基于深度学习的甲骨文字检测与识别

刘国英

(安阳师范学院 甲骨文信息处理教育部重点实验室,河南 安阳 455000)

摘要:甲骨文是汉字的源头和中华优秀传统文化的根脉。甲骨文研究对中华优秀传统文化的继承和发扬意义重大。然而,甲骨文的识读需要具备古文字学、考古学、历史学和文献学等多学科知识背景,普通大众参与难度极大。利用人工智能和模式识别技术进行甲骨文字的检测和识别,在一定程度上能够缓解这一问题。本文综述了人工智能和模式识别技术在甲骨文字检测和识别领域中的研究进展。首先,介绍了甲骨文字检测和识别的研究背景。其次,从传统方法的应用到深度学习技术的尝试等角度,分别详细介绍了甲骨文字识别和甲骨文字检测的研究进展,阐述了相关方法的技术细节、使用的数据集信息、以及基本性能。然后,从数据特性的角度出发,分析了当前甲骨文字检测和识别技术存在的不足,并阐述了常用数据集存在的问题。最后,对全文进行了总结,并指出了甲骨文字检测和识别领域仍需要解决的问题及未来的研究方向。

关键词:甲骨文;人工智能;模式识别;检测与识别

中图分类号:TP391.1;K877.1　　文献标识码:A　　文章编号:1001-0238(2020)03-0054-06

DOI:10.16140/j.cnki.ydxk.2020.03.011

一、前言

甲骨文是迄今为止中国发现的年代最早的成熟文字系统,是汉字的源头和中华优秀传统文化的根脉。2017 年 10 月 30 日,甲骨文入选联合国教科文组织“世界记忆名录”,其研究价值得到全世界公认。然而,甲骨学一直是一个没有多少人参与的冷门学科。主要原因有两个方面。首先,甲骨学研究门槛高,必需具备古文字学、考古学、历史学、文献学等多学科知识背景,高水平人才匮乏。其次,甲骨文认读难度大而难以被大众接受。目前,甲骨文仅仅停留在少数甲骨学专家的学术研究中,普通大众没有热情,社会上没有培育出相应的土壤,后备人才匮乏。这一问题已经引起党和国家的高度重视。2016 年 5 月习近平在全国哲学社会工作座谈会上指出要重视发展诸如“甲骨文等古文字研究”。在 2019 年 11 月 1 日,他在祝贺甲骨文发现和研究 120 周年的贺信中又进一步强调“在新形势下,要确保甲骨文等古文字研究有人做、有传承”。

在政府的大力倡导下,研究人员开始尝试利用计算机技术进行甲骨文字的检测与识别研究,以期在降低普通大众认读甲骨字难度的同时为甲骨学专家提供辅助工具。早期的甲骨文字检测和识别研究基本按照传统模式识别路线开展,特征提取和分类器设计等多个环节都需要人工参与。这对研究人员的经验要求较高,且流程较为复杂,应用也不够方便。近年来,深度学习技术尤其是自然场景目标检测和识别技术的快速发展为甲骨

[收稿日期] 2020-06-01

[基金项目] NSFC-河南省联合基金:基于深度学习的甲骨文字检测与识别研究(U1804153);NSFC-青年基金:语境和构形网络驱动的未识甲骨字场景语义预测(61806007)。

[作者简介] 刘国英(1979-),男,河南郑州人,安阳师范学院计算机与信息工程学院教授,主要从事数字图像处理研究。

文字检测和识别注入了新的活力。研究人员逐渐开始尝试使用深度神经网络进行甲骨文字的检测和识别研究,提出了一些新的方法,建立了多个甲骨文字检测和识别数据集。对这些研究内容进行详细梳理,将有助于加速甲骨文字检测和识别技术的发展,为甲骨学研究和甲骨文活化利用提供支撑。

本文组织结构如下: 第 2 节介绍甲骨文字识别的研究现状; 第 3 节介绍甲骨文字检测的研究现状; 第 4 节具体分析当前研究的困难; 第 5 节对全文进行总结并进行展望。

二、甲骨文字识别

1. 传统甲骨文字识别技术

早期的甲骨文字识别遵循了“数据预处理 + 特征提取 + 分类识别”的传统甲骨文字识别框架。复旦大学的周新伦和李锋等[1-2]将甲骨文字视为“线”和“点”构成的无向图,并基于图论方法提取多层次图特征,从而进行二级和三级甲骨文字识别。北京大学吕肖庆等人[3]将甲骨文字视为图形符号,并利用轮廓线曲率直方图获得的傅里叶描述子表示甲骨字形状,使用支持向量机(support vector machines ,SVM) [4]进行甲骨字识别。安阳师范学院栗青生教授[5]将甲骨文字抽象为无向图,并基于图同构判定算法进行甲骨字识别。刘永革等[6]利用甲骨字的分块直方图表示文字特征、基于支撑向量机进行甲骨字识别。江苏师范大学顾绍通[7]假设甲骨文异体字之间的拓扑结构具有不变特性,并利用最小距离对甲骨字拓扑结构编码进行等价关系判断,进而获得甲骨文字的识别结果。日本立命馆大学(Ritsumeikan University) L. Meng[8](p8-10)针对待识别甲骨文字图像和模板甲骨图像,分别使用霍夫变换和聚类来提取的直线特征点,并利用对应的最小距离进行一级识别,再使用模板匹配方法进行识别结果优化。这些方法主要针对甲骨文字的字形特征进行研究,并取得了有意义的结果。上述方法基本在较小的数据集上进行甲骨文字识别尝试,方法的稳定性、准确率和泛化能力均难以满足要求。

2. 基于深度学习的甲骨文字识别

目前,基于深度学习的甲骨文字识别都是监督方式的。它们需要大量的训练数据使深度神经网络学习甲骨单字的不同模式,从而实现对单字图像的自动识别。如图 1 所示,常见的甲骨单字图像主要有两类: 甲骨字模图像和拓片文字图像。

最早将深度学习技术应用于甲骨字模图像识别的代表性工作是 2016 年郭俊等[11]提出的多层次甲骨字符表示方法。该方法将基于稀疏自编码的中层表示特征和基于 Gabor 的低层表示特征结合起来描述甲骨字符,采用的甲骨字模数据集包含 291 个甲骨文单字,共有 20039 个样本。2018 年,Zhen Yang 等人[12]基于 LeNet 和 AlexNet 在一个含有 21373 个样本、涉及 39 个单字的甲骨字模数据集上进行了识别研究。同年,我们设计了一个简单的深度识别网络,并在包含 44868 个样本、共计 5491 个单字(含异体字) 的甲骨字模数据集上进行了验证[13]。甲骨字模数据集通过甲骨学专家手工描摹生成,不含任何噪声,上述识别方法均能达到 90% 以上的识别准确率。

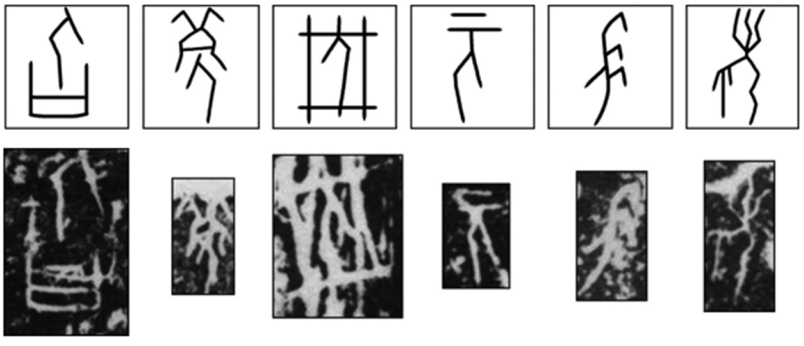

图 1　甲骨字模图像与拓片文字图像: 第一行为字模图像,第二行为对应拓片文字图像。

然而，甲骨字模图像与拓片文字图像有较大区别，后者噪声影响极其严重，识别难度要大得多。因此，真正意义的甲骨文字识别工作需要在拓片文字数据集上进行。2019 年，我校“甲骨文信息处理”教育部重点实验室和华南理工大学金连文教授团队[14]联合推出了目前标注样本最多的甲骨拓片文字数据集 OBC306，包含 306 个甲骨文单字，共计 309551 个样本，并通过 Inceptiion - v4、ResNet - 101、ResNet - 50、VGG16、AlexNet 等典型识别网络提供了基准识别率。河南大学王慧慧[15]根据不同甲骨字的样本分布构造了一组更为精细的拓片文字数据集，并利用非稀疏表示、深度学习和稀疏表示三类方法进行了识别实验。甲骨磨损、拓印噪声和甲骨纹理干扰对识别结果影响很大，从公开报道结果来看，整体识别率都不理想，最高的仅为 70% 左右。

三、甲骨文字检测

1. 基于传统方法的甲骨文字检测

基于传统方法进行甲骨文字检测的方法较少。安阳师范学院史小松老师进行了一定的尝试[16]。她使用基于连通分量的方法进行了甲骨字检测的探索，从一定程度上解决了纯手工提取甲骨字的误差和效率问题，并可作为进一步分析定位并识别甲骨字的基础。但该方法在复杂背景或噪声比较严重的拓片图像中，效果不太理想，存在拓片上部分残字或笔画定位不准确的现象。

2. 基于深度学习的甲骨文字检测

随着深度学习技术的发展，尤其是自然场景图像中目标检测技术的快速发展，促使研究人员逐渐开始利用深度学习技术进行甲骨文字的检测尝试。目前，基于深度学习的甲骨文字检测也都是监督方式的。它利用大量已标注甲骨文字位置信息的训练数据，采用一阶段或者两阶段的方式训练深度神经网络，进而实现拓片图像中甲骨文字的自动标注。

华南理工大学黄双萍教授团队[17]构造了一个甲骨文字检测数据集 OBCD，标注了 5838 幅甲骨拓片图像。他们结合使用基于区域的全卷积神经网络 R - FCN(Region based Fully Convolutional Network)[18]和特征金字塔模型 FPN(Feature Pyramid Networks)[19]进行甲骨字检测研究。日本立命馆大学 Lin Meng[20]构造了一个包含 330 幅图像的甲骨文字检测数据集，改进了 SSD[21]以提高较小字体甲骨字的检测准确率。我们在甲骨文字检测方面做了更为充分的基础性研究工作[22]。首先，构造了一个包含 9500 张图像的甲骨文字检测数据集。基于该数据集，分析了近几年有代表性的通用目标检测框架(包括: Faster R - CNN[23]、SSD[21]、RefineDet[24]、RFBnet[25]和 YOLOv3[26])的甲骨文字检测性能，并对性能最优的 YOLOv3 进行了改进。上述研究在字符分布均匀、大小相似、干扰较少的甲骨片上均表现良好。然而如图 2 所示，字符粘连、背景干扰、残缺字和重叠检测等问题使得一些甲骨字的准确检测十分困难。

图 2 甲骨字检测面临的主要困难

四、甲骨文字检测和识别研究存在的问题

1. 检测和识别方法存在的问题

与自然场景图像相比，甲骨图像数据有其自身的特点，具体来说，:

① 甲骨字的背景噪声严重。甲骨片在地下埋藏 3000 多年，经过长期的腐蚀变化，上面的文字变得模糊不清，获取的甲骨图像往往存在非常严重的背景噪声，拓印出来的拓片也常常出现严重的干扰拓痕。

② 残缺甲骨字较多。甲骨片在出土时很容易碎裂,甲骨字附近经常会出现断裂,因此产生大量的甲骨残字,这些残字与甲骨天然纹理非常相似,其检测与识别困难极大。

③ 甲骨字具有极强的不规则性。同一甲骨片上的甲骨字大小不一、方向随意、分布分散,极大增加了检测与识别的难度。

④ 甲骨文异体字出现频繁,但具有部分相似性。甲骨字出现在殷商时期,当时并没有统一的文字规范,再加上商周时期(约公元前 1600—公元前 256 年)跨度约 1300 多年,文字的演化比较明显,导致甲骨文中出现大量的异体字。很多异体字之间存在部分的相似性,这为异体字的识别提供了新的线索。

⑤ 甲骨字出现频率严重不平衡。在文献[28]研究的 56743 个甲骨字样本中,包含 1425 个单字。其中,常用字 366 个,次常用字 500 个,罕用字 559 个。甲骨字的字频结构极不平衡,为深度学习技术的使用带来困难。

然而,现有的检测和识别方法仅仅简单地将自然场景领域中的深度神经网络模型迁移到甲骨文字的检测和识别应用中来,并没有考虑甲骨图像数据自身的特点。这是现有方法不能取得令人满意的结果的主要原因。

2. 训练数据集存在的问题

训练数据集对监督方式的甲骨文字检测和识别至关重要。然而因为主观和客观的原因,现有的训练数据集仍存在一些较为严重的问题。具体来说:

① 训练数据集规模不大,难以保证深度神经网络的学习性能。检测数据集主要来源于《甲骨文合集》[27]的部分扫描数据,最大的数据集仅含 9500 多幅图像,数据多样性得不到保证。识别数据集基本局限于样本较多的几百个甲骨单字,覆盖不了已释的 1400 个左右甲骨单字,更覆盖不了所有的 4500 个左右甲骨单字。

② 训练数据集标注质量不高,严重影响深度神经网络的学习质量。甲骨学研究人员极少,现有训练数据集基本依靠非甲骨学领域人员依据工具书进行标注,因甲骨学知识欠缺使得残缺甲骨字、高噪干扰甲骨字、相近字形等难以标注,数据集质量不高。

③ 训练数据标注标准不统一,严重影响深度学习的甲骨学实际应用。受标注人员个人素质影响,训练数据标注的标准可能有所差别。没有释读的甲骨字在进行类别划分时没有统一标准,甚至存在巨大争议。考虑到甲骨学研究的实际水平和甲骨学专家稀缺的客观现实,短期内难以获取大规模、高质量的训练数据集。

表 1　用以深度学习的标注数据统计表

名称或出处	类型	规模	用途	说明
郭俊[9]	字模	单字 291、总数 20039	识别	
Zhen Yang[10]	字模	单字 39、总数 21373	识别	
刘国英[11]	字模	单字 5491、总数 44868	识别	含异体字
OBC306[12]	拓片文字	单字 306、总数 309551	识别	
张重生[13]	拓片文字	单字 469、总数 194804 字形 937、总数 155462 均衡 937、总数 85281	识别	均衡保证样本数目最多的不超过最少的二倍
OBCD[14]	拓片图像	5838 幅	检测	
Lin Meng[17]	拓片图像	330 幅	检测	
刑济慈、刘国英*[19]	拓片图像	9500 幅	检测	

四、总结、讨论与展望

本文简要回顾了甲骨文字检测与识别的研究背景,分别介绍了传统甲骨文字检测与识别的主要成果,指出了存在的问题与缺陷,并介绍了基于

深度神经网络的甲骨文字检测与识别研究现状，并分析了当前研究所面临的主要困难。

甲骨文字检测与识别是近年来才被学者关注的一个研究课题。从最初利用传统方法进行检测与识别，到当前深度学习技术在甲骨文字检测与识别上的应用尝试，在检测和识别的性能上均有明显提高。然而，仍有许多问题亟待解决:

(1) 噪声干扰问题。甲骨拓片图像的噪声与常见的高斯噪声、椒盐噪声等不同，难以用常规技术进行建模。从形态上看，这些噪声与甲骨字刻痕存在一定程度的相似性，对甲骨文字的检测与识别极为不利，严重影响了检测和识别的精度。

(2) 残缺甲骨字的检测与识别。甲骨字经常因甲骨片断裂而导致残缺，致使其字形特征与甲骨天然纹理极其相似，再加之残缺甲骨字训练样本量极少，其检测与识别困难非常大。

(3) 超大类别的甲骨文字识别问题。已知的甲骨单字有 4500 多个，对甲骨单字图像进行识别必须产生 4500 个左右的类别。而现有甲骨文字识别方法难以对甲骨字进行全类别识别，主要原因是部分甲骨字样本极少，有的甚至只有几个样本。这极大增加了识别难度。

(4) 甲骨文异体字的识别问题。甲骨文中严格意义上的异体字有 1032 组，其字形总数为 3085 个，占到了甲骨文字形总数的 49.5%[28]。甲骨文异体字出现非常频繁，同一甲骨字的不同异体字之间字形相差极大，对应的异体字识别非常困难。

(5) 检测和识别数据集的严重依赖问题。训练数据的束缚使得监督方式的深度学习难以发挥自身优势而陷入困境。事实上，通过数字化设备可以很容易获取大规模甲骨拓片图像数据。如果让深度神经网络学习数据自身的特性而不是学习难以获取的监督信息，则更有利于发挥深度神经网络强大的学习能力。

(6) 甲骨字构件的检测与识别问题。甲骨字具有明显的构件信息。甲骨字中二级构件有 291 个、三级构件有 61 个，频率不为 1 的基础构件为 497 个，这些构件通过不同方式构成甲骨字[28]。构件的识别能够为甲骨字的识别提供有用信息。然而，甲骨字构件之间空间关系复杂，包围关系、嵌套关系等对甲骨字构件的自动分析技术提出挑战。

总之，甲骨文字检测与识别在近几年来取得了一定的研究成果，但仍然有大量问题值得深入研究。本文通过对有关研究进展的回顾、分析和讨论，以期为有兴趣从事该项研究的研究人员提供全面的信息和研究思路，为早日甲骨文字检测与识别研究的实用化贡献力量。

[参考文献]

[1] 周新伦，李锋，华星城等. 甲骨文计算机识别方法研究 [J]. 复旦学报(自然科学版)，1996，(5)：481 - 486.

[2] 李锋，周新伦. 甲骨文自动识别的图论方法 [J]. 电子科学学刊，1996，(S1)：41 - 47.

[3] 吕肖庆，李沫楠，蔡凯伟等. 一种基于图形识别的甲骨文分类方法 [J]. 北京信息科技大学学报，2010，(Z2)：92 - 96.

[4] Vapnik V N，Lerner A Y. Recognition of patterns with help of generalized portraits [J]. Avtomat. i Telemekh，1963，(6)：774 - 780.

[5] 栗青生，杨玉星，王爱民. 甲骨文识别的图同构方法 [J]. 计算机工程与应用，2011，(8)：112 - 114.

[6] 刘永革，刘国英. 基于 SVM 的甲骨文字识别 [J]. 安阳师范学院学报，2017，(2)：54 - 56.

[7] 顾绍通. 基于拓扑配准的甲骨文字形识别方法 [J]. 计算机与数字工程，2016，(10)：2001 - 2006.

[8] L. Meng. Two - Stage Recognition for Oracle Bone Inscriptions. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)，vol. 10485 LNCS，pp. 672 - 682，2017.

[9] L. Meng. Recognition of Oracle Bone Inscriptions by Extracting Line Features on Image Processing [C]. ICPRAM，2017，pp. 606 - 611.

[10] L. Meng and T. Izumi. A combined recognition system for oracle bone inscriptions [J]. Int. J. Mechatron. Syst.，vol. 7，no. 4，pp. 235 - 244，2017.

[11] Jun Guo，Changhu Wang，Edgar Roman - Rangel，Hongyang Chao，and Yong Rui. Building Hierarchical Representations for Oracle Character and Sketch Recognition [J]. IEEE Transactions on Image Processing，2016，25(1)：104 - 118.

[12] Zhen Yang，Qiqi Wang，Xiuying He，Yang Liu，Fan

Yang, Zhijian Yin, and Chen Yao. Accurate Oracle Classification Based on Deep Convolutional Neural Network [C] // 2018 IEEE 18th International Conference on Communication Technology (ICCT), October 8 – 11, 2018, Chongqing, China. 2018: 1188 – 1191.

[13] Guoying Liu and Feng Gao. Oracle – Bone Inscription Recognition Based on Deep Convolutional Neural Network [J]. Journal of Computers, 2018, 13(12): 1442 – 1450.

[14] Shuangping Huang, Haobin Wang, Yongge Liu, Xiaosong Shi, and Lianwen Jin. OBC306: A Large – Scale Oracle Bone Character Recognition Dataset [C] // 2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20 – 25, 2019. IEEE, 2019: 681 – 688.

[15] 王慧慧. 大规模甲骨文数据集构建及算法研究 [D]. 开封: 河南大学, 2019.

[16] Shi Xiao song, Huang Yong jie, Liu Yong ge. Text on Oracle rubbing segmentation method based on connected domain [C] // 2016 IEEE Advanced information Management, Communicates, Electronic and Automation Control Conference, IMCEC 2016, Xian, P. R. China, 2016 . 10. 03 – 10. 05.

[17] 王浩彬. 基于深度学习的甲骨文检测与识别研究 [D]. 广州: 华南理工大学, 2019.

[18] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R – FCN: Object Detection via Region – based Fully Convolutional Networks [C] // Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5 – 10, 2016, Barcelona, Spain. 2016: 379 – 387.

[19] Tsung – Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21 – 26, 2017. IEEE Computer Society, 2017: 936 – 944.

[20] Lin Meng, BingLyu, Zhiyu Zhang, C. V. Aravinda, Naoto Kamitoku, and Katsuhiro Yamazaki. Oracle Bone Inscription Detector Based on SSD [C] // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol 11808 LNCS. [S. l.]: Springer International Publishing, 2019: 126 – 136.

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng – Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector [C] // Computer Vision – ECCV 2016 – 14th European Conference, Amsterdam, The Netherlands, October 11 – 14, 2016, Proceedings, Part I: Vol 9905. Springer, 2016: 21 – 37.

[22] Jici Xing, Guoying Liu, and Jing Xiong. Oracle Bone Inscription Detection: A Survey of Oracle Bone Inscription Detection Based on Deep Learning Algorithm [C] // Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing, AIIPCC 2019, Sanya, China, December 19 – 21, 2019. ACM, 2019: 39: 1 – 39: 8.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R – CNN: Towards Real – Time Object Detection with Region Proposal Networks [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2017, 39(6): 1137 – 1149.

[24] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single – Shot Refinement Neural Network for Object Detection [C] // 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18 – 22, 2018. IEEE Computer Society, 2018: 4203 – 4212.

[25] Songtao Liu, Di Huang, and Yunhong Wang. Receptive Field Block Net for Accurate and Fast Object Detection [C] //FERRARI V, HEBERT M, SMINCHISESCU C, et al. Lecture Notes in Computer Science, Vol 11215: Computer Vision – ECCV 2018 – 15th European Conference, Munich, Germany, September 8 – 14, 2018, Proceedings, Part XI. Springer, 2018: 404 – 419.

[26] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement [J]. CoRR, 2018, abs/1804. 02767.

[27] 郭沫若, 胡厚宣. 甲骨文合集 [M]. 北京: 中华书局, 1982.

[28] 陈婷珠. 殷商甲骨文字形系统再研究 [D]. 上海: 华东师范大学, 2007.

责任编辑: 郭　昱　胡洪琼