接近的字体选取为当前甲骨文的识别结果。

4. 实验结果分析

实验在 CPU 设备为 i5-5200,主频为 2.20 GHz,8 G 内存的 Windows 10 环境下利用 Visual Studio 2017 实现了上述中所提出的算法，设计了甲骨文字形自动识别系统，为了提高系统的运行速度，本系统采用了开发语言运行效率最高的 C/C++语言，甲骨文识别系统如图 6 所示。该系统的拓扑特征库的建立是根据《甲骨文字形字典》[16]上的甲骨文信息自主开发的甲骨文拓扑特征库，根据《甲骨文字形字典》收录的甲骨文信息，建立的拓扑特征库中共收录 1200 个甲骨文。并将建立的拓扑特征库上传到 Microsoft SQL Server 2017 数据库中，以便于甲骨文拓扑特征信息的更新和维护。甲骨文识别系统的识别流程为：首先对读取的甲骨文图片进行预处理，包括二值化、提取骨架等一系列操作。然后由对处理后的图像的提取相应的拓扑特征；再将提取的特征与甲骨文拓扑特征库中已建立的甲骨文特征对照，并进行权重计算，选取出最合适的甲骨文输出到图形界面上，并给出识别用时。实验选取了 320 个甲骨文字作为测试集进行识别测试，实验结果表明，根据本文提出的甲骨文识别算法，识别准确率达到 95%，平均识别每个甲骨文所花费的时间为 0.45 s，可以有效的对目前已释读的甲骨文进行识别。



Figure 6. Oracle recognition program  
图 6. 甲骨文识别程序

表 1 为本文提出的识别方法与其他已有的识别方法的比较。

**Table 1.** Comparison of identification methods

**表 1.** 识别方法比较

| 识别方法     | 识别准确率  | 识别用时   |
|----------|--------|--------|
| 图论和笔画特点  | 94%    | 未知     |
| 分形几何[17] | 94%    | 0.83 s |
| 图特征      | 92.27% | 17 s   |
| 拓扑配准     | 97%    | 4 s    |
| 基于 SVM   | 88%    | 未知     |
| 拓扑特征识别   | 95%    | 0.45 s |

## 5. 结论

甲骨文是我国的最古老的一种象形文字，它是中国汉字的起源。虽然甲骨文是我们中华名族的瑰宝，但是甲骨文毕竟是 3300 年前的产物，就算是在它的起源地——河南安阳，实际对甲骨文了解的人并不多，并且甲骨文同一个字有许多不同的版本，除了甲骨文考古领域的专家外，其他人想要学习并深入了解甲骨文并不是一件容易的事情。由于甲骨文的字形具有图画的特征，所以可以将甲骨文字转为图来进行处理，本文通过分析甲骨文的拓扑特征，并改进原先甲骨文识别的方法，在识别甲骨文的基础上提高识别的准确率。实验表明通过分析甲骨文的拓扑特征能够准确并且快速地识别甲骨文字，实现了甲骨文的识别功能。

## 参考文献

- [1] 景舜逸. 壮哉, 殷商甲骨文! [N]. 语言文字报, 2019-04-19(004).
- [2] 李振峰. 深于取象: 甲骨文字的意义生成及其文学意蕴[J]. 北方论丛, 2011(1): 1-5.
- [3] 蔡伟, 黄瑞芳. 借力甲骨文, 助推中学文言文教学[J]. 中学语文教学, 2019(4): 18-22.
- [4] 刘一曼. 考古学与甲骨文的释读[J]. 殷都学刊, 2019(1): 1.
- [5] 高亮, 吴建国, 卢英, 等. 书写文字识别算法研究[J]. 现代计算机(专业版), 2008(7): 23-25.
- [6] 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, 10(12): 4-6.
- [7] 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, 44(10): 4-5.
- [8] 刘永革, 刘国英. 基于 SVM 的甲骨文字识别[J]. 安阳师范学院学报, 2017(2): 54-56.
- [9] 冯炎. 基于对比度归一化的历史文档图像二值化算法[J]. 科学技术与工程, 2019, 19(1): 177-181.
- [10] 张九龙, 王夏妮, 张镇东, 等. 一种书法字骨架提取优化方法[J]. 西安理工大学学报, 2016, 32(1): 35-38.
- [11] 刘建忠, 刘心蓉. 一种基于简单背景变化规则的图像滤波方法[J]. 信息安全, 2014(8): 77-81.
- [12] 王要峰, 崔艳. 基于方向链码去除骨架图像毛刺算法[J]. 计算机应用, 2013, 33(s1): 193-194.
- [13] 叶福玲. 一种改进的图像骨架提取算法[J]. 西昌学院学报, 2018, 32(3): 99-101+131.
- [14] 黄河清. “拓朴学”探源[J]. 中国科技术语, 2019, 21(2): 68-71.
- [15] 刘建忠, 刘鑒, 张云东, 等. 二值图像中基于陡变度的拐角点集检测方法[J]. 信息安全, 2013(3): 12-16.
- [16] 陈济. 甲骨文字形字典[M]. 北京: 长征, 2004: 1.
- [17] 顾绍通. 基于分形几何的甲骨文字形识别方法[J]. 中文信息学报, 2018, 32(10): 2-4.

知网检索的两种方式:

1. 打开知网页面 <http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD>  
下拉列表框选择: [ISSN], 输入期刊 ISSN: 2161-8801, 即可查询
2. 打开知网首页 <http://cnki.net/>  
左侧“国际文献总库”进入, 输入文章标题, 即可查询

投稿请点击: <http://www.hanspub.org/Submission.aspx>

期刊邮箱: [csa@hanspub.org](mailto:csa@hanspub.org)