图像处理算法在秦简文字编建立过程中的应用与价值

吴峥

（湖南大学，湖南省长沙市，410006）

摘要：以往出土文献尤其是秦简的文字编建立过程中，人工手动进行图像裁切处理需要花费大量时间和精力，针对这一问题，本文以《岳麓书院藏秦简（四）》第一组简文为例进行实践和阐述，提出一套包含切条和切字的图像处理算法流程用以辅助人工处理，可以大大提升文字编的效率。

关键词：图像处理; 文字编; 秦简;

中图分类号：H　　文献标识码：B

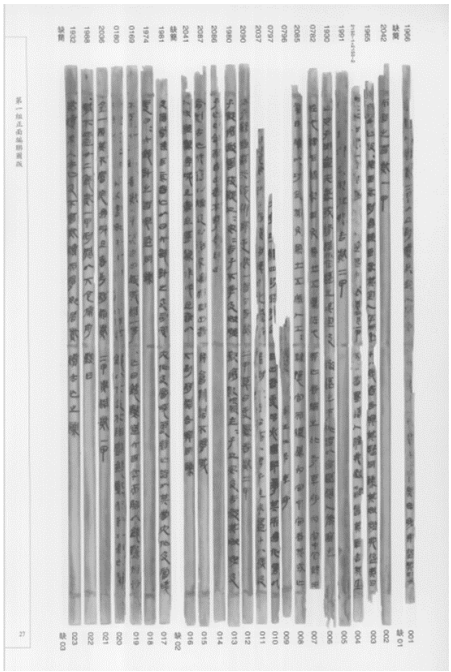
当前文字编整理工作的一般方法为穷尽性考察和收录出土材料中的文字，将红外扫描图像中的文字依序剪切出来，去除背景保留文字本身笔画部分。虽然借助了计算机的方式，但是研究者在建立文字编时仍需对原始图像进行手动裁切处理，此过程往往消耗大量时间和精力，少则需要数月，多则半年以上，而针对秦简图像自身特点的图像处理和识别算法研究目前仍然较少。

本项研究所探究的秦简图像处理算法可以较为迅速的裁切出单文字图像，以该算法为辅助进行图片裁切可以大大提升文字编的图像处理效率。在简牍文字图像处理过程中，“切条”和“切字”是两道重要工序。本文以原始书目 pdf 扫描文档中的红外编联图为待处理原始图像，经过“切条”“切字”和“二值化”等流程为得到单文字图像，即以往文字编整理所做工作，此处以计算机为主，人工校验为辅的方式进行。

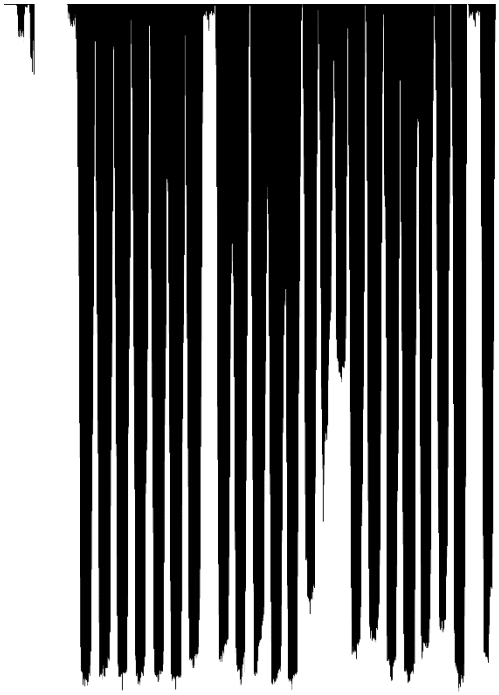
下面以《岳麓书院藏秦简（四）》为例列出从完整 pdf 图像到简条再到单字的处理过程与原理，最终得到经过二值化处理的黑白单文字图像。

一 切条

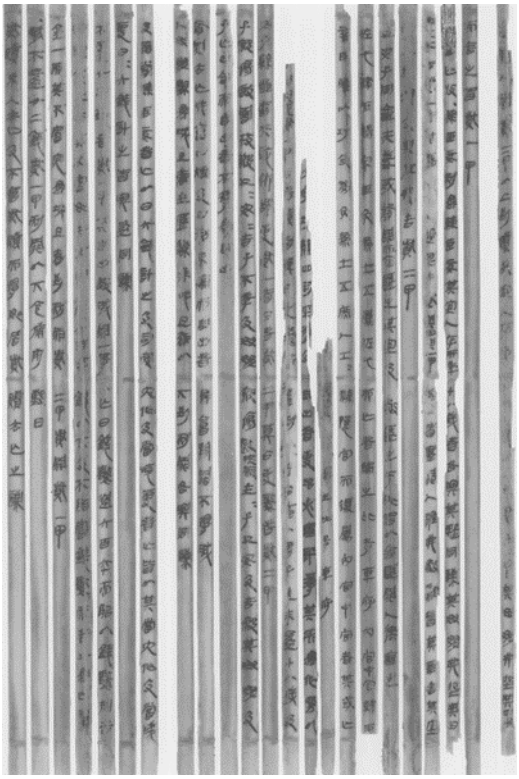
本文所取简文原图像来自《岳麓书院藏秦简四》中的编联图版红外扫描图像，因为已经经过灰度化处理，故不再考虑灰度化的问题。第一支简所在页面的完整图像如下：



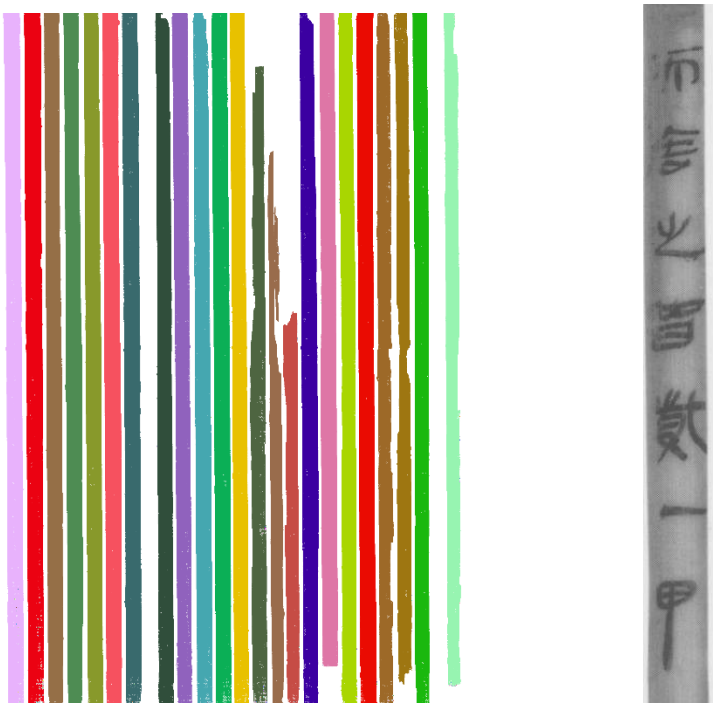
首先，采用基于统计的投影过滤方法将简文图像四周的简号、页码等过滤出去。其基本原理为提取页面每一个点的像素，计算灰度值，分别投影到图像的横轴和纵轴上，根据投影所示直方图敲定图像核心区域，将核心区域之外的图像裁切出去，保留简文高分布中心区域图像。同时，根据右侧图像的分布特点可知，简支和简支之间距离过于紧密且不是完全规则的矩形，故不宜继续采用此种方式进行下一步的裁剪切分。



裁切之后的效果如下：



观察图像可以发现，不同简支图像之间距离虽然紧密，但是并没有相互连接的情况，可以采用“连通域算法”将不同区域的简分别提取出来。对于二值图像，按照从左往右、从上往下的方式进行扫描，扫描到的点其像素值非 0 即 1。如果当前坐标没有像素则直接扫描下一坐标。如果当前坐标有像素，则须和它左侧的点及上方的点进行比较判断。若左侧和上方均没有像素，对当前坐标进行新的标记；若左侧或上方两点中的一个点有像素，该点（左点或上点）已进行过标记，当前点进行相同的标记；若左侧和上方的点均有像素且标记相同，进行相同标记；若左侧和上方的点均有像素但是标记不同，比较二值标记值的大小，以小值标记当前坐标，并且将已标记过大值的所有扫描过的点重新以小值标记。对扫描结果进行染色测试，发现各连通区域完美对应各简。以各简的左右侧极值点的横坐标之差作为图像宽度裁切，裁切效果如下图右（截取此简上半部分）：



二 切字

将各支简分离之后，进入切字步骤。岳麓书院藏秦简整体书写较为工整，对于大部分简可以直接采取定长窗口切分法，即原则上认为每个字所占空间大小相同，分配给其相同长度的窗格，裁切出相同大小的矩形。不过由于简文情况复杂，此方法亦存在明显不足，同时可以搭配上文提到的基于统计的投影过滤方法与连通图法进行处理。将图像投影到纵轴：

上图为简文纵向投影的横置图像，明显可以观察出该直方图的分布情况，对应上方简文“而舍之皆貲一甲”七个字。

三 二值化

图像二值化（Image Binarization）是将图像上的像素点的灰度值设置为 0 或 255，也就是将整个图像呈现出明显的黑白效果的过程。秦简所处年代久远，长期埋在地下，由于自然因素和人为因素等原因，字迹模糊的情况并不少见，图像和背景的区分不够清晰，二值化时阈值的选取便十分关键。

大津法（OTSU）是由日本学者大津于 1979 年提出的一种自适应的阈值确定的方法，又叫最大类间方差法。以灰度特征作为标准将图像分为背景和前景两部分，通过类间方差的比较确定最适合的灰度阈值。背景和前景之间的类间方差越大，说明构成图像的两部分的差别越大，当部分目标错分为背景或部分背景错分为目标都会导致两部分差别变小。因此，使类

间方差最大的分割意味着错分概率最小。对于图像 I(x,y)，前景（即目标）和背景的分割阈值记作 T，属于前景的像素点个数记作 NO（即灰度值小于阈值 T 的像素个数，此处设定前景为黑色），占整幅图像的比例记为ω0，其平均灰度μ0；背景像素点像素个数记作 N1，占整幅图像的比例为ω1，其平均灰度为μ1。图像的总平均灰度记为μ，类间方差记为 g，图像的尺寸为 M×N，则有：

$$\omega 0=NO/\ M\times N \quad (1)$$

$$\omega 1=N1/\ M\times N \quad (2)$$

$$NO+N1=M\times N \quad (3)$$

$$\omega 0+\omega 1=1 \quad (4)$$

$$\mu=\omega 0*\mu 0+\omega 1*\mu 1 \quad (5)$$

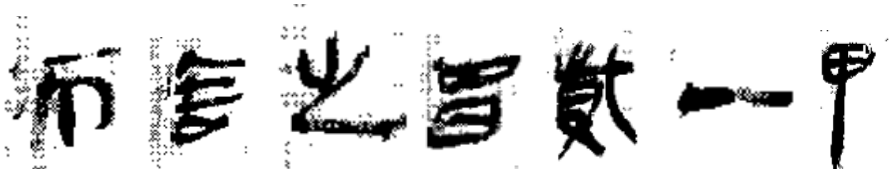
$$g=\omega 0(\mu 0-\mu)\hat{}2+\omega 1(\mu 1-\mu)\hat{}2 \quad (6)$$

将式(5)代入式(6)，得到等价公式：

$$g=\omega 0\omega 1(\mu 0-\mu 1)\hat{}2 \quad (7)$$

采用遍历的方法得到使类间方差最大的阈值 T，即为所求。

大津法将图像区分为前景与背景，与简文图像中的文字部分和竹简部分恰好可以对应，以文字作前景，竹简即背景区域。下图为使用大津法取阈值二值化后的图像，笔画清晰可见。对于部分图像背景之中的噪点，再稍加进行降噪、滤波处理即可。



综上，在文字图像提取过程中，首先使用基于统计的投影过滤法和连通图法将编联图版红外扫描图像进行“切条”，然后再综合使用上述方法与定长窗口切分法进行“切字”，最后用大津法求阈值得到文字的二值图像。利用该方法，可以大大提升工作效率，加速文字编图像整理的过程，为文字编研究节省大量时间和精力。

参考文献

[1]陳松長等.嶽麓書院藏秦簡的整理與研究[M].中西書局,2014(11).

[2]陳松長.嶽麓書院藏秦簡(肆)概述[J].出土文献研究,2015(00):23-26+8.

[3]陳松長.嶽麓書院藏秦簡.肆[M].上海:上海辭書出版社,2015(12).

[4]陳松長等.嶽麓書院藏秦簡(壹-叁)文字編[M].上海:上海辭書出版社,2017(06).

[5]肖攀.设计专用计算机软件辅助编纂古文字字编的设想[J].古文字研究,第三十二辑,2018(08):649-656.

[6]刘磊.基于内容的秦汉瓦当小篆文字识别方法研究[D].西北大学,2015.

[7]吴相锦.古文献文字图像分割与差异性比对算法研究[D].兰州交通大学,2016.

[8]张兰云.简牍文字提取与识别研究[D].西北师范大学,2017.

[9]李文英,曹斌,曹春水,黄永祯.一种基于深度学习的青铜器铭文识别方法[J].自动化学报,2018,44(11):2023-2030.

Application and Value of Image Processing Algorithm in the Process of Establishing Character Collection of Qin Bamboo Slips

Wu Zheng

(Hunan University, Changsha Hunan, 410006)

Abstract： In the past, during the process of character collection of the unearthed literature, especially of the Qin Bamboo Slips, it takes a lot of time and effort in manual image cutting. In order to solve this problem, a set of image processing algorithm flow, including Slips Cutting and words Cutting, which can assist manual processing and greatly improve the efficiency of character collection was proposed by this essay. The text of the first group in Qin Bamboo Slips Collected by Yuelu Academy Vol.4 would be used as example in this essay to illustrate the algorithms.

Keywords: image processing; character collection; Qin Bamboo Slips;