



基于 SVM 的甲骨文字识别

刘永革, 刘国英

(安阳师范学院 计算机与信息工程学院, 河南 安阳 455000)

[摘要] 甲骨文作为古文字还没有进入国家标准也没有进入国际标准, 所以甲骨文字在出版物上是以图片出现, 这给检索带来了困难; 同时使用大数据进行甲骨文考释的过程中, 需要大量的已标注的甲骨图像数据库, 而人工标注耗时耗力, 且只有甲骨文专家能够完成这项任务, 基于以上两个原因, 甲骨文字图片的识别变得越来越重要, 本文采用支持向量机分类技术研究甲骨文字图片的识别技术, 通过试验证明达到 88% 的准确率。

[关键词] 甲骨文; 支持向量机; 识别

[中图分类号] TP317.1

[文献标识码] A

[文章编号] 1671-5330(2017)02-0054-03

1 甲骨文字识别的研究意义

甲骨文作为古文字还没有进入国家标准也没有进入国际标准, 所以甲骨文字在出版物上是以图片出现, 这给检索带来了困难; 同时使用大数据进行甲骨文考释的过程中, 需要大量的已标注的甲骨图像数据库, 而人工标注耗时耗力, 且只有甲骨文专家能够完成这项任务, 基于以上两个原因, 甲骨文字图片的识别变得越来越重要, 国内外研究文字识别的成果很多, 但是研究甲骨文图像识别的不多, 一是因为甲骨文是古文字, 二是甲骨拓片上文字背景噪声比较大, 三是甲骨文异体字比较多, 所以甲骨文的图像识别有一定难度。

2 支撑向量机进行甲骨文字识别

支撑向量机(Support Vector Machine)是 Cortes 和 Vapnik 于 1995 年首先提出的, 它在解决小样本、非线性及高维模式识别中表现出许多特有的优势, 并能够推广应用到函数拟合等其他机器学习问题中。甲骨文字存在大量的异形体, 且有

很多甲骨字在已出土的甲骨片中只出现几次, 因此甲骨文字的识别需要一个满足小样本的识别方法。因此, 在项目执行过程中, 我们采用支撑向量机进行甲骨文字识别。

3 实验过程

在我们的前期研究中已经建立了甲骨文图文资料库, 该库中包含 6199 个已经经过甲骨文专家标示过的甲骨文字。我们从数据库中, 选择一些异形体出现次数较多或者样本数量较少的甲骨文字构造识别数据库。具体如下:

(1) 从数据库中, 挑选 15 个字符进行识别实验

‘大’, ‘耳’, ‘口’, ‘目’, ‘鸟’, ‘女’, ‘人’, ‘上’, ‘首’, ‘为’, ‘西’, ‘又’, ‘中’, ‘子’, ‘自’。

共计选择了 1290 个甲骨字进行识别, 每一个挑选的样本数量如表 1 所示。

[收稿日期] 2016-11-10

[基金项目] 国家自然科学基金项目(项目编号: 61572037); 河南省教育厅自然科学研究重点项目(项目编号: 14A520023); 河南省甲骨文信息处理重点实验室资助; 汉语海外传播河南省协同中心资助。

[作者简介] 刘永革(1966-), 男, 教授, 主要从事甲骨文信息处理、文字识别与文档分析; 刘国英(1979-), 男, 教授, 博士, 主要从事计算机图形图像研究。



表 1 15 个甲骨字符表

大	耳	口	目	鸟	女	人	上
80	28	18	55	38	70	112	8
首	为	西	又	中	子	自	
21	31	132	257	74	310	56	

训练集:随机从每一个字符对应的字符集合中选择 1/3 的样本作为训练样本;

测试集:整个数据集合作为测试集合。

(2) 甲骨字特征的提取

在研究中,我们采用如下步骤提取甲骨文字特征:

Step1. 原始图像归一化。试验中,采用最小规格化方法:

$$v' = \frac{vmin_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

试验中,我们将 new_min_A 和 new_max_A 分别设置为 0 和 1。

Step2. 对归一化后的图像提取骨架。试验中,采用数学形态学方法对图像进行细化。从而获取只有单像素宽度的甲骨字骨架。

Step3. 图像裁剪。计算骨架图像中甲骨字的最小外接矩形,并据此对图像进行裁剪,获取裁剪后的数字图像。

Step4. 文字特征提取。试验中,我们采用分块直方图的形式提取文字特征。具体来讲,将裁剪后的图像划分为 $M * M$ 个分块,统计每一块内部甲骨字像素的个数,再进行归一化以后作为对应文字的特征。

(3) 设计支撑向量机

首先,在研究中,我们用 C - SVM 作为分类器。对应的目标函数为:

$$\min \frac{1}{2} \|W\|^2 + C \sum_{i=1}^N \varepsilon_i$$

满足: $y_i [wx_i + b] \geq 1 - \varepsilon_i$ 且 $\varepsilon_i > 0$ 。其中, C 为调整代价系数。

其次,我们选用径向基函数作为核函数:

$$K(x,y) = \exp(-\gamma \|x - y\|^2)$$

接着,采用网格搜索和交叉验证的方法获取 C - SVM 参数。对应试验中选用的训练样本,选用的参数是: $C = 2, \gamma = 0.5$ 。

最后,多分类 C - SVM 选用一对一的投票策略进行甲骨字识别。

(4) 实验中选用的识别步骤

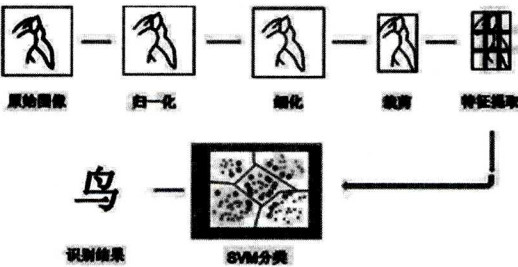


图 1 甲骨字识别步骤

如图 1 所示,打开一幅甲骨图像,依次进行归一化、细化、裁剪和特征提取后,送入 SVM 进行识别,获得最终的识别结果。

(5) 实验系统设计

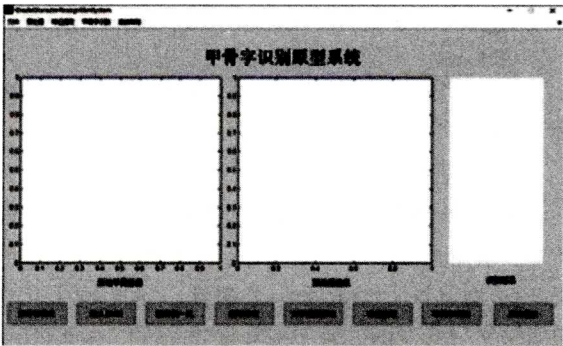


图 2 系统界面

研究中我们开发了如图所示的甲骨字识别原型系统。一个甲骨文字识别的例子如图 3。

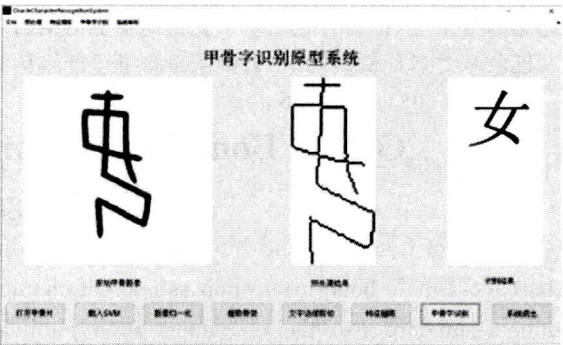


图 3 甲骨字识别示例图

(6) 实验分析

对测试数据进行分析,并获取对应的识别指标。研究中我们采用 Kappa 系数和识别准确率进行评价。针对数据集获取的混淆矩阵如表 2,

表 2 混淆矩阵表

字符	大	耳	口	目	鸟	女	人	上	首	为	西	又	中	子	自	总计
大	77	0	0	0	0	0	0	0	0	0	0	2	0	1	0	80
耳	0	17	0	1	1	1	1	0	0	0	0	0	0	7	0	28
口	0	0	15	0	0	0	1	0	0	0	0	0	0	2	0	18
目	0	0	0	43	0	1	0	0	1	0	0	0	0	10	0	55
鸟	1	0	0	0	22	0	3	0	0	1	2	2	1	6	0	38
女	0	0	0	0	1	53	5	0	1	0	1	1	0	7	1	70
人	0	0	0	1	0	0	102	0	0	0	0	3	0	6	0	112
上	0	0	0	0	0	0	0	6	0	0	0	0	0	2	0	8
首	0	0	0	1	1	0	0	0	12	0	2	0	0	5	0	21
为	0	0	0	0	0	0	2	0	0	24	0	2	0	3	0	31
西	0	0	0	0	0	1	0	0	0	0	116	3	2	10	0	132
又	0	0	0	0	0	0	1	0	0	1	0	252	0	3	0	257
中	0	0	0	0	0	0	1	0	0	0	1	3	65	4	0	74
子	0	0	0	0	1	1	0	0	0	0	10	2	0	289	7	310
自	1	0	0	0	0	0	0	0	0	0	1	0	0	11	43	56
总计	79	17	15	46	26	57	116	6	14	26	133	270	68	366	51	1136

根据该表计算得出 $\text{kappa} = 0.86$, 分类准确率 $\text{acc} = 0.8806$ 。

4 实验结果分析

从实验结果可以看出, 研究中采用的方法虽然有一定的准确率, 但是仍然不够高, 识别出的结果仍需要甲骨文专家进一步确认。这主要是因为甲骨文字异形体出现过于频繁造成的。在后续的研究中, 我们将针对异形体的识别问题重点攻关。

[参考文献]

[1] 王海燕, 王红军, 徐小力. 基于支持向量机的纳西东巴象形文字识别[J]. 云南大学学报(自然科学版), 2016(05): 730 - 736.

[2] 马然. 基于深度学习的自然场景文本识别系统的设计与实现[D]. 长春: 吉林大学, 2015.
[3] 焦微微. 脱机手写文字识别技术方法的研究[D]. 乌鲁木齐: 新疆大学, 2014.
[4] 张鹏, 谢晓尧. 基于改进的 C - 支持向量机的手写体数字高识别率方法研究[J]. 贵州师范大学学报(自然科学版), 2014(02): 95 - 98.
[5] 肖明, 曾莉. 基于 SVM 汉字识别方法的特征分析[J]. 数字技术与应用, 2011(10): 154 - 155.
[6] 李雷. 基于人工智能机器学习的文字识别方法研究[D]. 成都: 电子科技大学, 2013.
[7] 孙华, 李爱平. 支持向量机的古汉字识别研究[J]. 电脑知识与技术, 2013(18): 4296 - 4298.

Oracle Bone Inscription Recognition based on SVM

LIU Yongge, LIU Guoying

(School of Computer and Information Engineering, AnYang Normal University, AnYang 455000, China)

Abstract: Oracle bone inscription as ancient characters has not yet entered the national standard and the international standard, so the characters in a publication are pictures, which bring difficulty to the retrieval process, at the same time; the interpretation of Oracle bone inscription needs annotated image which annotation is time - consuming and only Oracle experts can complete it, based on the above two reasons, the image recognition characters become more and more important, this paper we use support vector machine classification technology to do research of Oracle bone inscription image recognition, the result of experiment is 88% accuracy accurate.

Key words: oracle bone inscription; SVM; recognition

[责任编辑: 江雪]