AI 是如何数字化《四库全书》的？

【马少平，王东】

我国历史源远流长。在漫长的历史发展过程中，历代文人墨客撰写了无数珍贵典籍，形成了我国所特有的灿烂文化。在计算机高速发展的今天，有必要利用现代化信息处理手段，对这些宝贵的文化遗产进行整理和研究，其中一个重要的问题就对古籍进行数字化处理。一方面，古籍数量庞大，存储和使用不便。

以文渊阁《四库全书》为例，全书共计３５７８种，６１４１函，３６３１５册，７９８９７卷。另有钦定《四库全书简明目录》3函、《四库全书总目》２０函、《四库全书考证》１２函、《四库全书分架图》４函、《古今图书集成》５７６函等 [1]。为了整理这部巨著，当年动员了 4 千多名学者，历时 10 年才手抄完成。即便是缩压成 1/4 大小的影印本，也重达 2.5 吨，售价几十万元。数字化以后，这些古籍可以通过网络在线阅读，方便检索，使用费用降低，极大方便了读者和研究人员。另一方面，数字化后的古籍为利用计算机技术对中国古典文化进行深入研究打下了基础，研究人员可以建立各种模型来探究古籍中蕴藏的知识宝藏。

图 国家图书馆收藏《四库全书》的书库

上个世纪 90 年代，基于清华大学提供的人工智能数字化方案，香港迪志文化和北京书同文有限公司历时两年时间，完成了《四库全书》的全部数字化工作[4]。该方案是如何实现的呢？

图 1. 数字化的《四库全书》图文光盘版

一，版面分析

数字化的第一步是对扫描出来的《四库全书》页面进行分析，分割出一个个汉字。看起来简单，但实际情况要复杂的多。例如，正常字体的正文之间可能会插入一些小字注释，有些相邻的字还可能有空间上的交叉。还有一些麻烦是当初那些编纂的官员们人为制造出来的。比如，为了拍皇帝的马屁，他们会故意弄了一些错别字，让皇帝审查的时候找出来，以彰显陛下英明神武。同时，为了表示对皇帝的尊敬，他们在书中提到臣子时会用小字，而提到皇帝时，字体会加大，而且故意突出出来。这些都给版面分析带来了很大困难。

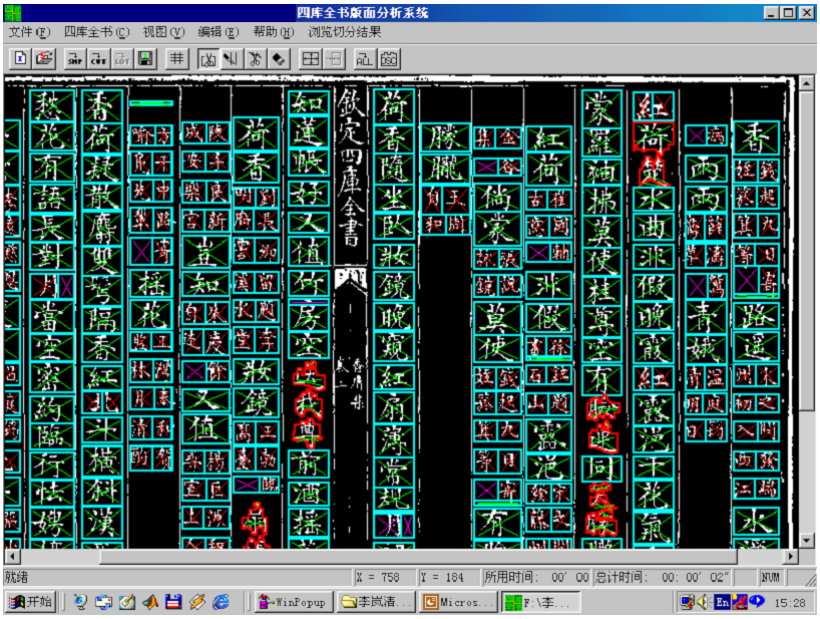

图 2. 《四库全书》版面分析结果示意图

二，非线性整形变换

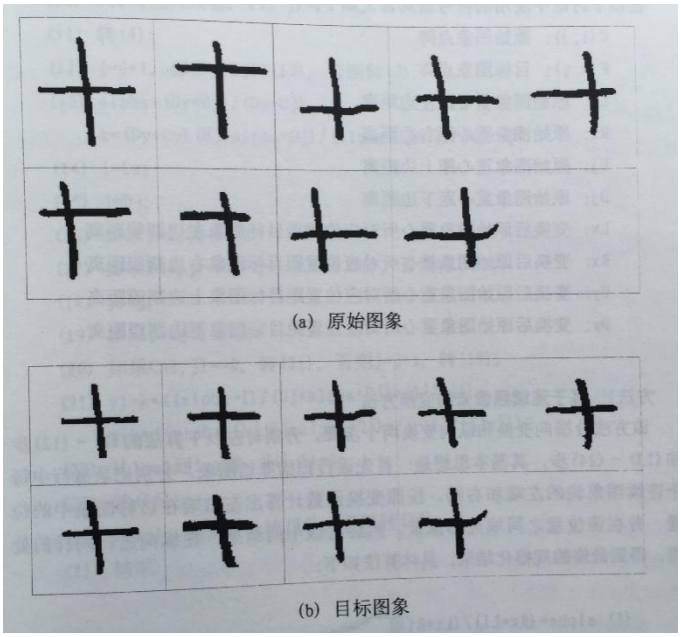

图 3. 非线性变换示意图

原本的《四库全书》由 3000 人抄写而成，虽然看起来比较工整，但是字体因人而异，差别比较大。为了提高识别率，系统首先采用非线性变换的方法对汉字进行归一化处理，使得同一个汉字看起来尽可能是一样的。图 3 是非线性变换的示意图，其中（a）图是原始“十”字，每个字差别较大，（b）图是变换后的“十”字，保留了原来的书写特征，但看起来有更好的一致性。

三，古汉字识别

对古汉字的识别采用近邻法，首先提取汉字图片的统计特征，再基于该特征计算马氏距离作为汉字间的相似度。因为古籍中的汉字缺少训练样本，研究人员采用了“滚雪球”式的增量学习方法：首先建立一个空的识别字典，每当遇到字典中没有的汉字时，由人工进行标注后加入字典中继续学习。另外，系统采用了基于非线性变换的样本增强技术，即反向应用前一节中提到的非线性变化方法，从一个汉字变换出多个变体，模拟不同人的书写方式，将这些变换出的样本一起送入学习系统进行模型训练。

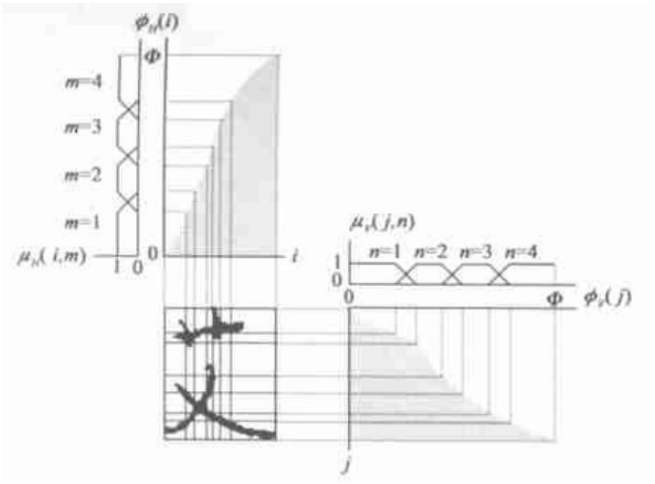

图 4：基于动态网格的统计特征提取[2]

在识别时，为了提高识别性能，系统同样利用了上述样本增强技术，对每一个汉字进行适当的变换，生成多个样本变体一起送入识别器，对识别结果进行综合评价，以提高识别性能。

四，人机协同的确认与校对

《四库全书》包含了近 3 万个不同的汉字种类，这是一个非常庞大的分类系统。由于训练样本有限，最终的首选识别率大概在 95%左右，但是十选识别率可以达到 99%以上。为此，研究者设计了一个人机协同的识别确认系统，机器提供汉字的图象以及最优的 10 选识别结果，由人工选择正确的汉字。这里并不需要操作者认识该古籍汉字，只要按照汉字的形状从 10 个候选中用鼠标点击正确的汉字即可，这样就可以实现正确率达 99%的高精度数字化。

最后，为了保证《四库全书》的数字化质量，系统提供了“横、纵”两种全局校对和一种局部重点校对。所谓横向校对，指的是按照正文顺序显示汉字图象和数字化结果，一一对照，由人工进行校对。所谓的纵向校对，是指系统将识别为同类汉字的扫描图象一屏一屏地显示出来，操作者只需要找出不一致的汉字图象即可。局部重点校对是指系统自动找出相似字、易混字和识别结果可信度低的字，重点进行校对。经过这样的校对后，抽查结果显示错误率小于万分之一，达到精品出版物的水平。

目前，《四库全书》已经可以在网上公开访问[3]，极大方便了读者阅读和学者研究。人工智能技术用于中文古籍数字化是一次有益的尝试，为古籍数字化开创了一条确实可行的路线。

[1]李芬林. 甘肃省图书馆对文溯阁《四库全书》的学术贡献[J]. 四库学, 2018(2).
[2]吴天雷 马少平，基于重叠动态网格和模糊隶属度的手写汉字特征抽取，电子学报，2004 年 02 期。
[3]http://www.sikuquanshu.com/main.aspx
[4]《四库全书》电子版问世，http://www.360doc.com/content/10/1224/09/803452_80861555.shtml