2024 年第十四届 MathorCup 数学应用挑战赛题目

B 题 甲骨文智能识别中原始拓片单字自动分割与识别研究

甲骨文是我国目前已知的最早成熟的文字系统，它是一种刻在龟甲或兽骨上的古老文字。甲骨文具有极其重要的研究价值，不仅对中国文明的起源具有重要意义，也对世界文明的研究有着深远影响。在我国政府的大力推动下，甲骨文研究已经进入一个全新的发展阶段。人工智能和大数据技术被应用于甲骨文全息性研究及数字化工程建设，成为甲骨文信息处理领域的研究热点[1]。

甲骨文拓片图像分割是甲骨文数字化工程的基础问题，其目的是利用数字图像处理和计算机视觉技术，在甲骨文原始拓片图像的复杂背景中提取出特征分明且互不交叠的独立文字区域。它是甲骨文字修复、字形复原与建模、文字识别、拓片缀合等处理的技术基础[2]。然而，甲骨拓片图像分割往往受到点状噪声、人工纹理和固有纹理三类干扰元素的严重影响[3]。且甲骨文图像来源广泛，包括拓片、拍照、扫描、临摹等，不同的图像来源，其干扰元素的影响是不同的。由于缺乏对甲骨文字及其干扰元素的形态先验特征的特殊考量，通用的代表性图像分割方法目前尚不能对甲骨文原始拓片图像中的文字目标和点状噪声、人工纹理、固有纹理进行有效判别，其误分割率较高，在处理甲骨拓片图像时均有一定局限性。如何从干扰众多的复杂背景中准确地分割出独立文字区域，仍然是一个亟待解决的具有挑战性的问题。

图 1 为一张甲骨文原始拓片的图像分割示例，左图为一整张甲骨文原

始拓片，右图即为利用图像分割算法[4]实现的拓片图像上甲骨文的单字分割。甲骨文的同一个字会有很多异体字，这无疑增加了甲骨文识别的难度，图 2 展示了甲骨文中“人”字的不同异体字。

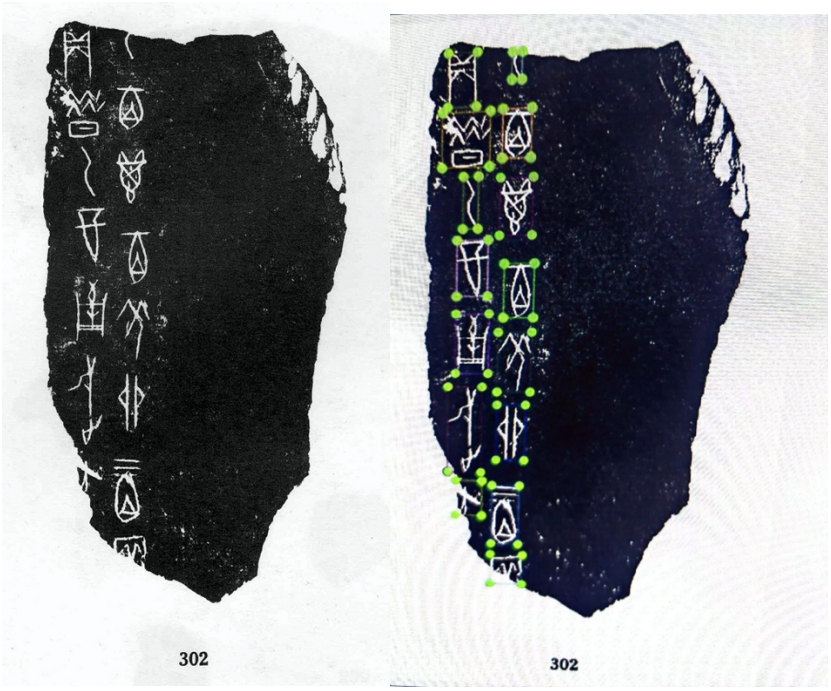

图 1 甲骨文原始拓片和自动识别单字分割情况

图 2 甲骨文中“人”字的不同异体字

现希望通过对已标记的甲骨文图像进行分析、特征提取和建模，从而实现对一张新的甲骨文图像进行单个文字的自动分割和识别。具体任务如下：

问题 1：对于附件 1（Pre_test 文件夹）给定的三张甲骨文原始拓片图片进行图像预处理，提取图像特征，建立甲骨文图像预处理模型，实现对甲骨文图像干扰元素的初步判别和处理。

问题 2：对甲骨文原始拓片图像进行分析，建立一个快速准确的甲骨文图像分割模型，实现对不同的甲骨文原始拓片图像进行自动单字分割，并从不同维度进行模型评估。其中附件 2（Train 文件夹）为已标注分割的数据集。

问题 3：利用建立的甲骨文图像分割模型对附件 3（Test 文件夹）中的 200 张甲骨文原始拓片图像进行自动单字分割，并将分割结果放在“Test_results.xlsx”中，此文件单独上传至竞赛平台。

问题 4：基于前三问对甲骨文原始拓片图像的单字分割研究，请采用合适的方法进行甲骨文原始拓片的文字识别，附件 4（Recognize 文件夹）中给出了部分已标注的甲骨文字形（不限于此训练集，可自行查找其他资料，如使用外部资料需在论文中注明来源），请对测试集中的 50 张甲骨文原始拓片图像进行文字自动识别，并以适当结果呈现。

参考文献

[1] 高旭.基于卷积神经网络的甲骨文识别研究与应用[D].吉林大学, 2021.

[2] Yabing S. Manifold and splendid: 120 Years of research on the oracle bone inscriptions and Shang history[J]. Chinese Studies in History, 2020, 53(4): 351-368.

[3] 宋传鸣,乔明泽,洪陋.边缘梯度协方差引导的甲骨文字修复算法[J].辽宁师范大学学报(自然科学版), 2023, 46(02):194-207.

[4] Zhang C, Zong R, Cao S, et al. AI-powered oracle bone inscriptions recognition and fragments rejoining[C]//Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 2021: 5309-5311.

附件说明：

附件下载链接：https://pan.baidu.com/s/1cp-DUwgS3c_d8EFhBwXRHw?pwd=pf1t

提取码：pf1t

附件 1： Pre_test 文件夹

三张甲骨文图像（.jpg）

附件 2：Train 文件夹

训练数据集，文件夹中包含 6150 张甲骨文图像（.jpg）和对应的标注文件（.json）。标注方式为矩形框，每个框用两个坐标表示（矩形的左上顶点和右下顶点）。如：

{"img_name": "w01906", "ann": [[111.0, 197.0, 152.0, 257.0, 1.0], [108.0, 283.0, 157.0, 335.0, 1.0], [108.0, 356.0, 144.0, 416.0, 1.0], [69.0, 192.0, 102.0, 301.0, 1.0], [106.0, 27.0, 139.0, 79.0, 1.0], [103.0, 77.0, 147.0, 163.0, 1.0]]}

其中“img_name”表示标注对应的图像，“ann”表示图像上的标注。“ann”中前四个数字为对应的坐标；末位为校验位，表示其是否为甲骨字：“1.0”表示为甲骨文，“0.0”表示其不是甲骨文。

附件 3：Test 文件夹

测试数据集，Figures 文件夹中包含 200 张未标注甲骨文图像（.jpg）和 Test_results.xlsx 文件。

附件 4：Recognize 文件夹

甲骨文识别训练集和测试集，训练集文件夹中包含 76 个常见甲骨文及其对应的异体字。测试集中包含 50 张甲骨文原始拓片。