湘西里耶秦简文字数字化检测技术研究与实现

陶珩，张钦科，刘浩男，戴苗鹏，李曙
（吉首大学信息科学与工程学院，湖南吉首，416000）

基金项目：本文受到国家级大学生创新创业训练计划项目（项目号：201910531053）、湖南省2019年大学生创新创业训练计划项目（项目号：S201910531007）和湖南省普通高等学校教学改革研究项目（项目号：HNJG-2020-0564）资助。

摘要：对里耶秦简文字模糊不清，不易识别的问题，提出了一种利用计算机检测的方法。在对里耶秦简图像进行预处理后，利用MSER算法和非极大值抑制算法，能有效地检测出里耶秦简图像上的文字区域，并成功的将其分割和框选出来，达到便于后续研究的目的。

关键词：里耶秦简；文字区域检测；区域特征提取

DOI:10.16589/j.cnki.cn11-3571/tn.2021.18.019

0 引言

2002 年，湖南湘西出土了 3.8 万多枚里耶秦简，数量庞大，是此前全国发现秦简总和的 10 倍，总计 20 余万字，内容涉及当时社会政治、经济、文化的各个层面，极大填补了秦代的历史文献和档案资料，让今人对秦朝政治和社会生活有了更加全面准确的认知。

里耶秦简出土于里耶战国秦代古城遗址，即现在里耶镇。里耶镇行政规划属湘西土家族苗族自治州龙山县，坐落于湖南省武陵山腹地，湘、鄂、渝、黔四省市在此交界，距离保靖城区 61 公里。里耶历史古老，早在 6000 年前便有人居住，但由于交通不便使其经济文化一直较为落后，直至清康熙年间始建街道和码头，雍正年间设置里耶塘，并渐成集市，一度繁荣。

研究里耶秦简可更深入了解秦朝社会的经济政治。但由于秦简的书写多采用竹子或木头，容易虫蛀霉变，且年代久远，在遗址的保存效果不好，发现出土时，秦简上的字已经出现了一系列模糊不清或笔画缺失等问题。

为有效识别秦简文字、避免耗费大量的人力物力地进行人工识别，有效结合现代图像识别处理技术，提高简牍文字的可识别数量和准确率，本文通过对里耶秦简照片进行处理，分割文字得到单文字区域并有效的显示出来，无需复杂的人为操作，效率比较高，实用性比较强。这对于后续里耶秦简文字的自动检测以及对缺少笔画的修复等的研究具有重要意义。

1 里耶秦简及其文字特征分析

里耶秦简出土数量庞大，不乏有清晰、辨识度高的简牍。书写字体绝大多数为古隶书文字，鲜有官方颁布的小篆文字。里耶秦简文字风格各异，大多起笔藏锋逆入，横竖线条粗细对比不明显，字形大多以长形居多，上紧下松，大小不一。里耶秦简书写格式严谨，从上往下错落有致[1-2]。

2 里耶秦简文字检测识别过程

现有文字识别检测的方法很多，如将所得字体骨架与原先所建字体库进行对比，得到字体库中与之近似度最高的文字作为识别结果；或利用专业软件橡皮擦、铅笔等功能对书法文字进行修复。前者方法需要在识别前建立完善的字体库，但里耶秦简文字因没有统一字体库而不适用于该方法，而后者方法需要人为的对书法图像进行删改与填充，需要花费大量的人力与时间，故实用性不佳。

本文提出利用最大极值稳定区域和非极大值抑制的方法，快速、准确率高地将文字区域获取并显示出来。

步骤（1）：文字图像拍摄获取。文字图像拍摄获取是将出土的秦简进行数字化。通过黑白摄像机、扫描仪、数码相机等设备，将秦简拍摄成可以读入计算机的图片，图片以 BGR 格式存放计算机中。该关键点在于拍摄获取图像的清晰度，即图片像素点矩阵，像素矩阵越大，其表现的里耶秦简画面信息量越多，后序处理效果将越好。清晰的书法图像更利于后序的处理，故获取高清晰度高保真的文字图片至关重要[3]。

步骤（2）：图像预处理。在进行文字区域检测之前，为保证检测结果的成功率，减少非文字信息对识别整体效果的干扰，对文字图像进行预处理。图像拍摄获取时多为彩色图像，为减少颜色对识别结果的影响，第一步将图片转换为灰度图。接着对灰度图进行滤波处理，目的是为了消除图像中的噪点，利用算子的平滑过程，降低图片噪声点与信息在图片中所占影响，达到忽略噪声点对整体图像影响的效果。滤波处理后对图像进行二值化，其原理在于通过阈值划分像素值为两个大小范围，将大于阈值的像素点与小于阈值的像素点分别处理为统一大小像素值（本文将大于阈值的像素点值大小设置为 255，小于阈值的设置为 0），得到非黑即白的二值化文字图像。

步骤（3）：形态学操作及均值化处理。进行平滑处理后的图像依然存在颗粒较大的噪点，且部分字的偏旁部首间隙大，易将这类字体检测成多个文字区域；非文字区域检测成文字区域。步骤（3）首先对平滑后的图像进行腐蚀的形态学操作，是再次将影响较大的颗粒进行缩小处理，使其颗粒变细成细小的噪声点。颗粒变成噪声点后，明显的大颗粒

已消失不见，对图像进行均值化处理，即可消除噪声点，极大程度上只保留图像中的文字区域。进行腐蚀操作得到的非实际上文字区域，故接着进行膨胀操作还原原先文字区域。因均值化处理得到的图像结果并非二值图像，为进行后序操作，再次对图像进行二值化处理。在步骤（3）处理后，可使得偏旁部首的字区域连接形成唯一的文字区域。

步骤（4）：最稳定极值区域算法处理。该算法类似于分水岭算法，该算法不断进行多次改变阈值的二值化处理，直到得到稳定的极值区域。该方法将区域寻找出并进行标记，在进行二值化处理的同时，标记区域的大小与像素值会发生改变，不断改变阈值的方法计算不同情况下该区域的稳定值，将最稳定范围内所有出现的可能作为算法最后处理的结果，因此会在计算后返回多个值。其结果为对一个文字区域进行重复的框选。

步骤（5）：非极大值抑制处理。在进行步骤（4）后，得到了需要的文字区域的同时，也出现多区域重叠出现现象，为能够正确显示，对所有标记区域进行非极大值抑制处理，保留最合适稳定极值区域，将多个极值区域处理成一个区域。目的是为了删除重复区域，只保留框选正确的正确区域。

3 实验处理过程与程序实现

本文选取多幅里耶秦简文字图像，分成高信噪比和低信噪比两类分别进行实验处理。

从两类分别选取一幅图作为实验示例。基于目前研究现状，本文代码采用 python 实现并完成图像处理具体步骤。

(1) 程序名称：Main.py。该程序读入图像，调用其他程序，对最终得到的文字区域完成矩形绘制过程，返回最终的处理结果。

(2) 程序名称：Pretreatment.py。该程序将完成图像预处理步骤，因多幅图像在进行二值化时需要的参数不同，该程序采用自适应的处理方法，对不同图像寻找其最适合的阈值。

(3) 程序名称：Morphological_ Average.py。该程序完成形态学操作以及图像的均值化处理。该程序将均值化与形态学运算的卷积核设置为 3×3 的大小，经实验该核适应大多数图像，但为达最佳处理结果，应对不同图像的卷积核大小进行相应的调整。

(4) 程序名称：Mser.py。计算最稳定极值区域，返回处理结果图像。

(5) 程序名称：Nms.py。非极大值抑制，得到最终的文字区域，将结果返回 Main.py 程序。

■ 3.1 高信噪比文字图像处理结果

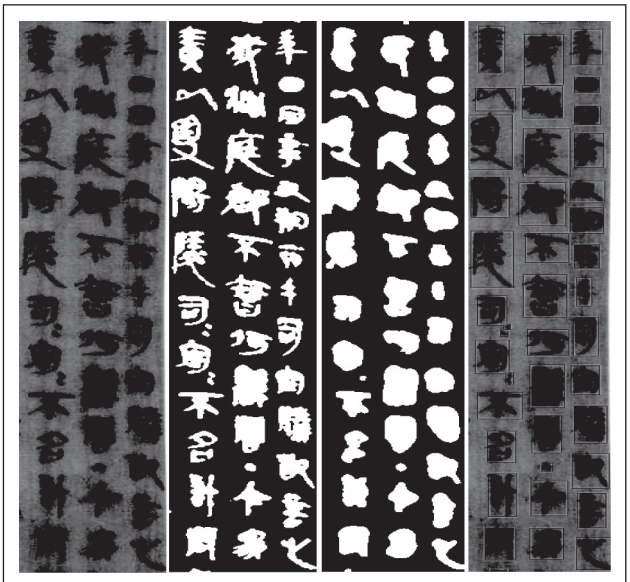

图 1 （a）原始里耶秦简文字图像；(b) 预处理后的图像；(c) 形态学操作与均值化处理后的结果；(d) 检测结果。

在该实验中成功将文字区域显示并在原图标识出来，且检测成功率较高，部分紧挨着的字体分割较好，但文字区域与非文字区域如图像上的墨点有很大的相似性，故不易区分，实验结果中也将墨点框选中。

■ 3.2 低信噪比文字图像处理结果

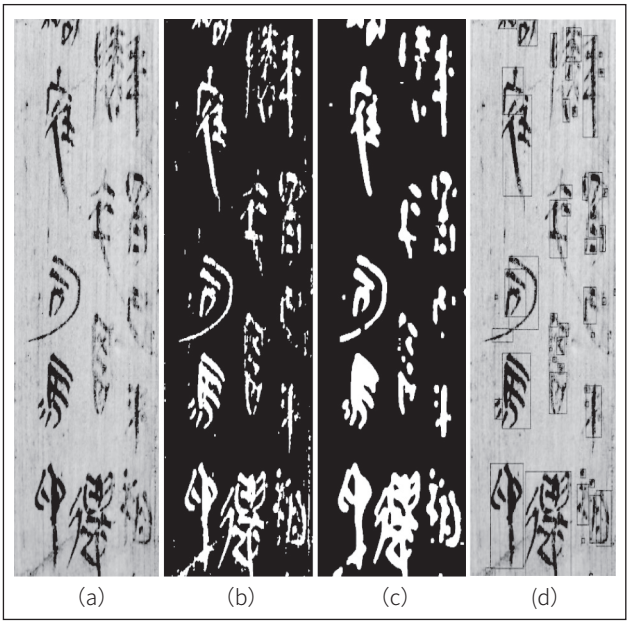

图 2 (a) 低信噪比里耶秦简文字图像；（b）预处理后图像；(c) 形态学操作与均值化处理结果；(d) 检测结果。

低信噪比的图像具有字体清晰度较低，非文字区域噪声点密集，在转化为灰度图像后，非文字区域存在许多与文字区域像素值接近的噪声点，这样的噪声点不易擦除，对实验

（下转第 25 页）

着峰值电压的增大，高压电极与接地电极间的电场强度增大，电子通过电场加速可以获得更多的能量，所以电子激发温度增大。当峰值电压从 8 kV 增大到 11 kV 时，电子激发温度从 4129 K 增大到了 4465 K。

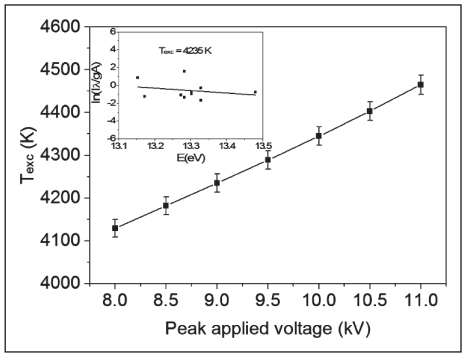

图 7　电子激发温度随峰值电压的变化

3 结论

利用预先产生的 Ar 等离子体射流在在高压电极周围产生了空气 /Ar 介质阻挡放电等离子体，并对其进行了电学特性和光谱特性诊断。实验结果表明，放电电流中出现了大量的丝状电流脉冲，并且最大放电功率为 61.21 W。另外，放电等离子体中存在 Ar*、OH*、N_2*、N_2^+、O* 等多种活性粒子，并且放电功率、N_2 转动温度和振动温度、电子激发温度随峰值电压的增大几乎线性地增大。

参考文献

* [1] Sadat H, Dubus N, Le-Dez V, et al. A simple model for transient temperature rise and fall in a dielectric barrier discharge reactor after ignition and shut down [J]. J. Electrostat., 2010, 68(1): 27-30.
* [2] Kogelschatz U. Dielectric-barrier discharges: their history, discharge physics, and industrial applications [J]. Plasma Chem. Plasma Process., 2003, 23(1): 1-46.
* [3] Massines F, Segur P, Gherardi N, et al. Physics and chemistry in a glow dielectric barrier discharge at atmospheric pressure: diagnostics and modeling [J]. Surf. Coating Technol., 2003,174-175: 8-14.
* [4] Liu S, Neiger M. Excitation of dielectric barrier discharges by unipolar submicrosecond square pulses [J]. J. Phys. D: Appl. Phys., 2001, 34(11): 1632-1638.
* [5] Konelschatz U, Eliasson B, Egli W. Dielectric-barrier discharges. Principle and applications [J]. J. Physique IV, 1997, 7(C4): C4-47-C4-66.
* [6] Massines F, Gherardi N, Fornelli A, et al. Atmospheric pressure plasma deposition of thin films by Townsend dielectric barrier discharge [J]. Surf. Coating Technol., 2005, 200(5-6): 1855-1861.
* [7] Dahiru U H, Saleem F, Zhang K, et al. Removal of cyclohexane as a toxic pollutant from air using a non-thermal plasma: Influence of different parameters [J]. J. Environ. Chem. Eng., 2021, 9: 105023.
* [8] Zhang L, Wang H, Luo H. Uncovering the inactivation kinetics of Escherichia coli in saline by atmospheric DBD plasma using ATR FT‐IR [J]. Plasma Process. Polym., 2020, 17(9): 1900197.
* [9] Uhm H S, Lim J P, Li S Z. Sterilization of bacterial endospores by an atmospheric-pressure argon plasma jet [J]. Appl. Phys. Lett., 2007, 90(26): 261501.

（上接第 57 页）

结果影响较大。对此，可利用灰度值的加权平均值作为阈值的改进二值化进行处理，该方法利用灰度值的加权平均值作为区分文字区域与非文字区域像素值的标尺，减少背景中噪声点的干扰的同时，极大程度上保留文字区域。

4 结束语

本文利用里耶秦简文字区域像素值接近的特点，对图像进行预处理、形态学操作及均值化处理、最稳定极值区域算法处理、非极大值抑制处理后，成功得到所需要检测的文字区域并将其显示。在检测高信噪比的文字图像时成功率较高，对多幅图共计 70 余字进行检测，共检测出个文字，成功率达。本文实验结果表明，本文算法对多数秦简文字能有效的分割与提取，有利于后期对里耶秦简文字识别工作的展开，较传统检测方法而言，本文所述方法大大的提高了检测的效率，在一定程度上为考古工作减轻来不少的工作量。

参考文献

* [1] 董飞 . 里耶秦简“笥牌”读札 [J]. 宝鸡文理学院学报 (社会科学版),2020,40(06):22-27+68.
* [2] 蒋伟男 .《里耶秦简（贰）》疑难文字补释（四则）[J]. 古文字研究 ,2020(00):476-479.
* [3] 张霄军 , 陈小荷 . 古文字自动识别过程及其程序实现 [J]. 中国文字研究 ,2006(00):37-41.