

汉字识别技术的研究进展分析

李阳娟

(莆田学院 工程实训中心, 福建 莆田 351100)

摘要:为全面深入地掌握汉字识别技术的研究现状与应用进展,基于 WOS 数据库中汉字识别研究领域的文献数据,利用 CiteSpace 可视化软件进行文献计量与知识图谱分析。结果表明:中国的高校和研究所是汉字识别研究的核心力量;汉字识别技术的发展历程由早期的传统方法 OCR 技术演化为中期的联机手写汉字识别,再到近期基于深度学习的汉字识别研究;研究前沿为手写汉字识别、混合语言模型、生成对抗网络和场景文本识别;汉字识别技术将朝着高效率及智能化方向发展。

关键词:汉字识别;光学字符识别;深度学习;手写体汉字

中图分类号:TP391 **文献标志码:**A **文章编号:**1671-1807(2022)04-0180-08

光学字符识别(OCR)技术指对图像中的文字进行检测并识别出文字内容。OCR 是人工智能和计算机视觉领域的热点研究方向。2020 年 4 月,国家发改委明确将人工智能等新技术基础设施列入“新基建”范围,从政策布局上加速了人工智能产业的不断成熟。随着人工智能的飞速发展以及中文汉字的广泛使用,汉字识别技术有着重要的研究价值和社会意义,在金融、政府、医疗和教育等多个领域有极其广阔的应用前景,如车牌识别、机器翻译、卡证票据识别和无人驾驶系统等^[1-5]。

由于汉字本身字符量多、字形结构复杂且有较多的相似字等特点,汉字识别的研究仍存在着许多的困难与不足之处^[6]。因此,充分了解近年来汉字识别领域的研究现状及发展趋势对相关领域的研究者有一定的参考价值。本文以 2001—2021 年 Web of Science(WOS)数据库中汉字识别的文献数据作为研究对象,利用文本可视化软件 CiteSpace^[7]绘制知识图谱并对文献内容进行分析,从量化视角挖掘汉字识别领域近几年的研究现状、热点和研究前沿,为汉字识别的研究与发展提供参考。

1 数据来源及研究方法

选取 WOS 核心数据库中 2001—2021 年汉字识别领域的相关文献作为文献数据来源,采用的检索式为 TS=(Chinese character recognition OR Chinese text recognition) AND SU=(Computer),文献类型为 Article 或 Review 或 Proceedings Pa-

per,排除掉非计算机视觉领域的研究内容后获得 1 119 篇文献,检索时间为 2021 年 9 月 20 日。CVPR、ICCV 和 ECCV 这 3 个权威的计算机视觉领域的国际会议收录了该领域最新且重大的研究成果,因此会议论文也是文献数据的重要来源。

CiteSpace 软件主要基于共引分析理论,可以绘制和可视化科学知识领域的知识结构,探寻学科领域演化关键路径及知识转折点^[8]。本文利用 CiteSpace 5.8. R1 对采集到的文献数据进行文献计量分析,并对机构与作者、关键词和共被引文献进行可视化知识图谱分析,以全面深入地掌握汉字识别领域的研究进展情况与前沿发展趋势。

2 汉字识别领域的发文趋势及研究力量分析

2.1 发文趋势分析

对 2001—2021 年汉字识别领域的发文情况进行统计分析,如图 1 所示。汉字识别领域在这些年的发展过程中出现了两个明显的增长高峰。第一个高峰出现在 2008 年,文献数量在 2008 年以前缓慢增长,之后呈现下降趋势。第二个增长高峰在 2017 年,文献数量从 2012 年急速增长直到 2017 年,之后维持在较高的水平。通过文献调研可知其主要原因:第一个增长时期是由于智能手机等设备的流行,汉字识别领域迎来了应用高潮;第二个增长时期则由于卷积神经网络的发展带来了技术的革

收稿日期:2021-12-09

基金项目:福建省教育厅中青年科技项目(JAT200515);莆田学院校内科研项目(2019030)。

作者简介:李阳娟(1987—),女,福建莆田人,莆田学院工程实训中心,实验师,硕士,研究方向为计算机应用。

新,在汉字识别领域引发了一轮新的发展热潮。

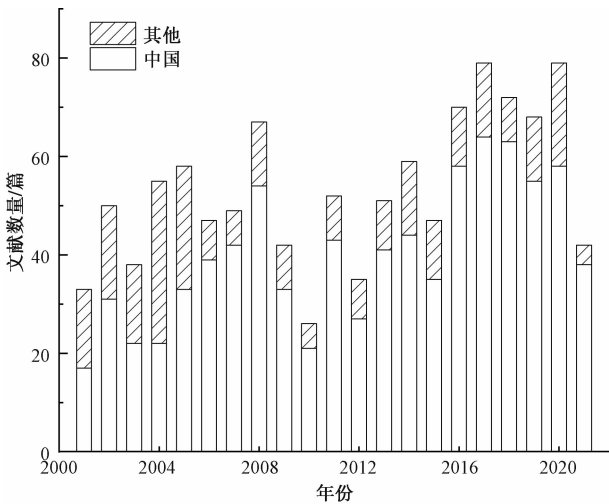


图 1 汉字识别研究领域的发文趋势

2.2 研究力量分析

对研究机构及作者的发文情况和合作关系进行分析,可以得到汉字识别的研究力量分布,有助于发现研究汉字识别的权威机构及有影响力的作者。从汉字识别研究领域的国家或地区的贡献来看,使用汉字的国家或地区成为汉字识别领域的研究主力,如图 1 所示,中国发表的论文数量占总论文的 75% 以上,在汉字识别研究领域起主导作用。

表 1 列出了发文量前 10 位的研究机构,这些机构共发文 398 篇,占全部文献量的 32%,说明了汉字识别领域的研究力量比较集中,绝大多数是中国的高校和研究所。中国是汉字使用大国,同时也是汉字识别研究的核心力量。中国科学院以 127 篇的发文量位居榜首,明显高于其他研究机构,清华大学和华南理工大学紧随其后,这 3 所院校及研究机构是当前汉字识别研究领域的重要基地。利用 CiteSpace 软件绘制各个机构间的合作网络图谱,如图 2 所示。图中每个节点代表着相对应的机构,节点越大表示发文量越多,节点间的连线代表机构间的合作关系,连线的粗细与合作次数成正比。图中有 3 个明显的大的节点:中国科学院、清华大学和 华南理工大学,与统计结果相符。中介中心性是衡量节点在网络中重要性的一个指标,数值越大表明节点在网络中越重要^[9]。在图 2 中,中国科学院的中介中心性的数值达到 0.15,且节点处有圆圈标识,说明了该节点在图谱中的重要性高,与其他机构有着较强的合作关系。

作者合作图谱显示了汉字识别研究领域所有作者的发文数量与合作关系。在 CiteSpace 软件中

表 1 发文量前 10 位的研究机构

机构	发文量/篇	中心性
中国科学院	127	0.15
清华大学	73	0.06
华南理工大学	66	0.03
哈尔滨工业大学	31	0.03
中国科技大学	23	0
香港理工大学	17	0.03
华中科技大学	16	0.02
北京邮电大学	15	0.01
河北大学	15	0
新加坡国立大学	15	0.04

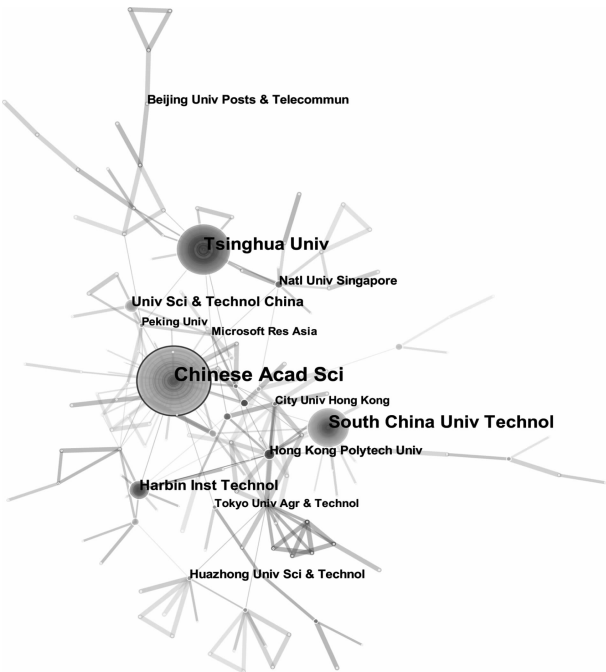


图 2 研究机构合作网络图谱

绘制作者合作网络图谱,如图 3 所示。图中有 3 个较大的节点:刘成林(CHENGLIN LIU)、金连文(LIANWEN JIN)和丁晓青(XIAOQING DING),节点越大表明发文量越多。围绕这 3 个高产的核心作者形成了 3 个较大的作者合作网络群,分别是中科院的刘成林教授团队、华南理工大学的金连文教授团队和清华大学的丁晓青教授团队。刘成林教授团队主要研究文档图像分析、模式识别与机器学习,该团队发布了大型手写汉字数据库 CASIA-OLHWDB 和 CASIA-HWDB,包含了孤立的手写字符和连续的文本,可用于各种手写文档分析任务的研究,促进了汉字识别领域的学术研究和技术研发^[10]。金连文教授团队主要的研究方向为基于深度学习的文档图像分析与理解,所研发的联机手写识别技术已在搜狗等知名公司获得了规模化应用。丁晓青教授团队主要研究智能图文信息处理与图

像分析,创造性地将信息熵理论引入模式识别,从而解决汉字识别的众多问题,形成的 TH-OCR 系列文字识别软件产品广泛应用于国内外多家知名公司。这些研究团队的合作关系显著、取得成果较多且研究力量较强。还有一些其他的作者合作网络,但是发量较少,形成的合作网络较小。从作者合作网络图谱的整体上看,有较强合作关系的作者发量较多。



图 3 作者合作网络图谱

3 汉字识别领域的关键词分析

关键词是对文章主题内容的精炼和概况,文献数据中的关键词出现的频次越高,说明该主题的研究热度越高。引入 CiteSpace 软件进行关键词共现知识图谱分析,如图 4 所示。图 4 中频次越高的关键词节点半径越大,研究热度也越高,节点间的连线代表了各关键词间的关联强度。图中最大的几个节点为识别、字符识别、特征提取、联机 and 分割。

表 2 列出了关键词频次前 20 位的高频关键词。可以看出大部分的高频关键词在 2003 年之前出现,如:特征提取、字符分割、字符识别、光学字符识别(OCR)、手写汉字识别、模式识别等。此阶段的研究重点在利用模式识别的模型算法来进行汉字识别及手写汉字识别。2005 年新增高频关键词:联机。2009 年新增高频词:卷积神经网络(CNN)、手写识别。2015 年新增高频词:深度学习。这些高频关键词及其出现时间反映了汉字识别领域的热点演变趋势。将高频关键词所涉及的研究热点总结

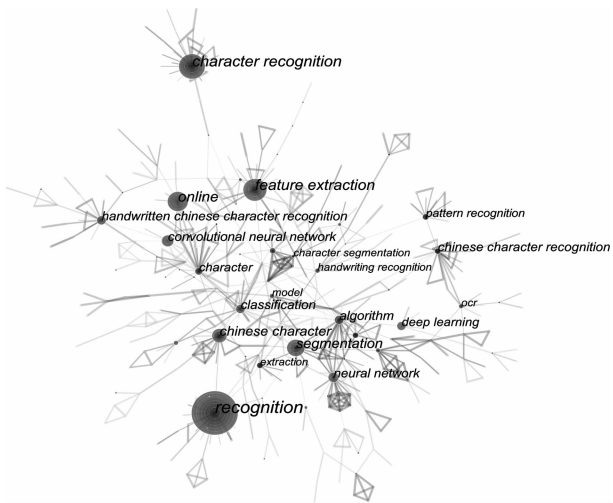


图 4 关键词共现图谱

为以下 3 点:基于传统方法的 OCR 技术、基于深度学习的 OCR 技术和手写汉字识别。

1)基于传统方法的 OCR 技术。汉字识别是模式识别的一个重要的研究领域。汉字识别通常采用传统的 OCR 技术流程:先预处理目标图片,再对文字区域进行字符切分,对单字符进行特征提取,最后利用分类器分类并识别出文字。其中特征提取和分类识别是汉字识别中的重点和难点。Liu 等^[11]利用传统的统计模式识别方法进行单字符手写汉字识别,在联机手写数据库 OLHWDB 和脱机手写数据库 HWDB 上获得最高的准确率为 93.95%和 90.71%。传统的 OCR 技术在单字符的汉字识别中识别精度较高,但对于复杂背景等情况的文字识别精度较低。

2)基于深度学习的 OCR 技术。卷积神经网络(CNN)是典型的深度学习网络结构。深度学习以神经网络为主要模型,无须手动提取特征,由神经网络自动提取特征并识别出结果。在 2012 年的 ImageNet 图像识别大赛上,Krizhevsky 等^[12]设计了深度卷积神经网络 AlexNet,并以较大的优势赢得了冠军。自此以后,CNN 网络发展非常迅速,相继产生了许多著名的网络框架,如 VGGNet^[13]、GoogleNet^[14]和 ResNet^[15]等,卷积神经网络在图像识别领域获得了巨大成功。借鉴和采用深度学习的汉字识别技术省略了人工的特征提取过程,简化了传统 OCR 技术的流程,在识别精度上有了质的飞跃。在 2013 年 ICDAR 举办的汉字识别竞赛中,基于卷积神经网络的联机手写识别和脱机手写识别的准确率分别达到了 97.39%和 94.77%,超过了采用统计模式识别的方法^[16]。目前,基于深度学习

的汉字识别技术已成为汉字识别领域的主流技术^[17-19]。

3) 手写汉字识别(HCCR)。由于汉字的类别繁多、相似字多且书写风格随意等特点,手写汉字识别的难度很大,所以一直是汉字识别研究领域的热点问题^[20]。手写字体识别有联机(online)和脱机(offline)两种方式。联机手写汉字是在智能手机等

电子设备上人工书写而成,书写过程中能够即时获得笔画轨迹特征,使得汉字更易被识别。而脱机手写汉字识别处理的是经扫描仪或摄像头等设备采集到的图片,由于没有笔画笔顺等信息,而且图片中的文字背景或分辨率等存在着一定的噪声干扰,脱机手写汉字识别比联机手写汉字识别更具有挑战性。

表 2 频次前 20 的高频关键词列表

关键词	中文含义	频次	中心度	年份
recognition	识别	159	0.15	2001
character recognition	字符识别	89	0.12	2001
feature extraction	特征提取	79	0.09	2001
online	联机	72	0.04	2005
segmentation	分割	54	0.2	2001
chinese character	汉字	49	0.15	2001
convolutional neural network	卷积神经网络	41	0.08	2009
classification	分类	40	0.36	2001
handwritten chinese character recognition	手写汉字识别	38	0.1	2002
algorithm	算法	38	0.25	2001
chinese character recognition	汉字识别	36	0.13	2001
neural network	神经网络	35	0.09	2001
character	字符	31	0.27	2001
deep learning	深度学习	31	0.04	2015
pattern recognition	模式识别	24	0.06	2001
extraction	提取	21	0.03	2001
model	模型	21	0.13	2002
character segmentation	字符分割	20	0.03	2001
handwriting recognition	手写识别	20	0.17	2009
ocr	光学字符识别	20	0.05	2003

4 汉字识别领域的引文分析

4.1 主题演进分析

利用 CiteSpace 软件进行共被引文献分析,将共被引文献按照研究方向进行聚类,聚类的主题反映了研究前沿的领域^[21]。科技发展日新月异,及时地识别研究领域的前沿有助于推动该领域的研究发展。在 CiteSpace 中进行共被引文献分析,通过 LLR 算法提取聚类主题,生成共被引时间线图谱,如图 5 所示,共生成了 19 个聚类。时间线图谱展现了汉字识别领域随着时间推移而不断演进的研究前沿动态。图 5 中,节点越大表示文献被引用的频率越高。通过聚类文献的时间跨度来探究研究领域的兴起、繁荣和衰落过程。有些研究领域短暂兴起后淡出,而有些研究领域顺应时代发展而被学者重视,开展深入研究并创造出许多高影响力的成果。为了更加直观地分析前沿主题的演进过程,将各聚类名称及时间跨度罗列在表 3。结合时间线图谱分析,可将汉字识别研究的主题演化按照起始时间大致划分为早期、中期和近期 3 个研究阶段。

1) 早期阶段。聚类 #1、#9、#10、#11、#14、#16、#17、#19 这 8 个聚类是早期的研究主题,它们是从 2000 年之前延续或刚兴起的研究主题,其中只有聚类 #1 修正二次判别函数(MQDF)受到的关注最多,热度持续时间最长,直到 2013 年结束。其他聚类的文献引用频次不高,影响力不足。在此阶段中,通常采用传统的 OCR 技术,它有一个成熟的技术流程:图像预处理、特征提取和分类识别。常用的分类方法有掩模匹配法、支持向量机、MQDF 等^[22-25]。传统的 OCR 技术在单字符的汉字识别中识别精度较高,但对文本行识别或复杂背景文字等情况的识别精度较低,且难以取得突破性进展。

2) 中期阶段。聚类 #5、#6、#7、#12、#18 这 5 个聚类主题在 2006 年附近开始逐渐受人关注,其中聚类 #6 联机、聚类 #7 字符串识别这 2 个聚类的高被引文献较多,关注度较高。在此期间随着智能手机的流行,基于智能设备的联机手写汉字识别迎来了应用高潮,不断产生新的应用和技术需求。OCR 技术从单字符的识别转变为字符串及文本行

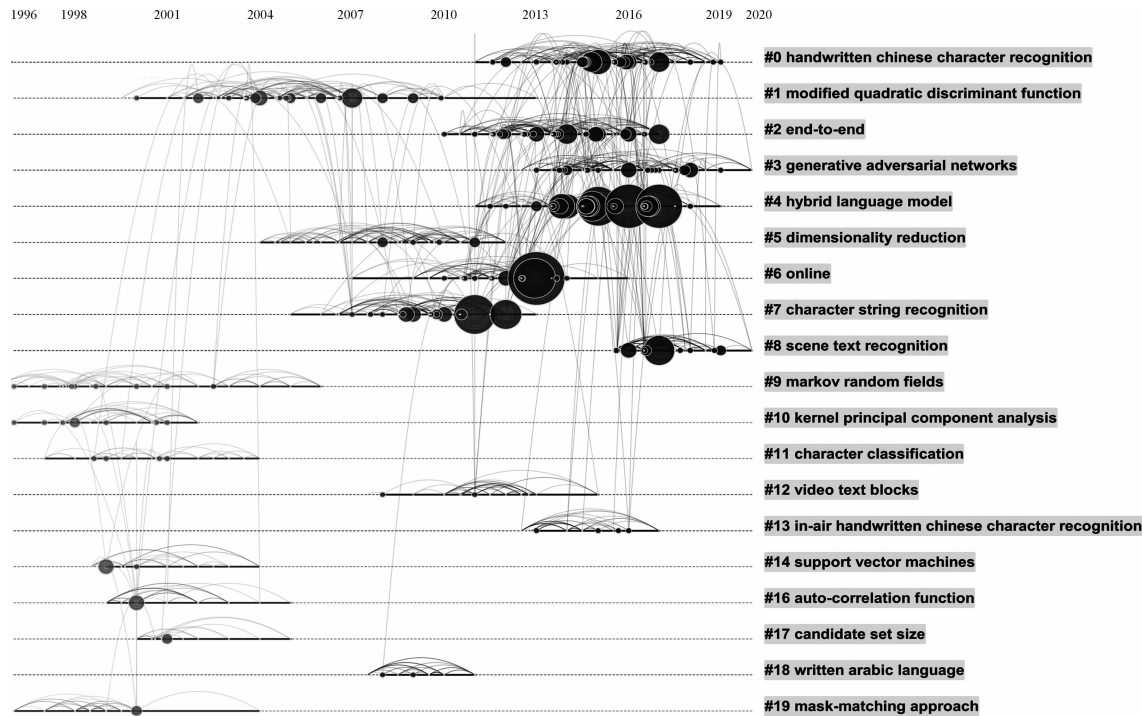


图 5 共被引文献时间线图谱

的识别,通过利用文本行的序列信息,提升识别效果^[26-28]。

表 3 共被引聚类的时间跨度

发展时期	聚类编号	聚类名称	起始时间	结束时间
近期	# 0	手写汉字识别	2011	2019
	# 2	端到端	2010	2017
	# 3	生成对抗网络	2013	2020
	# 4	混合语言模型	2011	2019
	# 8	场景文本识别	2016	2020
	# 13	空中手写汉字识别	2013	2017
中期	# 5	降维	2004	2012
	# 6	联机	2007	2016
	# 7	字符串识别	2005	2013
	# 12	视频文本块	2008	2015
	# 18	手写阿拉伯语	2008	2011
早期	# 1	修正二次判别函数	2000	2013
	# 9	马尔可夫随机场	1996	2006
	# 10	核主成分分析	1996	2002
	# 11	字符分类	1998	2004
	# 14	支持向量机	1999	2004
	# 16	自相关函数	1999	2005
	# 17	候选集大小	2000	2005
	# 19	模板匹配法	1996	2004

3)近期阶段。聚类# 0、# 2、# 3、# 4、# 8、# 13这 6 个聚类主题在 2010 年后开始出现且研究热度持续到近几年,这些主题有较多的高被引文献,影响力高^[29-31]。因此,这几个聚类主题是需要重点关注的研究前沿领域。在此期间,汉字识别技

术出现重大变革,由传统方法的 OCR 技术识别转变成基于深度学习的 OCR 技术。随着深度学习的不断发展以及计算设备能力的不断提升,基于深度学习的汉字识别取得极大进展。在应用上向社会生产生活各领域渗透,产生的应用领域有脱机手写汉字识别、空中手写汉字识别、特定场景文本识别及自然场景文本识别等。

4.2 研究前沿分析

如果一篇论文的引用频次突然呈现急速增长,表明该研究领域受人关注或具备里程碑意义,也代表了该研究领域的前沿。CiteSpace 具备探测引文突现功能,在共被引分析的基础上,通过“Burstness”进行突变率检测,得到汉字识别领域引用频次激增的参考文献^[32]。为了探测近期的研究前沿,应重点关注引文突现时间持续到最近的文献,共有 7 篇文献,详见表 4。这些文献分布在聚类# 0 手写汉字识别、# 3 生成对抗网络、# 4 混合语言模型和# 8 场景文本识别这 4 个聚类中,说明这 4 个聚类主题是近期的研究热点及前沿,将这 4 个聚类的详细信息罗列在表 5。

聚类# 0 手写汉字识别是汉字识别研究领域开展最早且发展时间最长的研究主题。聚类中突变率最高的文献是 2015 年发表在 Nature 上的深度学习综述^[33],介绍了各种深度学习的模型及发展现状,包含 CNN、DBN 和 RNN 等。其中,CNN 在处

理图像、视频、语音等方面带来了突破性的进展,而RNN则在文本和语音等顺序数据方面大放异彩。在以CNN为代表的深度学习技术的协助下,手写汉字识别能力大步提升。

聚类#4混合语言模型,该聚类中还包含空中书写、GoogLeNet、cnn压缩、手写体中文文本识别等研究内容,与聚类#0的研究内容存在交集。聚类#4中含有4篇突发文献,说明了该聚类是关注度较高的研究前沿。突变率最高的文献是微软研究院的He等^[15]于2016年发表的关于ResNet网络模型的设计及应用,该模型通过增加相当的深度来提高识别准确率,且能够有效降低深层CNN模型的梯度消失问题。华南理工大学研究团队^[34]和中科院研究团队^[35]都利用了传统的特征提取方法与神经网络相结合的混合语言模型来提高手写汉字识别性能。由于深度神经网络的训练过程既耗时又非常消耗计算机资源,使得它们无法在便携式设备中部署。Xiao等^[36]设计了一种降低计算成本的算法和9层CNN模型用于脱机手写汉字识别,在保证识别精度的前提下大大提高了计算速度。

聚类#3生成对抗网络是一种深度学习模型,包含生成模型和判别模型,两者之间互相对抗学习以提升模型精度^[37]。该聚类还包含了人工智能、输

入自然图像、密集卷积网络等研究内容。与聚类#8场景文本识别的研究内容有部分重叠。聚类#3中的突现论文介绍了一种汉字生成的方法,该文利用RNN模型实现了一个富有挑战性的任务,即教会机器自动书写汉字^[38]。

聚类#8场景文本识别,不同于发展地比较成熟的特定场景的OCR技术,自然场景中的文字具有字体多变、排列不一且文字背景复杂等特点,一直都是汉字识别研究领域的难点^[39]。聚类#8中的突现论文突变率排在第2名,高达12.37,表明了该研究主题的关注度越来越高。该文提出了一种端到端的CRNN网络,将检测和识别过程在同一个网络框架中进行,实现了基础特征的共享,既能够减少重复计算又能提高特征质量,实现了自然场景文字的高效识别^[40]。

综上所述,通过引文的突现探测,得到了7篇近期的突现文献,这些文献的关注度激增,反映了汉字识别领域的研究前沿。这些文献所涉及的研究内容有深度学习、深度残差网络、传统方法与深度学习相结合的混合模型、CNN加速与压缩、汉字生成、端到端OCR模型、手写汉字识别与场景文本识别。归属于以下4个前沿聚类主题:手写汉字识别、混合语言模型、生成对抗网络和场景文本识别。

表 4 基于引文突现的前沿分析

文献	频次	突变率	突现时间	研究内容	研究机构	聚类
LeCun Y, 2015, NATURE, V521, P436 ^[33]	25	6.47	2017—2021	深度学习	Facebook 人 工 智 能研究院	#0
He K M, 2016, PROC CVPR IEEE, V0, P770 ^[15]	59	13.37	2017—2021	深度残差网络 ResNet	微软研究院	#4
Zhong Z Y, 2015, PROC INT CONF DOC, V0, P846 ^[34]	42	8.84	2017—2021	传统方法与 GoogLeNet 相结合	华南理工大学	#4
Zhang X Y, 2017, PATTERN RECOGN, V61, P348 ^[35]	53	11.38	2017—2021	传统方法与 convNet 相结合	中科院	#4
Xiao X F, 2017, PATTERN RECOGN, V72, P72 ^[36]	21	6.93	2018—2021	CNN 模型加速与压缩	华南理工大学	#4
Zhang X Y, 2018, IEEE T PATTERN ANAL, V40, P849 ^[38]	16	6.96	2019—2021	汉字生成	中科院	#3
Shi B G, 2017, IEEE T PATTERN ANAL, V39, P2298 ^[40]	33	12.37	2019—2021	端到端的卷积循环神经网络(CRNN)	华中科技大学	#8

表 5 研究前沿的聚类信息

聚类	聚类主题	聚类数量	轮廓值	聚类成员
#0	手写汉字识别	71	0.886	任务分析、车牌检测、文本检测、特征提取
#3	生成对抗网络	53	0.887	人工智能、输入自然图像、密集卷积网络、对抗性图像生成
#4	混合语言模型	48	0.876	空中书写、GoogLeNet、cnn 压缩、手写体中文文本识别
#8	场景文本识别	38	0.901	多场景古汉语文本、透视矫正、自动翻译、任意排列中文文本

5 结论

将 WOS 核心合集数据库中的汉字识别领域的

相关文献作为数据来源,对 2001—2021 年的发文趋势、研究力量、热点关键词和引用文献进行分析和

可视化,从量化的角度总结了汉字识别领域的研究现状、研究热点和研究前沿。综合上述研究,可得到如下结论:

1) 汉字识别领域在 2001—2021 年的发展历程中出现了两个明显的增长高峰:2008 年和 2017 年。中国发表的论文数量占总论文的 75% 以上,在汉字识别研究领域起主导作用。对研究力量的分布进行分析,中科院的刘成林团队、华南理工大学的金连文团队和清华大学的丁晓青团队在汉字识别研究领域取得的成果较丰硕,团队的合作关系显著。结果显示有较强合作关系的作者发文量较多。为了促进汉字识别研究领域的发展,应加强研究机构之间或者研究者间的合作交流,资源共享及共同进步。

2) 高频关键词反映了汉字识别领域主要的研究热点:汉字识别技术和手写汉字识别。汉字识别技术由传统的 OCR 识别技术转变为基于深度学习的识别,并成为了近几年重要的研究热点。

3) 基于共被引聚类时间图谱的分析,得到汉字识别领域的前沿主题演化趋势:由早期基于传统方法的 OCR 技术研究演化为中期的联机手写汉字识别与文本行识别,再到近期基于深度学习的汉字识别研究。

4) 对近期的引文突现进行分析得知,研究前沿的主题为手写汉字识别、混合语言模型、生成对抗网络和场景文本识别。

随着深度学习技术在 OCR 领域的应用,汉字识别的性能得到了显著提升,OCR 技术将朝着高效率、智能及一体化发展。汉字识别的应用范围从简单的印刷体识别,逐步演进到手写文本识别与自然场景文本识别等复杂情形。

参考文献

- [1] MEMON J, SAMI M, KHAN R A. Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR) [J]. IEEE Access, 2020, 8: 142642-142668.
- [2] HE M X, HAO P. Robust Automatic recognition of Chinese license plates in natural scenes [J]. IEEE Access, 2020, 8: 173804-173814.
- [3] 王润民, 桑农, 丁丁, 等. 自然场景图像中的文本检测综述 [J]. 自动化学报, 2018, 44(12): 2113-2141.
- [4] 刘荟悦. 基于深度神经网络的印刷体文字识别 [D]. 济南: 山东大学, 2020.
- [5] 唐结玲. 基于视频流的文本识别研究与实现 [D]. 成都: 电子科技大学, 2019.
- [6] 王馨悦. 离线中文手写体识别研究 [D]. 合肥: 中国科学技术大学, 2019.
- [7] CHEN C M. Science mapping: a systematic review of the literature [J]. Journal of Data and Information Science, 2017, 2(2): 1-40.
- [8] 李杰, 陈超美. CiteSpace: 科技文本挖掘及可视化 [M]. 2 版. 北京: 首都经济贸易大学出版社, 2017.
- [9] 吴丹, 向筱茜, 冀晨辉. 中国科技创新研究热点及演化的可视化分析 [J]. 科技和产业, 2021, 21(9): 1-6.
- [10] LIU C L, YIN F, WANG D H, et al. CASIA online and offline Chinese handwriting databases [C]//International Conference on Document Analysis & Recognition. IEEE Computer Society, 2011: 37-41.
- [11] LIU C L, YIN F, WANG D H, et al. Online and offline handwritten Chinese character recognition: Benchmarking on new databases [J]. Pattern Recognition, 2013, 46(1): 155-162.
- [12] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2012, 25(2): 1097-1105.
- [13] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014: 1409. 1556.
- [14] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015: 1-9.
- [15] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2016: 770-778.
- [16] YIN F, WANG Q F, ZHANG X Y, et al. ICDAR 2013 Chinese handwriting recognition competition [C]//2013 12TH International Conference on Document Analysis and Recognition (ICDAR). Washington: IEEE, 2013: 1464-1470.
- [17] 周甲甲. 基于深度学习的汉字识别方法研究 [D]. 武汉: 华中师范大学, 2019.
- [18] LIN D Z, LIN F, LV Y P, et al. Chinese character CAPTCHA recognition and performance estimation via deep neural network [J]. Neurocomputing, 2018, 288: 11-19.
- [19] WANG K L, YI Y H, TANG Z W, et al. Multi-scene ancient Chinese text recognition with deep coupled alignments [J]. Applied Soft Computing, 2021, 108(4): 107475.
- [20] 金连文, 钟卓耀, 杨钊, 等. 深度学习在手写汉字识别中的应用综述 [J]. 自动化学报, 2016, 42(8): 1125-1141.
- [21] 李嘉明, 赵阔, 屈挺, 等. 基于知识图谱的区块链物联网领域研究分析 [J]. 计算机科学, 2021, 48(S1): 563-567.
- [22] LIU C L, JAEGER S, NAKAGAWA M. Online recognition of Chinese characters: The State-of-the-Art [J]. Pat-

- tern Analysis & Machine Intelligence IEEE Transactions on, 2004, 26(2):198-213.
- [23] WANG Y W, DING X Q, LIU C S. MQDF retrained on selected sample set[J]. Leice Transactions on Information & Systems, 2011, 94(10):1933-1936.
- [24] SU T, ZHANG T, GUAN D. Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text[J]. International Journal of Document Analysis & Recognition, 2007, 10(1):27.
- [25] GAO T F, LIU C L. High accuracy handwritten Chinese character recognition using LDA-based compound distances[J]. Pattern Recognition, 2008, 41(11):3442-3451.
- [26] WANG Q F, YIN F, LIU G L. Handwritten Chinese text recognition by integrating multiple contexts[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(8):1469-1481.
- [27] DU J, HUO Q. A discriminative linear regression approach to adaptation of multi-prototype based classifiers and its applications for Chinese OCR[J]. Pattern Recognition, 2013, 46(8):2313-2322.
- [28] SU T H, ZHANG T W, GUAN D J, et al. Off-line recognition of realistic Chinese handwriting using segmentation-free strategy[J]. Pattern Recognition, 2009, 42(1):167-182.
- [29] MESSINA R, LOURADOUR J. Segmentation-free handwritten Chinese text recognition with LSTM-RNN[C]//2015 13th International Conference on Document Analysis and Recognition (ICDAR). Nancy; IEEE, 2015:171-175.
- [30] WU C P, FAN W, YUAN H, et al. Handwritten character recognition by alternately trained relaxation convolutional neural network[C]//2014 14th International Conference on Frontiers in Handwriting Recognition. IEEE, 2014:291-296.
- [31] XIE Z C, SUN Z H, JIN L W, et al. Learning spatial-semantic context with fully convolutional recurrent network for online handwritten Chinese text recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, 40(8):1903-1917.
- [32] 王金丽, 樊勇, 张辉. 区块链文献主题发现及演化研究[J]. 计算机工程与应用, 2020, 56(20):1-8.
- [33] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521:436-444.
- [34] ZHONG Z Y, JIN L W, XIE Z C. High performance offline handwritten Chinese character recognition using GoogLeNet and directional feature maps[C]//2015 13th International Conference on Document Analysis and Recognition (ICDAR). Nancy; IEEE, 2015:846-850.
- [35] ZHANG X Y, BENGIO Y, LIU C L. Online and offline handwritten Chinese character recognition: A comprehensive study and new benchmark[J]. Pattern Recognition, 2017, 61:348-360.
- [36] XIAO X F, JIN L W, YANG Y F, et al. Building fast and compact convolutional neural networks for offline handwritten Chinese character recognition[J]. Pattern Recognition, 2017, 72:72-81.
- [37] 王万良, 李卓蓉. 生成式对抗网络研究进展[J]. 通信学报, 2018, 39(2):135-148.
- [38] ZHANG X Y, YIN F, ZHANG Y M, et al. Drawing and recognizing Chinese characters with recurrent neural network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4):849-862.
- [39] 白翔, 杨明钊, 石葆光, 等. 基于深度学习的场景文字检测与识别[J]. 中国科学:信息科学, 2018, 48(5):531-544.
- [40] SHI B G, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11):2298-2304.

Analysis on the Research Progress of Chinese Character Recognition Technology

LI Yangjuan

(Engineering Training Center, Putian University, Putian Fujian 351100, China)

Abstract: In order to keep a comprehensive and systematic understanding of the progress of Chinese character recognition, the literature data of Chinese character recognition in Web of Science database are selected and knowledge mapping analysis is made by CiteSpace visualization software. The research results show that Chinese universities and research institutes are the core force of Chinese character recognition. The development of Chinese character recognition technology has evolved from the early traditional OCR technology to the online handwritten Chinese character recognition in the middle period, and then to the deep learning in recent research of Chinese character recognition. The research frontiers topics are handwritten Chinese character recognition, hybrid language model, generative adversarial networks and scene text recognition. Chinese character recognition technology will develop towards high efficiency and intelligence.

Keywords: Chinese character recognition; OCR; deep learning; handwritten Chinese character