简析古文字识别研究的几个认识误区

刘 志 基

（华东师范大学　中国文字研究与应用中心，上海　200062）

摘要：古文字识别，属于人工智能范畴，故该项研究理应该带动古文字的研究应用进入人工智能化时代。基于这一认识，我们综合评估了二十余年来的古文字识别研究论著，认为其中存在几个阻碍研究走向成功的认识误区：将识别对象定位为可以经二手处理、少量选择的古文字，而不是原貌保真的古文字；将识别任务定位为仅仅识别出字形对应的今日简化字，而不是识别对象的各种信息的迄今学界认知；将识别的基本思路定位为通过少量对象的特征归纳去匹配全部对象，而不是识别对象的逐个唯一身份认定。

关键词：古文字；文字识别；数字化
中图分类号：H121　　文献标识码：A　　文章编号：1000-1263(2019)04-0089-07

古文字识别的研究已有 20 多年的历史，但迄今为止，我们还未看到面向社会公众的成功的古文字识别工具面世[①]。因此，评估一下既有研究的得失，找出问题所在，对于完成这项研究而言，是很有必要的。

既有古文字识别研究，主要从技术的角度，设计了一些将古文字原形图像识别成今日汉字字符集中某一个单位的方案。虽然所采用的具体方法林林总总，各有短长，但总体研究成绩已经证明，开发古文字识别工具在技术上是可以做到的。我们在这一方面的研究，也验证了这一点[②]。那么，阻碍古文字识别研究成功的因素是什么呢？综合评估研究现状以后，我们认为，问题在于既有研究在思想认识尚存在若干误区。仅就误区所涉及的几个方面谈一点管见，敬请方家批评指正。

一　古文字识别的对象

文字识别的对象自然是文字形体。而这种文字形体可以细分为两种：一是在识别工具实际使用中的识别对象，即根据各种识别需要经光电转换装置（如照相机、扫描仪等）变为电信号后送入计算机进行识别的文字对象；二是在文字识别工具制作中需要完成识别信息提取的对象，即所谓识别工具内储存的“识别字典”中的每个单位。因为只有先识别出“识别字典”中每个对象的特征信息，它们才能在识别工具使用中与前一种识别对象进行特征比对，进而完成识别。由于我们讨论的是文字自动识别的工具研制，所以所讨论的识别对象不仅指前者，也包含后者。

一般人们所说的“古文字”的具体内涵，可以有很大的差异，对古文字识别而言，以下几种差异是非常重要的：是一手的古文字（包括原始古文字材料的拓本或照相），还是今人摹写的二手古文字；是历史上真实存在的古文字，还是后人的仿古或夹杂自己“创意”的古文字书法作品；是完整的古文字字形

基金项目：教育部重点研究基地重大项目“系列古文字专题数据库建设”（18JJD740004）
作者简介：刘志基，男，1955 年生，上海人，教授，研究方向为汉字发展史、古文字数字化。
① 本文成稿于 2018 年 11 月，当月在复旦大学中文系以学术讲座的形式发表过论文主要观点。同年 12 月，在笔者的 2018 年“华东师大终身教授学术报告”上宣读过全文。截止投稿之时，网上有名为“字鉴”的文字识别软件，可以识别较多书法字迹。但对于古文字，笔者使用后发现识错率较高。
② 见课题组论文《基于编码的甲骨文识别技术研究》，陈婷珠等撰，《中国文字研究》，2019 年第 1 期。

集，还是其中的部分选择。我们认为，对古文字识别而言，在以上各个选择中，只有前一选项是正确的，因为只有这样，才能确保识别对象是原貌保真的。而后一选项则恰恰相反，会造成古文字识别对象不真实进而导致识别不能真正获得成功。遗憾的是，既有相关研究则多做后一种选择。

要求材料的第一手性质，是传统古文字研究的基本原则，而这一原则对于古文字识别来说，非但没有过时，而是更加重要。这是因为传统古文字研究中审视古文字材料的是人眼，而人眼对于古文字字形本质性的视觉信息具有一定选择能力，因此二手材料对研究的负面影响有可能由此得到一定程度化解。然而在古文字识别中，审视古文字字形材料的是机器的“眼”，要求机器的眼也具备对古文字字形关键信息的选择能力至少在目前是不现实的。因此，一手古文字字形材料中的非本质信息，也一样会被机器之“眼”所重视，也会成为识别中的一种依据。这种依据一旦被二手处理过滤掉的，识别对象就可能成为假材料。

然而，目前的古文字识别研究中却都很注重对识别对象的二手化处理，有的是以今人摹写的古文字为对象，如李锋等（1996）直接设定“我们识别的甲骨文是模（笔者按：“模”当为“摹”之误）写的”。其所列甲骨文字形类参考文献为高明《古文字类编》和徐中舒《甲骨文字典》，这两部工具书上的甲骨文字形均为今人摹写，可以确定该项研究所确认的甲骨文识别的对象是摹写的字形。周新伦等（1996）描述其识别过程“是先用扫描器输入待识别的摹写甲骨文”。陈丹等（2008）所设计的识别系统则是在“该系统在样本库中，临摹手写录入了左安民的《细说汉字——1000 个汉字的起源与演变》一书中的 735 个甲骨文字。”

有的研究者则把识别对象转换为字体，如周新伦等（1996）“根据目前已考证识出的 2860 个甲骨文，用手持式扫描器输入字符，在 386 微机上建立 1430kB 的甲骨文样本字形库（64x64 点阵）及对应的 270kB 汉字字库（32X32 点阵）。”很显然，字体相对于摹本，是二手化处理的加强版，大致来说，依照保留图像信息的丰富程度的从高到低，图像格式可分三种：彩色、灰度、黑白。摹本可以是这三种格式中的任何一种，而造字只能选择其中保留图像信息最低的黑白格式字形图像作底本。

识别对象之所以不应包含后人仿写的“古文字”，首先是因为识别这种所谓“古文字”意义不大，因为这种识别并不能带给我们历史文字的真实信息，对于我们解读历史文献，了解古史古文化也不会有真实的帮助。另一方面，这样做实际上是要去完成一个不可能完成的任务。比如有的研究者将识别系统识别对象定位为“认识名胜古迹中的古文字”，以增加人们“旅游的乐趣”（陈珺等 2017）。这样定位，显然会遇到如下问题：今日留存于名胜古迹中的真实古文字并不多见，大部分貌似古文字者只是后人（包括今人）的字迹，其中不乏偏离篆法传统的“创意”之作，有的甚至是除了书写者本人外无人能识的生造者，或者是写了错字别字的“古文字”，要让识别系统去识别出这样的“古文字”，除非是识别工具预先处理过这种随时随地会被“创意”出来的“古文字”的样本，否则并无可能。

既有研究都只选定了数量有限的古文字作为识别对象来开发古文字识别工具。李锋等（1996）“抽取 1035 个甲骨文字符进行自动识别实验”。但其试图完成的任务则是“要自动识别不同写法的 3000 多个甲骨文单词。”

周新伦等（1996）则“根据目前已考证识出的 2860 个甲骨文”，“在 386 微机上建立 1430kB 的甲骨文样本字形库（64x64 点阵）及对应的 270kB 汉字字库（32X32 点阵）”。

陈丹等（2008）的其识别系统“样本库”所包含的是“左安民的《细说汉字——1000 个汉字的起源与演变》一书中的 735 个甲骨文字。”

吕肖庆等（2010）则“选择了部分甲骨文文字的图片进行分类实验。一个典型实例中包含了八个古文字，“保，禾，牛，女，人，天，羊，祝”的相关图片，每个古文字选择了 30 张图片作为一类，每类选择 20 张图片用作训练，其余 10 张用作测试。”

李尚婕等（2013）的甲骨文自动识别研究所建的用以与识别对象进行比对匹配的“甲骨文数据库存储文字量为 5000 字”。

顾绍通（2016）“基于拓扑配准的甲骨文字形识别方法”研究中所用的是“含有甲骨文字形 3673 个”的“TTF 格式字库”。

刘永革等（2017）的“基于 SVM 的甲骨文字识别”研究，建立了甲骨文图文资料库，“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”。

上述从事甲骨文识别的各家研究，用作识别对象的甲骨文字数尽管多寡不一，但是相对真实识别环境中可能出现的甲骨文字的数量来说都少了很多。学界公认，目前已被发现存世的甲骨有十五至十六万片，每片有字甲骨上一般总有数量不等的少至个位数，多至数百的字形，仅就我们研发的甲骨文数据库统计，甲骨文字形数已超过一百一十万[①]。客观来说，上述研究者将识别对象限定为少数量古文字，有受限于已有的研究条件的原因，有的学者也希望通过增加识别对象的样本数量也提升研究效果，但却并未意识到限定识别对象这种研发思路的本身存在什么问题。如陈丹等（2008）对其研究结果作如下自我评估：“目前虽然限于条件，样本的数量偏少，结果有待进一步验证，但是根据上述结果可知，本文所提出的方法对于古文字的识别是有效的。”不难看出，这种被认定为“有效的方法”，在本质上就是试图以一小部分古文字为对象，去研制能够成功识别大批量古文字的古文字识别系统。而正是古文字识别研究的一个误区。识别对象之所以不能只选取部分，主要是因为针对小范围材料的调查来实现大范围材料的识别，只能设法在小范围材料中抽象归纳出能够覆盖大范围材料的字单位特征，但是这种意图就古文字识别而言既难以实现，又足以取消其应有意义。关于这一点，将在后文具体说明。

二　古文字识别的任务

迄今的古文字识别研发方案，均仅以将相关古文字字形归属为哪个字单位为最终目标。如“识别的结果在显示器上用甲骨文及对应的汉字显示在输入字符的右上角”（周新伦等 1996），“根据移动终端预存的古文字与简体字的绑定关系，查找出与匹配度最高的古文字对应的各简体字……”（陈珺等 2017）。有的研究者已发现识别任务这样定位会发生识别不清问题：因为“有些甲骨文字可能有不止一个与之同构的甲骨文”（栗青生等 2011）。然而，问题远不止此。

首先，仅仅完成这样的任务，古文字识别有可能产生的效益将被大大缩水。就其潜在的社会需求而言，理想的古文字识别，除了能确定字形归属为哪个字以外，还应该同时包含与这个字形相联系的全方位各层次的有认识价值的信息，比如这个字的意义、语境、出处、文献类型、时空属性等内容的识别。只有这样，才能真正通过识别，让人们获得古文字的汉字史、汉语史以及其他历史文化信息，大大扩展识别结果应用范围。这种包含系统信息识别，实质上就是通过识别来全面呈现学界对相关字形的各种认知，因而不仅在普及层面锦上添花，对各种专业研究，特别是古文字古文献研究业内已无识字障碍的人群而言，一定是特别被期待的识别效果。

换个角度来说，即便是为了识别出古文字字形属于哪个字，也需要依赖这种系统信息的识别来助力。在实际的识别过程中，作为识别对象的古文字字形一般都不会孤立出现，总附有这样或那样的关联信息，比如：甲骨文，是哪种著录，甚而其中的哪一片；金文，是哪个断代的，或哪个器铭；楚简，是包山还是郭店，甚而其中哪一篇哪一简。而每一种关联信息，都可以在特定的角度为相关识别对象增添一种识别特征，因而在识别中就可以相应增加条件限定，以缩小识别中比对的范围进而提升识别效率。值得注意的是，有的研究者已经注意到了语境在古文字识别中的作用，针对“甲骨文模糊字形的识别”，提出“利用语境分析生成的候选字库得到对应的甲骨文语义构件向量，然后结合基于 Hopfield 网络的识别结果计算待识别的甲骨文模糊字的匹配度，根据匹配度确定目标甲骨字”（高峰等 2014）的方法。具体来说，就是通过语境调查，确定相关字群的语境关联性，进而通过这种关联性，依据相邻字的出现频度来缩小被识别字的候选字群，进而帮助识别获得成功。虽然这一方案不无有待商榷之处[②]，但就兼顾字形以外的其他关联信息来实现古文字识别这一点来说，还是值得肯定的。总之，识别中引进的关联信息越多，越有助于古文字识别的效率提升。

进而言之，古文字识别作为最先进的文字处理手段施用于具有数千年历史的古文字，理应担负起助推古文字的研究和应用全方位与时俱进的历史使命。仅仅满足于将古文字字形对应出简化字，则如牛鼎

① 该数据库为国家社科项目“出土古文献语料库建设研究”成果，具体介绍详见后文。

② 甲骨文虽然是程式化程度较高的占卜文字，但程式化导致的关联字群，还是按类组有不同分布特点。该方案既然关注“甲骨文语义构件向量”，就应该在类组层面来寻找确定关联字群的“向量”，而不是将各不同类组的甲骨文混为一体来做这一调查。另外，仅关注识别对象的“甲骨字的前驱及后继甲骨字”，也有局限性。

烹鸡。汉字古文字研究，虽然有着悠久的历史，但是限于历史条件，在字处理方式上长期以来落伍于时代发展，进而导致资料获取查找的困难和问题研究中难以引进新的手段和理念。本世纪以来，随着中文信息化的发展，古文字数字化研究和相关数据库建设有所推进，为解决古文字研究的历史局限起到了一定的积极作用。然而，目前的古文字数字化建设尚存在一个盲点，虽然人们已经可以利用古文字数据库检索古文字的字形、文句、考释等信息，但是要获得检索的成功必须有一个前提，那就是必须预先确定要检索的是什么字。如果面对的是一个不认识的字，便无从利用古文字数据库来查找它的任何信息。而古文字识别的意义，就在于它可以消除古文字数字化现存的这个关键性盲点，即不借助其他条件，直接以古文字字形原貌为对象的释读检索。理论上说，抹掉这一盲点，就有可能盘活数字化营造的古文字大数据系统，进而推动古文字研究大踏步前进。

三　古文字识别的基本思路

既有古文字识别实现的基本思路，都是试图以一个文字单位的一部分字形为调查分析的对象，从中找出这个字的构形的共有特征，然后用这种特征去匹配该文字单位的所有字形，进而达到识别的目的。

李锋等（1996）设计了一种基于图论编码的甲骨文三级识别特征码，试图以此种技术实现甲骨文识别。其一级识别特征码，即甲类图论编码（六种图特征按块数、网孔数、端点数、三叉点数、四叉点数、五叉点数、六又点数的顺序排列起来得到的序列），二级识别特征，即乙类图论编码（邻点子图割集码），三级识别特征码，即端点方向编码。

周新伦等（1996）“ 提出一种两级分类的识别方法：首先，将待识字符抽象为一种图，并提取其拓扑特征进行第一级识别；然后，给出一种广义笔划定义，并在此基础上提取有关的特征进行第二级识别。”

陈丹等（2008）“ 提出了一种用于识别联机手写古文字的方法，详细介绍了所采用的笔画特征和字元特征，主要包括在古文字书写变形中具有较高稳定性的 7 种笔型特征,以及笔画交叉点、字元相对位置特征等等。并在此基础上,介绍了所构建的一个古文字识别的原型系统 ”。

吕肖庆等（2010）为探索与甲骨文自动识别相关的甲骨文自动分类问题 “ 提出了一种基于曲率直方图的傅里叶描述子———FDCH，并将这种特征进一步用于对甲骨文文字的分类 ”。具体是 “ 将甲骨文的文字轮廓提取出来后，首先需要计算轮廓上各点尽可能准确的曲率 ”，“ 为了描述图形的分布情况，我们利用各采样点与中心点的距离和角度关系构造直方图。”然后 “ 对其进行了二维傅里叶变换 ”，“ 最终得到基于曲率直方图的傅里叶描述子 FDCH 具有平移、旋转、尺度不变性 ”。

张霄军等（2006）认为 “ 特征提取过程 ” 是古文字自动识别的一个必要步骤，而 “ 笔画特征具有明显的抗干扰性、方向性和普遍性 ”，“ 由于任何一个汉字的每一笔画间都存在一个相对的位置关系，我们可以利用笔画间的特征信息来实现整字的识别 ”。

栗青生等（2011）提出 “ 甲骨文识别的图同构方法 ”，其 “ 甲骨文的图抽象 ” 的方式是 “ 从甲骨文中抽象出顶点和边并对这些顶点进行标号 ”，然后进行编码。

试图通过归纳一小部分古文字的构形特征，来研制能够成功识别所有或者大批量古文字的古文字识别系统，之所以被我们认定为一个认识误区，首先是因为归纳特征的做法是与前文言及的古文字识别的任务定位相背离的：既然将识别任务确定为 “ 要求识别出识别对象的各种具体信息 ”，识别对象就只能落实到具有各种具体信息的，即有语境、有载体，有时空属性的文献实际用字，而不能是排除上述具体信息的抽象的 “ 字 ”。

退一步说，即使归纳特征的做法可以认可，但具体实施起来也并不可行。古文字识别研究的归纳特征的做法，显然是因循今日手写汉字识别技术定式的产物。汉字构形经楷化演变后，在笔画、构件及整字各层面都形成了严格规范的标准形体，所有人学习写字的过程实际也就是摹写这种统一的标准形态的过程，所以一般的手写字迹都是基于这种统一标准而形成。因此，根据统一标准归纳特征去匹配不同手写字迹进而实现识别具有很大的可行性。相关开发研究也正是循着这一思路而开展的：“ 在联机手写汉字识别中，笔划特征关键点的有效提取对笔划的正确获取有着直接影响,进而也关系到最后整个字符识别的结果。”（孙阳光、何坚韧 2012）。目前，手写输入已成为人们使用手机或电脑的一种常用方式，而这种

成功很容易引导古文字识别依然因循特征归纳法。但是，先秦古文字是在未经“书同文”，也未见“正字法”的社会环境下出现的手写文字，因此与隶变楷化后的汉字形态截然不同，古文字无论在线条程式还是结体方式等构形要素上都没有一种绝对标准范式。“笔画特征具有明显的抗干扰性、方向性和普遍性”，“任何一个汉字的每一笔画间都存在一个相对的位置关系，我们可以利用笔画间的特征信息来实现整字的识别”（张霄军、陈小荷 2006）之类都是不切古文字实际的推测。我们的实验表明，人们很难归纳出只属于某个古文字单位的特征，进而用这种特征去匹配这个字单位所属的所有古文字字形。如果一定要去做这件事，那就只能把识别对象设定成一个小范围的字群。如前文所述，目前的各家“成功”的研究，正是这样的前提下完成的。然而，在每个书写者都能充分自由营构字迹的背景下，针对小范围识别对象的“成功”特征归纳，是极有可能因为新的识别对象加入而瞬间崩盘的。更何况，即使针对小范围的识别对象，“特征归纳”所能达到的识别率也多不能令人满意。李锋等（1996）以 1035 个甲骨文为识别对象，“三级累计识别率达 92.27%”；周新伦等（1996）以 2860 个甲骨文为识别对象，“提出的算法，其识别率达 94%”顾绍通（2016）基于拓扑配准的甲骨文字形识别方法根据“图论和笔划特点”，以 2860 个甲骨文为识别对象的识别率为 94%，根据“图特征”对 1035 个甲骨文的识别率为 92.27%。而有的研究者也在某种程度上意识到了此种研究思路的问题，如陈丹等（2008）对研究方案有如下自我评估：“本系统有些方面还有待改进。例如：在字库设计方面没有考虑到古文字的异体字，每个汉字只存有一个文字作为标准样本，从而在一定程度上降低了系统的识别率和应用范围。”吕肖庆等（2010）“选择了部分甲骨文文字的图片进行分类实验。一个典型实例中包含了八个古文字，“保，禾，牛，女，人，天，羊，祝”的相关图片，每个古文字选择了 30 张图片作为一类，每类选择 20 张图片用作训练，其余 10 张用作测试。”而其实验结果表明，有一半字分类准确率未达 100%，其中 3 字的分类准确率为 80%，1 字准确率 90%。作者的总结是：“通过对分类错误结果的分析，我们也发现该方法目前对多级轮廓或是多个并列的轮廓，不易得到准确的特征描述，这将是我们下一步改进算法的研究方向。”刘永革，刘国英（2017）“采用支持向量机分类技术研究甲骨文字图片的识别技术”选取 15 个字符的“1290 个甲骨字进行识别”，“试验证明达到 88%的准确率”。其“实验结果分析”中坦承：“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。”

既然各种归纳特征的“算法”面对小范围的识别对象都无法获得满意的识别率，那么这种识别思路的合理性也就难以得到证明了。不宜走“归纳特征”的路线，同时识别对象又应该落实到古文字材料的每一个文献用字上，那么，古文字识别研究的基本思路就应该是通过把每一个识别对象与其他对象区分开来以实现识别，这也就是需要赋予每个可能被识别的古文字字形以唯一的身份识别码。就技术而言，这种逐个分开的做法没有根本性障碍。实验表明，现在的图像处理技术，完全可以做到将逐个古文字字形图像转换成一串与其图像唯一性相匹配的数字。关于这一点，我们在既有研究中也可以看到方向正确的具体做法，如陈丹等（2008）设计的“整字特征识别”：

整字特征是在字元识别的基础上，分析字元之间相对位置关系进行编码。字元的相对位置，实际上指的是字元重心的相对位置，其计算方法如下：

（1）将整个输入区等分为如下 9 个部分：

1	2	3
4	5	6
7	8	9

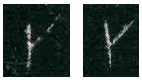
图 6　输入区域等分图

（2）依次计算每个字元的重心，确定该字元落在以上 9 个区域的哪个部分，则该部分有字元的数目增加 1；

（3）统计落在这 9 个部分的字元的数目，得到一个 9 位的字符串，即代表字元间的位置关系。

研究者是在所谓“字元”特征识别提取的基础上提出这一方法的，而“字元”层次的设置实际可以上移到整字层面。我们的甲骨文识别课题组尝试运用 MATLAB 软件的古文字图像编码识别方法，采用 3*3 的图像分割比例（与陈丹等“输入区域等分图”同）直接读取整字图像，即可精确获得每个古文字

字形的唯一数据编码。如村中南甲骨 39 片有两个“ 卜 ”字，尽管出自同一刻手的同一书写过程，字形十分相似：



但依然可以被编码有效区分。前一“ 卜 ”字被读取的唯一编码为如下九组数据：

0	2.25595614589922	2.12945393211048
0.179211469534050	18.9542483660131	1.12797807294961
0.432215897111533	5.94560404807084	0

后一个“ 卜 ”字唯一编码为如下九组数据：

0	3.89393939393939	3.90151515151515
0	22.2500000000000	4.19696969696970
0	4.70454545454545	0

两者绝对差值 11.6272736856228。如果以 3*3 的图像分割比例读取的字形数据尚不能满足区分古文字字形唯一值的精度要求，尽可以进一步细化这种分割比例：8*8，16*16，甚至 32*32，直到能够满足需要为止。

当然，逐个分开唯一认定的研发路线，将完全改变既有古文字识别的研究模式，攻关的重点由研究“ 算法 ”以找出不同字形的字单位归并特征，变为海量识别对象载体的数字化转换及其在数字平台中的信息分析、整理标注。而这种研究模式的改变，实际上就是要把古文字自动识别研究拉回到古文字数字化的既有发展轨道上来。

古文字自动识别必须有相应的大数据计算机平台作为它立足的基础。如果没有这个基础，无论创意出何种“ 算法 ”技术也都是枉然。这就如同鲜榨果汁，榨汁机研制得再好，如果没有匹配这种机器的鲜果原料供给，也出不了一滴榨汁。而在另一方面，在不能清晰了解作为榨汁对象的鲜果原料的情况下，又如何能够研制出真正管用的榨汁机？就此来看，古文字识别研究，只是纠结于“ 算法 ”技术的设计，却不注重支持识别的古文字大数据平台建设，又是一个思想误区。

然而，连带而来的一个问题就是，完成如此海量的工作有没有可行性。答案也是肯定的。古文字数字化研究，起步于上世纪末本世纪初，已历经 20 年的发展，由于研究者的坚持和国家政府科研项目的持续投入支持，已经取得了可观的成绩。仅以 2017 年结题的国家社科重大项目“ 出土古文献语料库建设研究 ”为例，该项目结题成果得到全国社科办如下评价：“‘ 出土古文献语料库建设研究 ’课题组，建成《战国简帛文》《商周金文》《殷商甲骨文》三个网络数据库，为迄今最完备的古文字网络数据库。”[①]该项目结题报告之《成果简介》称：项目最终成果“ 先秦古文字网络数据库 ”，“ 在材料收录上以穷尽目前已公布古文字资料为原则，在数据库功能开发上根据出土古文献的材料价值特点，注重多种功能的实现。如多层次保真，关联相关考释研究信息，呈现古文字材料字形和用字特点，揭示字际关系等。”很显然，作为先秦古文字的主体甲骨文、商周金文、战国楚简帛文字，材料已经实现了全面的数字化转换，而所有文献用字的释字、释义、语境、字际关系、构形分析、出处、文献类型、时空属性等信息已经得到整理标注，这种研究成绩，为古文字识别研发提供了所需要的古文字数字平台。毫无疑问，全国乃至全球范围内此类古文字数字化建设成果并不在少数，以此例之，可以认定，新的研发模式不存在可行性的问题。

语言文字自动识别已属于人工智能范畴，作为最先进的文字处理手段施用于具有数千年历史的古文字，理应担负起助推古文字的研究和应用全方位与时俱进的历史使命。据此，我们认为古文字识别应该带动古文字研究进入人工智能化时代。人工智能对于海量材料数据的记忆、运算具有超强能力，并能模拟人的意识、思维，利用大数据来进行机器学习、深度学习，进而延伸和扩展人的智能。人的古文字研究的基本思维方式，就是全面把握文字的形、音、义及各种相关属性来发现和解决问题。而制约古文字

① 全国社科办：《2017 年度国家社科基金基础类重大项目结项成果概述》，全国哲学社会科学规划办公室网站 http://www.npopss-cn.gov.cn/n1/2018/0319/c219545-29875808.html，2018 年 03 月 19 日。

研究进步的一个最大问题，就是人脑的记忆能力相对海量材料的识记要求捉襟见肘，由此，思辨、发现的空间也被大大压缩。因此，推动人工智能的触角广泛覆盖古文字原始材料和研究资料，通过古文字识别来打通古文字数字化的各类数据关联，对于推动古文字研究走向智能化具有至关重要作用，乃是古文字识别任务的重中之重。正是基于这一认识，我们才认定既有古文字识别研究存在某些认识的误区。而这些认识误区，虽然涉及不同方面，但实际又是具有内在联系：非原貌保真的识别对象、抽象字单位的识别任务、特征归纳的识别思路无疑是互为因果的，可以归纳为一种成体系的古文字识别思想认识。与之相应，我们在对既有古文字识别研究的分析评论中实际提出的若干认识也是自成体系的，那就是以客观存在的逐个古文字文献用字的保真形态为对象，采用逐个对象唯一身份认定的识别方式，通过识别，获得各识别对象具有认识价值的系统信息。

【附记】2019 年 5 月，华东师范大学中国文字研究与应用中心发布了该机构的古文字识别首个成果“商周金文智能镜”。本文提出的研发思路，在该识别程序开发中得到落实和验证。

参考文献：

陈丹、李宁、李亮　2008　古文字的联机手写识别研究,《北京机械工业学院学报》第 12 期。
陈珺等　2017　一种古文字识别系统及方法，专利申请号：201710614296.X。
高峰、吴琴霞、刘永革、熊晶　2014　基于语义构件的甲骨文模糊字形的识别方法,《科学技术与工程》第 10 期。
顾绍通　2016　基于拓扑配准的甲骨文字形识别方法,《计算机与数字工程》第 10 期。
李锋、周新伦　1996　甲骨文自动识别的图论方法,《电子科学学刊》第 12 期。
李尚婕等　2013　一种甲骨文自动识别方法，专利申请号：201310480306.7。
栗青生、杨玉星、王爱民　2011　甲骨文识别的图同构方法,《计算机工程与应用》8 期。
刘永革，刘国英　2017　基于 SVM 的甲骨文字识别,《安阳师范学院学报》第 2 期。
吕肖庆、李沫楠、蔡凯伟、王晓、唐英敏　2010　一种基于图形识别的甲骨文分类方法,《北京信息科技大学学报》第 12 期。
孙阳光、何坚韧　2012　联机手写汉字识别系统中特征提取方法研究,《计算机光盘软件与应用》第 24 期。
张霄军、陈小荷　2006　古文字自动识别过程及其程序实现,《中国文字研究》第七辑，广西教育出版社。
周新伦、李锋、华星城、韦剑　1996　甲骨文计算机识别方法研究,《复旦学报》（自然科学版）第 10 期。

A Brief Analysis of Several Misunderstandings in the Study of Ancient Chinese Character Recognition

LIU Zhi-ji

(Center for the Study and Application of Chinese Characters, East China Normal University, Shanghai 200062, China)

Abstract: The ancient script recognition belongs to the category of artificial intelligence, so the recognition research should drive the research and application of ancient characters into the era of artificial intelligence. Based on this understanding of the recognition research into the era of AI, we have comprehensively evaluated the ancient Chinese character recognition research for more than 20 years and found that there are several misunderstandings that hinder the success of the research. These misunderstandings include: 1) the recognition object as a small selection that can be processed by second-hand, instead of the original fidelity; 2) The recognition task only to the simplified characters of today corresponding to the glyphs, rather than the so-called academic cognition of various information of the recognition object; 3) The basic idea of identification to match all the objects through the feature induction of a small number of objects, instead of identifying the individual identification of the objects one by one.

Key words: Ancient script;Text recognition; Digitizing