基于人工智能的甲骨文识别技术与字形数据库构建*

门　艺　张重生

【摘　要】已有的甲骨文工具书在整理模式、编排方式、检索方法、呈现手段等方面积累了经验，但也有字形分合无标准、检索困难、隶定方式混乱等现象。纸制工具书不利于复用、优化和改进，且无法对每个字形的相关资料穷尽性枚举并提供相关的上下文信息。利用新一代人工智能技术处理数量庞大、表义复杂、字形多变的甲骨文字，训练深度学习模型，实现在全张拓片上对甲骨文字逐一自动定位、识别，可依据字形自动进行分类，在此基础上建设材料完整、语段信息方便查看、开放式编辑的甲骨文字形数据库。进一步设计相关甲骨文识别与相似性检索软件，使字形数据库成为以图片检索甲骨文的便利工具。

【关键词】甲骨文；字形；人工智能；数据库

【作者简介】门艺，女，河南大学黄河文明与可持续发展研究中心研究员、硕士研究生导师，研究方向为古文字学、甲骨学；张重生，河南大学计算机与信息工程学院教授、博士研究生导师，研究方向为人工智能、机器学习。（河南 开封　475000）

拓本、摹本或照片是甲骨文的一般著录形式，这些图像文件为学者们研究甲骨文的重要第一手材料。为了使甲骨文的内容更便于一般研究者阅读，学者们通过摹写和转写的形式，编辑甲骨文释文和各种工具书，是为甲骨文研究的第二手材料。甲骨文的材料相当庞大，仅《甲骨文合集》就著录有 4 万多片，总字数达 110 多万字。如此巨大的数量，人工编纂极为耗时费力；并且由于各种原因，人工对甲骨文的处理经常出现重复劳动的现象，研究者大量的时间和精力都浪费在了校验第一手材料、重复检索、重新梳理材料上。用新的科技手段帮助人工处理甲骨文，将大大缩减甲骨学者从事相关研究的时间成本。在 IT 时代，学者们选择造字、将各种甲骨文资料输入电脑，以便在电脑上呈现和检索，初步实现了甲骨文的电子化。在新一代人工智能时代，甲骨学者可以选择更为智能的方式，利用人工智能技术，对海量的甲骨文图像进行自动文字定位、提取、识别和检索，使得对第一手材料的收集和整合自动、智能、便捷。本文首先综述甲骨文字整理方法的经验与不足；然后结合当前的新一代人工智能技术，提出甲骨文字识别的相关构思；最后，从实现路径进行详述，具体介绍基于人工智能的甲骨文自动识别相关技术，以及甲骨文字形数据库的自动构建。

一　甲骨文字整理方法的经验与不足

甲骨文字的整理仅是甲骨文整理的一个方面，由于人工智能识别甲骨文是以文字为主要识别对象，并在此基础上自动生成甲骨文字形库，因此本节仅就甲骨文字整理方法等技术层面进行述说。

（一）甲骨文字形的呈现方式

就目前所见到的材料来看，展示甲骨文字的方式主要有摹写、造字和剪字三种，其中摹写是使用最普遍的一种方式。

*　基金项目：本文为黄河文明省部共建协同创新中心/黄河文明与可持续发展研究中心项目“甲骨信息标注与人工智能辅助缀合研究”（编号 2020K05）的阶段性成果。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

1. 摹写

摹写出甲骨文的字形，是甲骨文得以呈现的最便捷形式。孙海波先生的《甲骨文编》就因摹写精良而为学界使用多年，李宗焜先生的《甲骨文字编》也是摹写的形式，对甲骨文字形体的把控非常好，现在多为学术界使用。摹写的方式有一点最大的不足，即在拓片不清或字形较复杂的情况下，同一个字形在不同的摹写者手中会呈现不同的面貌，摹录的字形反映的是摹写者对字形的认识和看法，即带有主观意识的摹写字形，这样的字形还有可能会成为学者进一步讨论的基础。

2. 造字

计算机的普及对甲骨文的研究带来了新的机遇和挑战，尤其是大量古籍数据库的使用，使甲骨文研究人员看到了广阔的研究前景和方便使用的美好愿景。然而，横亘在研究人员和甲骨文数据库之间的是甲骨文字形的处理。如何使甲骨文字形能像现代汉字一样在电脑中直接显示出来，并可以进行检索，就成为计算机专业人员和甲骨学者共同关心的问题。

经过学者们不懈的努力，已有很多研究机构推出了自己的甲骨文字体文件，比较常用且与甲骨文释文库相关联的字体文件主要有：香港汉达文库甲骨文字体，浙江师范大学甲骨文字体，安阳师范学院甲骨文字体，华东师范大学甲骨文字体等。以前三种字体为基础，分别出版了甲骨文字编类工具书，即《甲骨文字形表》《甲骨文字新编》《殷墟甲骨文编》。

造字一般采用一款造字软件，在电脑上用鼠标或电子笔摹画出甲骨文的形体，把这些造出来的字形做成字体文件，使用者在安装字体文件后便可以在自己的电脑终端显示出甲骨字形。造字实际上就是以计算机进行甲骨文字形的摹写。因此造字除了有摹写的缺点（字形不准确，包含主观字形）以外，还有字形过于统一的缺点，不利于使用者认识甲骨文的原貌。

3. 剪字

既然摹写和造字都不能反映甲骨文的原貌，研究者们就想出了直接把甲骨文字从拓片上剪切下来编写甲骨文字形工具书的方法。各类图形处理软件的广泛使用，给直接从原材料图像上切字编辑甲骨文工具书提供了便利。2009 年出版的《新甲骨文编》即是这样的一部工具书，刘钊先生在斟酌处理甲骨文字形方法后，决定使用“电脑切割原形后加以黑白翻转，并去掉字形无关的部分”。王蕴智教授《甲骨文可释字形总表》也是通过切字形式编辑的甲骨文字工具书，与《新甲骨文编》不同的是，直接从拓片上切下文字后，没有进行黑白翻转，而是把甲骨文字在拓片上的原形和原始尺寸展现给读者。用剪字的方法处理甲骨文字，基本可以反映甲骨文字形的原貌。在拓片上直接切字，保证了甲骨文字形的真实性，对原始拓片的要求颇高，对于一些甲骨表面残蚀严重的拓本，还是高水平摹本的清晰度更高，更能反映字形结构。同时直接在拓片上剪切原篆字，也不免会出现因去除拓片上残渣痕和其他文字的痕迹而致该字缺笔变形的现象。

甲骨学者们对甲骨文字形的逼真度要求比较高，为了适应这种要求，甲骨文字工具书的编纂者们以更加清晰的图像摹写，或造字时通过映射扫描来建构甲骨文原型字，或直接将甲骨拓片上的原篆字剪切下来使用。这些处理甲骨文字手段的优缺点，都为人工智能甲骨文识别技术提供了借鉴。

（二）甲骨文字的收集、标注和检索

1. 甲骨文字的收集

已经著录的甲骨片上的文字，都应是收集整理的对象。由于甲骨著录书也在不断的出版，新材料不断涌现，所以甲骨文字工具书的收字范围往往标记所使用材料的年限。可是甲骨文的总字数非常庞大，有些单字的字频超过了 2 万次，在传统的纸媒出版物中，逐一收录所有的甲骨文字不太现实，势必引起“资料的齐全与使用的方便”之间的矛盾。因此大部分甲骨文字工具书对于字频较高的文字，都是选择有代表性的形体进行收录。如“卜”字，《甲骨文字编》仅出 3 个字形，下注“‘卜’字见于甲骨无虑千万，大略不出此三形，举此以概其余”。《甲骨文合集》中“卜”出现的次数为 2 万次以上，字形的

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

确没有太大的变化，而无名组和何组各有一类字体中“卜”字兆枝那一笔总是向下作“卜”形，比较有特色，对甲骨文的字体划分和组类之间关系的探索也许有一定的价值。全面而穷尽性的收集材料中的所有文字，对于窥见字形的全貌，借以考察和对比甲骨文在各个不同时期不同组类的变化情况是有用处的。人工来做这样全面收集的工作则是耗时费力，对操作者的学问长进而言意义也不大。

2. 甲骨文字的标注

甲骨文字编类工具书一般以展示甲骨文字形体为主，不释义、不附辞例。标注的信息主要是来源出处，便于使用者核对第一手材料，以及查看辞例。近年出的字编类工具书有的还标明字的组类情况，这对查看字形在不同组类的变化以及一段时间内的演变情况是有一定帮助的。还有一些字编类工具书有一些简单的按语，以及字形字义的简单分析。如果要检索辞例的话，还要再查相关工具书或核对原片。而《殷墟甲骨刻辞类纂》等检索语段的工具书由于篇幅的限制，对于用例非常多的文字及其语段仅是节录，也不能很方便地查到所需要的全部材料。把字编、字典、语段检索以及其他需要检索的内容汇于一篇，是每个古文字研究者的愿望，而在纸媒的条件下实现起来会非常困难。李守奎先生曾经说：“如果给字编所收字都附上辞例，注明每个字的用法，字编就兼具了字书的功能，对读者会有极大的便利，这是毫无疑问的，但制作这样的字编工作量非常大，字编的篇幅会大量膨胀，出版单位和读者能否承受都尚未调查。”

3. 甲骨文字的检索

如何快速简便地找到所需要的甲骨文字，各种甲骨文工具书在编排以及检索方法上进行了不断的探索。最初《殷虚书契考释》对已识字是《尔雅》式的以义类聚合文字，这种编排方式并未被以后的甲骨文字工具书所沿用。研究者们更注重文字形体方面的关联，因此从之后的《簠室殷契类纂》《殷墟文字类编》开始，便以《说文解字》为序，以《说文》部首系联文字，可以按《说文》字序检索甲骨文。学者在按《说文》字序排列甲骨文时一般说来已识字为正编，未识字为附编，然而对于已识字和未识字“研契诸家没有一个共认的成文的标准”，不同的编纂者对已识字和未识字的把握不同导致各甲骨文字编收字、编排的不同，使用者和编纂者之间因同样的原因导致检索时的诸种不便。有些甲骨文工具书为了方便使用者，编制了隶定字笔画索引、可识字音序索引等。而隶定古文字是一个比较复杂的问题，也没有形成统一的标准，用笔画检索的效率也不是很高。音序索引则仅适用于可识字。

为加强甲骨文形体之间的联系，学者们按照甲骨文本身的特点制定了甲骨文部首，把甲骨文字形按自然形态分类与编排。《殷墟卜辞综类》《殷墟甲骨刻辞类纂》《甲骨文字形表》《甲骨文字编》等甲骨文工具书均按甲骨文部首排列，除《甲骨文字形表》外，其余三种编制了“原篆偏旁索引”，即可根据一个甲骨文中所包含的任何一个偏旁找到该字，这种以形索形的方法使用起来非常方便。《甲骨文字编》还提供了与《殷墟甲骨刻辞类纂》和《甲骨文编》的对照表，用户体验极佳。刘钊先生 2018 年完成的国家社科基金重大项目《甲骨文已识字、有争议字和未识字综理表》也提供了一个各个甲骨文字编工具书之间的对照表。对于将来大数据的关联非常有用。

做到对已有材料中甲骨文字的穷尽性搜集，标注相关信息，减少与第一手材料之间的差异，可随时调取第一手材料复验，以各种条件检索项目，聚合所需材料。这样的理想研究手段，目前可以通过基于人工智能的甲骨文识别技术做到。

二　人工智能识别甲骨文的理论依据与设想

如果按某种字编或释文关联甲骨文图像，实现一部分的检索功能，那仅是书本材料的电子化，而非人工智能。要使人工智能成为甲骨文研究的利器，第一步、也是最关键的技术是得教会人工智能认识甲骨文。现在这项技术已经具备一定的数据基础和理论依据，使我们在人工智能识别甲骨文方面

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

提出一系列的设想。

(一) 大数据的基础

当前的人工智能技术是大数据驱动的、数据饥饿的,非常依赖于大量的数据作为训练原料。近年甲骨文大数据云平台的建设如火如荼,完整的甲骨文图像原始资料和原文释文等资料最终将汇聚在云平台上。这构成了人工智能识别甲骨文的大数据基础。

以拓本为主的甲骨文著录书是世人得以认识甲骨文的基本材料,如今已经出版的甲骨文著录书基本全部扫描成数字图像,还有国家图书馆"甲骨世界",台北"中研院"史语所的"甲骨文数位典藏"等图像资料可供参考。在甲骨文著拓方面存在有重片、重复著录、缀合等现象,因此在已经出版的甲骨文数量统计方面数字也是变动的。学者们通过人工的观察与梳理,整理出很多重片和重复著录,并且标注了缀合的片号,建立起这些图像之间的联系,为进一步统计甲骨文的数量奠定了基础,也为人工智能识别甲骨文字提供了基本的素材。

自 20 世纪 90 年代开始,计算机专业人员便与甲骨文研究人员合作建设甲骨文数据库,香港汉达文库甲骨文库的建成是一个标志性的成果。其后的甲骨文数据库建设基本都采用汉达文库的建库模式,在汉达文库的基础上不断增加新出版的甲骨文材料,吸收新缀合甲骨文成果不断扩容。目前最大最全的甲骨文数据库当数浙江师范大学陈年福教授所做的甲骨文原文释文数据库,该库收集了 32 部甲骨文著录书中近 8 万片原文和释文材料,多达 340 万字。为建设此库,陈年福教授以造字的方式梳理和摹造了 13000 多个甲骨文不同形体,甲骨文原文通过安装"浙师大甲骨文"字体得以在电脑中呈现,并可实现全文检索。甲骨文原文释文数据库为甲骨文语料的各项研究提供了便利的检索条件和结果,必将推动甲骨文语言词汇研究的深入,也为人工智能识别甲骨文字提供了可关联的大量矢量字形。

(二) 新一代人工智能技术的重要突破

近些年的新一代人工智能技术的重要突破,给甲骨文的智能识别提供了解决问题的新思路。新一代人工智能所依托的主要技术是深度学习。深度学习使用深度神经网络,利用大量的数据(大数据)进行迭代训练,自动提取关键特征,得到高准确度的模型。最近几年,基于深度学习的场景文字定位与识别方法也取得了重要进展,其性能较传统方法有了显著提升。其中,场景文字定位方法可分为基于传统区域推荐的方法、基于区域推荐网络的方法、基于分割的方法以及基于区域推荐与分割的混合方法。尤其是混合方法,采用整体化思想避免阶段错误的累积,能够较好地进行精细化检测。场景文字识别方法可分为卷积网络循环网络结合的方法和注意力机制的方法。这些研究成果为基于人工智能的甲骨文自动识别研究的顺利开展提供了重要参考,而穷尽性的甲骨文原篆字形数据库正是其数据来源和支撑。

(三) 人工智能识别甲骨文的主要构想

人工智能识别甲骨文并不是让计算机做文字学家的工作释读甲骨文,而是让计算机通过图像匹配等方式,识别甲骨文字,与已有的数据比对,以便快速找到所需要的材料,或在很短的时间内提供大量的关联材料,并进行简单的分析。图 1 给出了所设想工作的主要步骤,包括如下七个部分:其中,第一部分是人工对部分拓本图像进行按字标注,包括每个文字的区域及对应的原文和释文(已完成);第二部分利用人工智能中场景文字定位的相关算法,对第一部分中已经标注的图像训练文字检测模型,以便对未标注的文字自动定位相关的文字区域(已完成);第三部分基于第二部分的成果,对已经定位到的文字区域中的原篆字图像进行相似性计算并归类;第四部分由人工再次确认第三部分中的结果,确保其正确性;第五部分使用第四部分和第一部分中的数据,利用人工智能技术中的场景文字识别算法训练文字识别模型;第六部分针对很多字形的样本量严重不足的问题,设计针对性的、适用于样本量很少的文字的识别算法;第七部分旨在利用第二、第五、第六部分的算法成果及模型,开发相关的甲骨文识别与检索系统和可视化界面。下面将依次对这些部分进行介绍。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

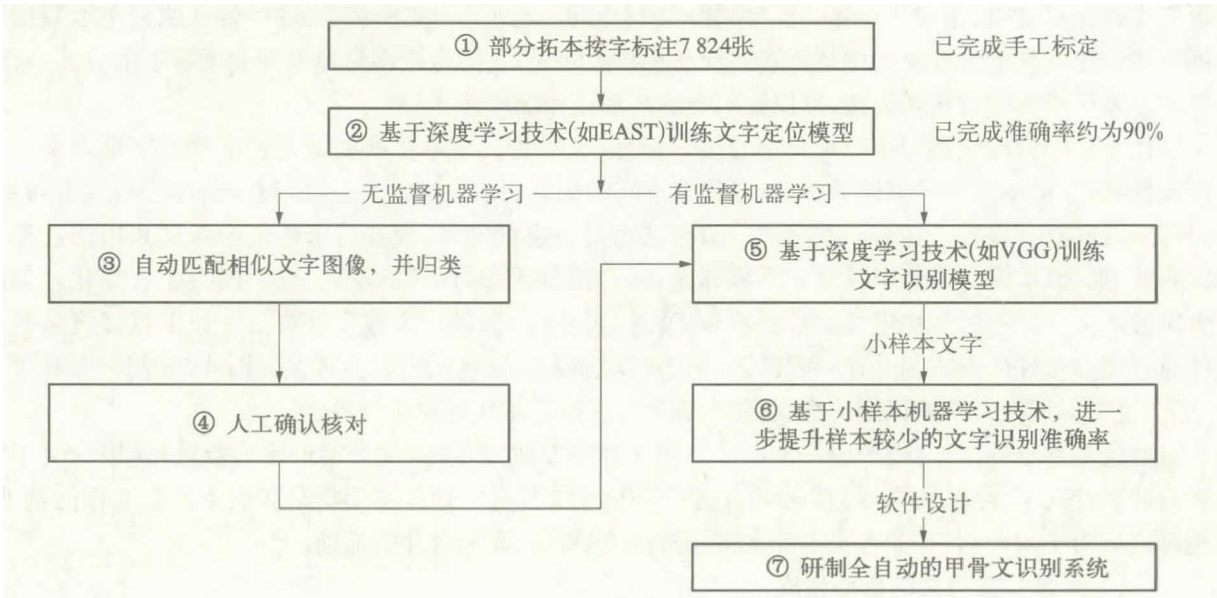


图 1　人工智能识别甲骨文技术路线图

三　基于人工智能的甲骨文自动识别技术与字形数据库构建的实现路径

本部分主要依据图 1 中所设想工作的主要步骤，在技术路线上分别进行陈述。

（一）甲骨文定位

我们的第一和第二步工作，已经实现了对甲骨文拓片上文字的自动定位。首先，两名甲骨文专家和十名甲骨文方向的研究生，耗时一个月有余，人工标注 7824 张甲骨文拓片；然后，根据目前经典的场景文字定位算法实现对拓片中的甲骨字的自动定位。通过深度学习的迭代训练，计算机可以将未标注的甲骨拓片中的文字部分通过矩形框的形式标示出来，其精度已达到 90%左右，较好地实现了拓片中甲骨文字的定位。不准确的结果多为构件间缝隙较大的字形。对于这个问题，我们需要依靠人机结合的形式进行校正。对于一些有争议为一个还是两个字的，我们也将标注说明采用的某家说法或自己的判断。

矩形框定位甲骨文字后，通过计算机程序可以自动进行剪切，并保留单个甲骨文字的位置信息，以便将来与甲骨文原文释文数据库的数据进行关联。带有信息的甲骨文单字裁切结果，我们称之为原篆字，先以甲骨片为单位归类存放。

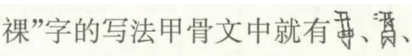
定位和裁切所使用的甲骨拓片，包括所有已出版的甲骨著录书中的所有拓片，对于重复著录和重片的情况，根据定位和裁切的结果，进行自动合并，选择更加清晰的字迹予以保留。通过本部分的操作，可以形成一个穷尽性的比较原始的甲骨文原篆字形资料库，为下一步的工作做好资料准备和据以核校的基础。客观上，通过该步骤的资料对比，还可以发现和校验重片，对甲骨文的重片和重复著录的情况可以有一个比较清晰的认识。

（二）甲骨文原篆字比对与归类

原始的甲骨文原篆字形资料库仅是切出了甲骨片中的文字，而没有对文字进行比对和归类，本步骤的工作就是用人工智能进行图像比对和筛选，自动把相同的字形归入同一个文件夹。人工智能识别甲骨文是以形体为主要的识别方式，在梳理和归类甲骨文字时把字形作为第一标准。通过原篆字

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

形图像的比对碰撞、相似性计算，把字形相同的自动归类到一个文件夹，实现甲骨文原篆字形数据库的初步构建。其中会涉及到调整识别同异度参数的问题，我们将按逐渐降低同异度要求的方式，对甲骨文形体进行多轮对比和筛选，以保证归类的客观性和准确性。

把字形作为梳理和归类甲骨文字的第一标准，是甲骨文字本身所要求的。甲骨文字形表义具有严谨性，两个形体上差异细微的字，可能代表不同的词，如：人与匕，犬与豕，旬与云，又与左，大与尹，目与臣，视与见，等等。还有一些字形以前被认为是一字的异体，现在看来他们的意义和用法还是有区别的，如帝、示等。当然甲骨文字形异体繁多，且因刻手、时代等的不同，也不断发生着变化。例如大家公认的"裸"字的写法甲骨文中就有  等多种写法。如果直接将某些形体通过增加条件的形式归并在一起的话，则势必增加人工智能识别的主观性。因此，我们宁愿麻烦一点儿，也要保证人工智能识别的客观，使其成为有普遍适用性的原始材料。

经过本部分的工作，将形成一个以字形为主体的基础字形库，这个字形库的数量据《甲骨文字形表》和《甲骨文字导航》估算，约有 8000 个至 13000 个文件夹。这些字形将是我们下一步工作的基础，也可以成为学术公器，作为各家考释文字、进行字形观察、异体合并的基础。

（三）甲骨原文释文的自动生成

在步骤 1 中，我们标注的甲骨片，主要是做了框出甲骨片上的文字，以及标出对应的原篆矢量字。这些矢量字使用的主要是陈年福先生所做的浙师大甲骨文字体文件，字体文件中没有的字形或不准确的字形，我们拟进行不规则数据的均衡化，并结合陈年福先生的字库生成模拟的字形图像，即根据原形再造新的矢量字。甲骨文字体库的主要用途是输入、在计算机终端显示甲骨文，但字库并不具备智能识别拓本图像中甲骨文的功能。在第一步和第四步的工作之后，我们会以字形为单位训练文字识别模型，利用深度学习技术，给未标注的甲骨文字自动匹配相对应的矢量字，并自动生成出未标注甲骨片上的矢量字释文。

释文的重要性在于可以给出字与字之间的搭配关系，即提供语境的信息。在释读甲骨文字的过程中，不同或相近的字形在甲骨文中的用法和意义是否有区别，往往要在语境之下进行判断，而不能仅仅依靠字形的相似与否。字形和语境，是辨别和梳理甲骨文字的基础。人工智能识别甲骨文的基础是字形，在区分和关联字形的基础上，充分利用计算机的语法分析功能，为人工智能识别甲骨文提供更多的识别角度。

对于样本量较小的字形，我们拟使用仿射变换、增加噪声等技术，最终实现对原篆字图像中文字的自动识别。

最后，我们将利用上述研究形成的数据、算法和技术，研制对应的甲骨文识别与文字检索软件，为广大甲骨学爱好者提供便捷的服务。

四　人工智能甲骨文数据库的展望与应用

利用人工智能构建原篆字形数据库，具有自动、智能、快速的特点，有望从根本上解决甲骨学者穷尽性搜集甲骨字形原始资料的迫切需求，为甲骨文研究节约大量搜集整理资料的时间。而且，本课题形成的算法和技术易推广复用到其他古文字如金文、陶文、盟书、玺印、简帛文字等，能够直观呈现某个文字在不同时期的构形变化和演变过程，对未识字的考释将有积极的推动作用。

上述甲骨文原篆字形数据库、甲骨文自动定位和识别技术，有利于甲骨学者对甲骨文拓本原始资料的智能检索、归类和识别，快速找到所需资料并自动归类，大大减少相关的人工操作，极大提高学者们的工作效率。另外，依托本文中的原篆字形数据库和相关人工智能技术，可以开发相关的手机端 App 产品，如利用扫一扫等操作实现对甲骨文的快速定位和识别，这些应用对于甲骨文学者的科学研

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

究，对于甲骨文的科普、甲骨文爱好者的学习，必有较大的帮助。

附记：人工智能甲骨文字识别在实验阶段，河南大学黄河文明与可持续发展研究中心及河南大学文学院、计算机与信息工程学院2016级、2017级十多位研究生进行了辛苦而认真的标注工作，谨表谢忱！

【参考文献】

[1] Bai Fan & Cheng Zhanzhan & Niu Yi & Pu Shiliang & Zhou Shuigeng. *Edit Probability for Scene Text Recognition*. CVPR, 2018.
[2] 陈梦家.殷墟卜辞综述[M].北京：中华书局，1988.
[3] 陈年福.甲骨文字新编[M].北京：线装书局，2017.
[4] 韩江苏.殷墟甲骨文编[M].北京：中国社会科学出版社，2017.
[5] Li Hui & Wang Peng & Shen Chunhua & Zhang Guyu. *Show, attend and read: A simple and strong baseline for recognising irregular text*. AAAI, 2019.
[6] 李守奎.关于古文字字编体例的一些思考[J].华夏文化论坛(第一辑)，2006.
[7] 李宗焜编著.甲骨文字编[M].北京：中华书局，2012.
[8] 刘钊主编.新甲骨文编[M].福州：福建人民出版社，2009.
[9] 沈建华，曹锦炎编著.甲骨文字形表[M].上海：上海辞书出版社，2008.
[10] Shi, Baoguang & Bai, Xiang & Yao, Cong. *An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition*. IEEE Trans. PAMI, 2017.
[11] 孙海波.甲骨文编[M].北京：哈佛燕京学社，1934.
[12] 唐兰.古文字学导论(增订本)[M].济南：齐鲁书社，1981.
[13] 王襄.簠室殷契类纂[M].天津：河北第一博物院，1920.
[14] 王蕴智主编.甲骨文可释字形总表[M].郑州：河南美术出版社，2017.
[15] 姚孝遂主编.殷墟甲骨刻辞类纂[M].北京：中华书局，1989.
[16] 赵诚.甲骨文字学纲要[M].北京：中华书局，2005.
[17] Zhou Xinyu & Yao Cong & Wen He & Wang Yuzhi & Zhou Shuchang & He Weiran & Liang Jiajun. *EAST: An Efficient and Accurate Scene Text Detector*. CVPR, 2017.

AI-driven Oracle Bone Inscription Recognition and Glyph Database Development

Men Yi Zhang Chongsheng

(Key Research Institute of Yellow River Civilization and Sustainable Development, HeNan University, KaiFeng 475000, China; School of Computer and Information Engineering, HeNan University, KaiFeng 475000, China)

Abstract: Researchers have achieved significant progress in collecting and investigating the oracle bone inscriptions (OBIs), and the research outcomes were usually published and spread in form of reference books. The community has accumulated lots of experiences in the arrangement, retrieval and representation of OBIs. Yet, there still exists many other problems, such as the inconsistency in the separation and union of the character patterns, the difficulty in retrieving OBIs, and the discrepancy in 'liDing' (vectoring-drawing) the OBI characters/glyphs. Reference books are easy to

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

be disseminated, but the results summarized in the books are hard to be reused and improved; moreover, due to the page size limit, these books cannot include the complete material and related contextual information of the OBIs. To this end, in the work we adopt the new generation of AI techniques for OBI recognition and database development. In specific, we will train deep learning models for text detection and recognition over the OBI rubbings. Upon this, we will design algorithms to automatically build the glyph database with all the materials and context information, using unsupervised or semi-supervised image similarity computation methodology. Finally, we will also develop a software to facilitate the recognition and retrieval of the OBI characters and the related materials.

Key words: oracle bone inscriptions; character patterns; AI; database

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net