基于深度学习的场景文字检测综述

姜　维[1],张重生[2],殷绪成[3]
(1. 华北水利水电大学信息工程学院,河南郑州 450045;2. 河南大学计算机与信息工程学院,河南开封 475001;
3. 北京科技大学计算机与通信工程学院,北京 100083)

摘　要： 近年来,基于深度学习的场景文字检测技术取得重要进展. 本文综述了该技术在 2014 ~ 2018 年间的最新工作,将其分为传统区域建议方法、文字建议网络方法、基于分割的方法以及文字建议网络与分割的混合方法,并对各类方法的优劣进行分析. 本文还展望了未来发展趋势,指出未来研究热点.

关键词： 深度学习；场景文字；检测定位
中图分类号： TP391　　**文献标识码：** A　　**文章编号：** 0372-2112 (2019)05-1152-10
电子学报 URL： http://www.ejournal.org.cn　　**DOI：** 10.3969/j.issn.0372-2112.2019.05.024

Deep Learning Based Scene Text Detection: A Survey

JIANG Wei[1], ZHANG Chong-sheng[2], YIN Xu-cheng[3]
(1. *School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou, Henan* 450045, *China*;
2. *School of Computer and Information Engineering, Henan University, Kaifeng, Henan* 475001, *China*;
3. *School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing* 100083, *China*)

Abstract: In recent years, deep learning based scene text detection have achieved significant progress. The paper reviews state-of-the-art methods in the field from 2014-2018. We categorize existing methods into traditional Region Proposal based method, Text Proposal Network method, segmentation based method and hybrid method based on Text Proposal Network and segmentation with detailed analysis of pros and cons for the four methods. Finally, we point out research trends and focuses in this field.

Key words: deep learning; scene text; text detection

1　引言

文字是人类最重要的信息载体,记载了几千年的人类文明和历史[1],通过计算机进行文字识别具有重要价值. 20 世纪 80 年代,国内清华大学丁晓青团队开始了文档识别的研究,取得了丰硕成果. 但是,至今自然场景文字(简称场景文字)检测与识别问题仍尚未解决. 图 1 对比了传统 OCR 和场景文字检测与识别的研究对象和难点. 其中,场景文字问题的难点在于如下几个方面:(1)背景复杂;(2)文字颜色多变;(3)光照条件的不确定性;(4)文字排列的不确定性;(5)文字类型、字体与大小的不确定性;(6)文字位置的不确定性.

基于以上难点,2014 年之前的方法[2~5]无法有效解决问题,因此国内外研究者尝试使用深度学习技术解决问题. 本文目标是对基于深度学习的场景文字检测成果梳理、分类和对比,进而分析该领域的发展趋势,帮助研究者系统了解领域内相关算法与技术. 如图 2 所示,本文将基于深度学习的场景文字检测方法分为传统区域建议的方法、文字建议网络的方法、基于分割的方法以及文字建议网络与分割的混合方法四种类型并进行详细阐述.

2　主要算法介绍

场景文字检测问题的研究大致可分为两个时期:传统方法时期和深度学习时期. 本文重点在于深度学习时期研究成果,下面先介绍传统方法时期的主要工作.

2.1　传统方法时期

在该时期,研究者主要使用人工设计特征与传统分类器,多数算法如图2所示包含两个阶段,即文字候

收稿日期:2018-10-11;修回日期:2019-01-14;责任编辑:马兰英
基金项目:国家自然科学基金(No. 61601184, No. 61562058);河南省高等学校重点科研项目(No. 16A520018)

图 1　传统OCR与场景文字的研究难点对比

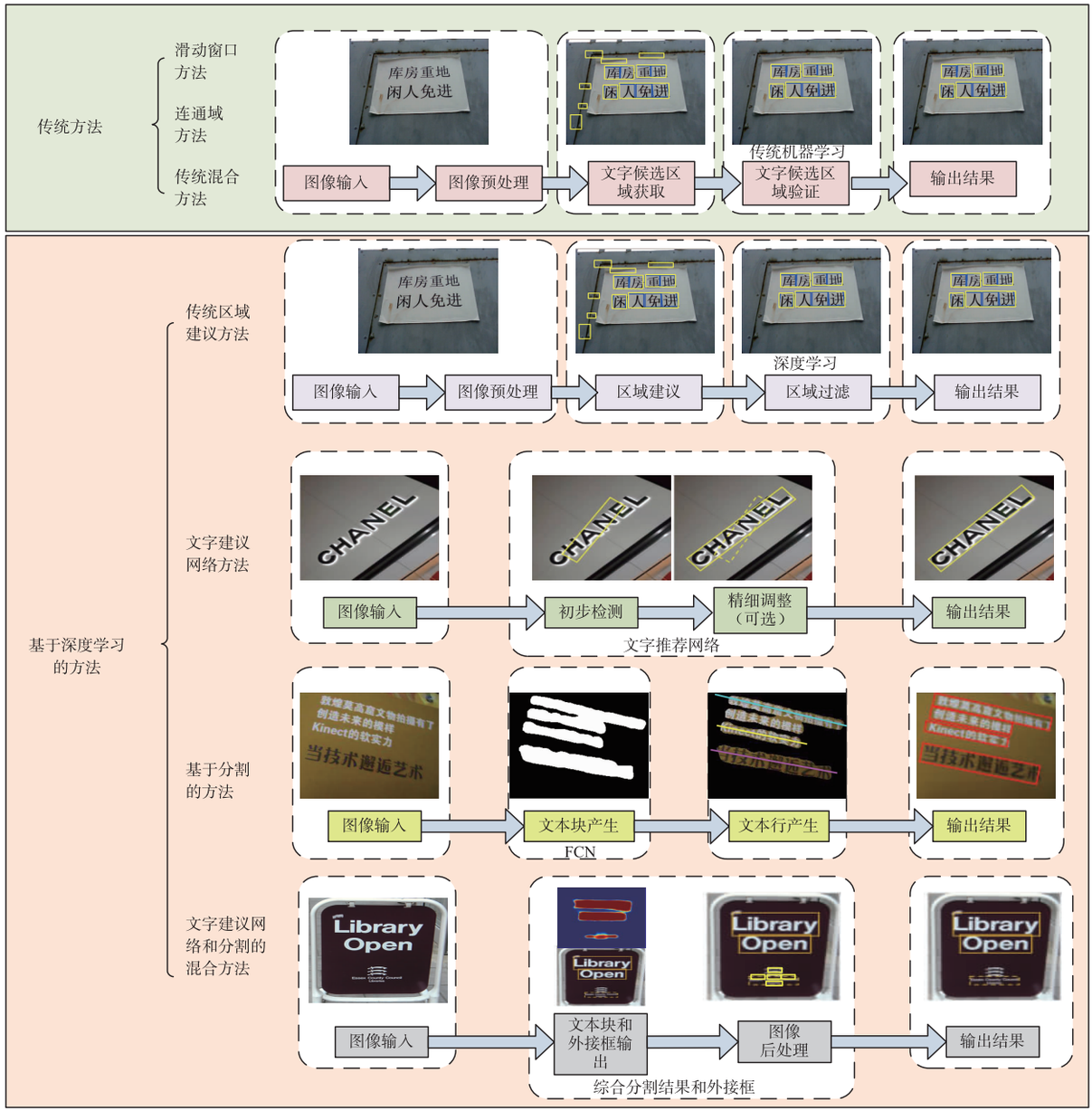


图2　场景文字检测方法分类和结构图

选区获取阶段与验证阶段. 在文字候选区获取阶段中,根据获取候选区的不同方式可以分为滑动窗口方法(Sliding-window Approach)[6~8]、连通域方法(CC Approach)[9~12]与二者混合的方法(Hybrid Appoach)[13~15].

在文字候选区域验证阶段,研究者使用的是传统机器学习方法,但其存在问题:人工设计特征区分能力不足;浅层分类器无法适用复杂场景.

2.2 深度学习方法时期

在该时期,本领域产生传统区域建议方法、文字建议网络方法、基于分割的方法以及区域建议与分割的混合方法,如图 2 所示. 以上方法的提出时间和方法思路的来源如下表 1 所示.

表 1 四种方法的简介

名称	提出时间	思路来源
传统区域建议的方法	2011	传统方法 + 深度学习
文字建议网络的方法	2016	似物性建议
基于分割的方法	2016	图像语义分割
文字建议网络与分割的混合方法	2017	文字建议网络与分割方法的结合

2.2.1 传统区域建议的方法

该类方法源于似物性建议(Object Proposal)[16],通过滑动窗口或连通域分析产生建议区域,然后再对其进行过滤且合并文本行,最终找到文字区域外接框. 如图 3 所示,该类方法延续了传统方法分阶段的思想,但使用深度神经网络作为分类器,这是与传统方法不同之处. 传统方法中的滑动窗口与 MSER 方法在区域建议中广泛使用.

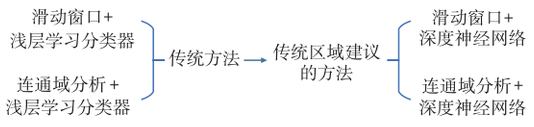


图3 传统方法与传统区域建议的方法对比

(1)滑动窗口方法 此类方法有两种模式:第一种是采用卷积神经网络结合滑动窗口做区域建议,然后采用传统分类器消除虚警率;第二种是采用传统特征与分类器结合滑动窗口做区域建议,然后采用卷积神经网络消除虚警率. 文献[17~19]属于第一种模式,文献[20~22]属于第二种模式,此类方法是在传统方法时期的滑动窗口方法的基础上,使用卷积神经网络做文字与背景的区分.

(2)连通域方法 在连通域分析中,MSER 是最为有效的方法之一. 所以在此类方法中,研究者多通过 MSER 或者增强 MSER 方法进行区域建议,然后采用传统分类器结合卷积神经网络进行背景滤除,最终得到场景文字检测的结果. 文献[23~26]属于此类方法,基本参照的是以上思路. 综上,传统区域建议方法相较于传统方法阶段有很大提升,但依然存在问题,优缺点如下表 2 所示. 在本类方法中,只有文献[26]所述方法可以检测多方向场景文字.

表 2 传统区域建议方法的优劣

名称	优点	缺点
传统区域建议的方法	延续传统方法的思路,结合深度学习方法,容易设计且性能较传统方法更优	依然是分阶段方法,错误容易累积; 依然采用传统方法思路,未充分利用深度神经网络; 多数只考虑了水平排列文字,而忽略倾斜或者弧形排列等多方向文字.

2.2.2 文字建议网络方法

2015 年,用于目标识别的 Faster R-CNN(Regions with Convolutional Neural Network)[27]提出,给本领域带来突破. Faster R-CNN 是基于区域建议网络(Region Proposal Network)进行工作,采用的是整体化思想,将目标检测的各阶段整合进入了深度神经网络. 该方法可以避免阶段错误的积累,也可将区域建议转入到 GPU 计算,提高了算法运行速度. 在 Faster R-CNN 后,SSD 与 YOLO 等网络结构相继提出. 这些网络模型在获取检测外接框时采用了不同方式,分别是间接回归与直接回归方式. Faster R-CNN 属于间接回归,需要对初步回归结果做二次调整;SSD 与 YOLO 属于直接回归,回归结果无需二次调整. 因此,源于以上模型的文字建议网络方法也分为两种类别.

(1)间接回归 Faster R-CNN 的思路进入了场景文字检测领域,形成了文字区域建议网络方法. 该类方法大致分为三个部分:1、特征提取:通过卷积方式获取整幅图像的特征图;2、初步检测:划分图像的网格(grid cell),在网格中心计算锚点框(anchor box)的置信度,得到初步检测结果;3、精细调整:在初步检测结果上,使用回归(regression)方法或其他方式精细调整检测外接框(bounding box),最终得到检测结果.

按照以上所述结构,Zhong[28]在 2016 年提出 DeepText 方法,将各阶段整合进入网络结构. Tian[29]提出 Connectionist Text Proposal Network(简称 CTPN),使用长短期双向记忆模型处理文字建议序列,进而得到文字区域外接框与置信度. Ma[30]提出 Rotation Region Proposal Networks(简称 RRPN),该网络为锚点增加了 6 个方向的建议,可检测任意方向的直线排列的场景文字. Liu 提出 Deep Matching Prior Network[31](简称 DMPNet),在粗预测阶段使用共享蒙特卡洛方法粗略估算,精调整阶段使用四边形和 *Ln* 范数精确回归. 该网络模型可检测任意方向的直线排列的场景文字,但不包括

弧形排列场景文字.

针对弧形排列场景文字,Liu 在 2017 年构建了针对弧形排布的场景文字的数据集 SCUT-CTW1500[32],设计了训练样本的 14 点等分标注法. 在此基础上,使用 Resnet-50[33] 与长短期双向记忆模型完成弧形排布的场景文字检测. 2018 年,Liu[34] 提出共享检测和识别信息的快速端对端的场景文字检测识别方法.

(2)直接回归　Faster R-CNN 是基于区域建议的方法,而 SSD 与 YOLO 是基于网格与锚点实现的,所以不需要二次回归. 基于此原理,Gupta[35] 等使用人工合成场景文字设计并训练基于 YOLO[36] 的深度神经网络. Liao[37] 基于 SSD 分别提出“TextBoxes + +”网络模型,“TextBoxes + +”可检测任意方向的直线排列的场景文字. Shi[38] 参考 SSD[39] 提出“SegLink”网络模型,该模型先产生文字“片段”(segment),再建立“连接”(link),最终根据连接和片段输出任意方向直线排列的场景文字. Liao[40] 将任务分为分类和回归,分类采用旋转不敏感特征,回归采用旋转敏感特征,按照类似 SSD 网络结构,检测任意方向的直线排列的场景文字.

综上,本类方法是整体性方案,输入图像后可直接给出场景文字外接框的相关几何属性. 该类方法一定程度摒弃了传统方法的既有思路,更充分利用了深度学习技术,但也有相应优缺点,具体如下表 3 所示.

表 3　文字建议网络方法的优劣

名称	优点	缺点
文字建议网络方法	采用整体性思想,避免各阶段错误的积累,计算速度较快	该类方法的输出是外接框与外接框的置信度,因此在一些情况下检测出的外接框不够精确,且无法调整. 受外接框的表示方法的影响,会损失一部分文字的几何信息.

2.2.3　基于分割的方法

2015 年,Long[41] 提出全卷积神经网络,并将其用于图像的语义分割. 语义分割不仅把图像分割为多个区域,而且对分割区域进行了分类. 基于以上原理,本类方法将场景文字的检测定位问题,转化为场景文字与背景的语义分割问题. 如图 2 所示,该类方法首先完成场景文字与背景的语义分割,得到场景文字块区域;然后精细分割获取文字行;最终输出文字行的位置和几何属性. 该类方法相比文字建议网络,有其优势,但也存在固有缺陷:如图 4 所示当相邻文字行距离过近时,分割结果会发生黏连. 为了改善分割效果,研究者纷纷提出自己的改进方法,按照改进方式不同,大致分为两类:多信息融合与多阶段级联.

(1)多信息融合　该类方法融合多种不同信息,以达到精确分割场景文字与背景的目标. Yao[42] 提出基于

图4　相邻文字行分割结果黏连问题[42]

多通道信息的场景文字检测方法. 该方法以文本置信图、字符置信图和字符连接方向图等通道信息训练全卷积神经网络. Polzounov[43] 提出 WordFence 的概念,表示场景文字边界附近的区域. 该方法(简称 WDN Recognition)不仅分割出了文字核心区域,也分割出了文字边界区域. Xue[44] 发展了 WordFence 的概念,将文字边界分为上下左右四种边界. Long[45] 提出以圆心在文字中心线上的圆盘序列形状描述文本形状的场景文字检测方法(简称 TextSnake). 该方法基于全卷积网络,借助圆盘中心连线的切线角度与圆盘半径等信息,可以检测任意方向与任意形状排列的场景文字.

(2)多阶段级联　该类方法通过多个全卷积网络的级联,达到精确分割场景文字与背景的目标. Zhang[46] 先利用全卷积网络得到文字块,然后在文字块中计算文本行的方向和获取文本行候选,再估算字符中心,最终检测多方向排布的场景文字. He[47] 提出 Cascaded Convolutional Text Network(简称 CCTN),该网络包含粗检测网络和精检测网络. Tang[48] 提出一种由检测网络、分割网络与分类网络组成的级联结构,用于场景文字的检测和分割. He[49] 提出基于多尺度全卷积与级联实例感知分割的场景文字检测方法. Deng[50] 提出基于实例分割的场景文字检测方法(简称 PixelLink),该方法首先进行语义分割,然后对文字候选像素,进行文字类别和连通预测,最终得到场景文字的检测结果.

综上,基于分割的方法的检测结果较为精准. 但是,方法本身也是有其自身优劣,具体如下表 4.

表 4　基于分割的方法的优劣

名称	优点	缺点
基于分割的方法	分割结果天然包含类别信息、位置信息与几何信息. 方便对各种排列与方向的场景文字的检测. 经过多阶段级联或多信息融合,分割较为精细. 输出文字外接框与分割结果,便于精细调整.	初始分割结果较为粗糙,不够精确. 多阶段级联的分割方法容易累积错误,后处理通常比较耗时. 多信息融合的分割方法,计算信息种类多,计算耗时.

2.2.4　文字建议网络和分割混合的方法

文字建议网络方法速度较快,采用整体性思路,可避免多阶段错误累积,但检测结果有时不够精确,且外接框表示方法会损失文字几何信息;基于分割的方法的结果天然具有类别、几何与位置信息,但通常需要多

信息融合或多阶段分割,计算量大,速度慢. 因此,有研究者尝试将两种方法融合. Zhou[51]基于 PVANet[52]设计 EAST(Efficient and Accuracy Scene Text detection)网络模型,其输出的结果包括文字置信图、旋转外接框和外接四边形. He[53]提出“direct regression”的概念,并基于此设计包含卷积特征提取、多级特征融合、多任务学习和非极大值抑制的后处理等四个阶段的网络模型. Qin[54]提出一种全卷积语义分割网络与文字建议网络级联的场景文字检测定位方法. Lyu[55]提出基于角点定位与区域分割的场景文字检测定位方法. 该方法只做了场景文字检测工作,因此他[56]基于 Mask R-CNN 模型,提出一种针对场景文字的 text spotting 方法.

综上,该类型的混合方法既可以采用整体化思想,避免阶段错误的累积;又可以较好地进行精细化检测,融合两种方法的优势.

2.2.5 讨论分析

在以上四种方法中,传统区域建议方法从 2011 年开始研究,文字建议网络的方法和基于分割方法是当前的主流方法,文字建议网络和分割的混合方法在 2017 年出现,能够很好对两种方法融合并扬长避短,将是研究的新兴方向之一. 表 5 与表 6 是本文调研算法在各个数据集上的性能评价. 两个表共给出 7 个数据集的评价,其中 ICDAR 2013 多为水平排列场景文字,难度较低,研究者将逐渐转向包含多方向场景文字的 ICDAR 2015 和其它数据集;COCO-Text 数据集较大,难度最高,将得到研究者更多关注. 值得说明,以上数据集均为拉丁文字数据集,公开场景中文数据集缺乏,因此 RCTW 与 CTW 数据集值得关注.

表 5 算法在各数据集表现 1

方法	年份	分类	ICDAR2005			ICDAR2011			ICDAR2013 FOCUS			ICDAR2015 INCIDENTAL		
			P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
FOTS[34]	2018	2	–	–	–	–	–	–	–	–	0.93	0.92	0.88	0.9
Xue 方法[44]	2018	3	–	–	–	–	–	–	0.92	0.87	0.89	–	–	–
Mask TextSpotter[56]	2018	4	–	–	–	–	–	–	0.95	0.89	0.92	0.92	0.81	0.86
TextSnake[45]	2018	3	–	–	–	–	–	–	–	–	–	0.85	0.8	0.83
RRD[40]	2018	2	–	–	–	–	–	–	0.88	0.8	0.84	–	–	–
PixelLink[50]	2018	3	–	–	–	–	–	–	–	–	–	0.86	0.82	0.84
TextBoxes + +[37]	2018	2	–	–	–	–	–	–	0.92	0.86	0.89	0.88	0.79	0.83
Lyu 方法[55]	2018	4	–	–	–	–	–	–	0.92	0.84	0.88	0.9	0.8	0.85
RRPN[30]	2017	2	–	–	–	–	–	–	0.95	0.88	0.91	0.84	0.77	0.8
Tang 方法[48]	2017	3	–	–	–	0.9	0.86	0.88	0.92	0.87	0.9	–	–	–
He W 方法[53]	2017	4	–	–	–	–	–	–	0.92	0.81	0.86	0.82	0.8	0.81
EAST[51]	2017	4	–	–	–	–	–	–	–	–	–	0.83	0.78	0.81
Ma 方法[26]	2017	1	0.82	0.73	0.77	–	–	–	0.91	0.8	0.85	–	–	–
He D 方法[49]	2017	3	–	–	–	–	–	–	0.93	0.79	0.85	0.76	0.54	0.63
DMPNet[31]	2017	2	–	–	–	–	–	–	–	–	–	0.73	0.68	0.71
SegLink[38]	2017	2	–	–	–	–	–	–	–	–	–	0.73	0.77	0.75
Qin 方法[53]	2017	4	–	–	–	–	–	–	0.9	0.83	0.86	0.79	0.65	0.71
WDN Recognition[43]	2017	3	–	–	–	0.64	0.92	0.75	0.65	0.92	0.76	–	–	–
CCTN[47]	2016	3	–	–	–	0.88	0.79	0.84	0.9	0.83	0.86	–	–	–
Text – CNN[24]	2016	1	0.87	0.73	0.79	0.91	0.74	0.82	0.93	0.73	0.82	–	–	–
Zhu 方法[25]	2016	1	–	–	–	–	–	–	0.86	0.74	0.8	–	–	–
Gupta 方法[35]	2016	2	–	–	–	0.94	0.77	0.85	0.92	0.76	0.83	–	–	–
Yao 方法[42]	2016	3	–	–	–	–	–	–	0.89	0.8	0.84	0.73	0.59	0.65
Zhang2016[46]	2016	3	–	–	–	–	–	–	–	–	–	0.71	0.43	0.54
CTPN[29]	2016	2	–	–	–	0.89	0.79	0.84	0.93	0.83	0.88	0.74	0.52	0.61
DeepText[28]	2016	2	–	–	–	0.85	0.81	0.83	0.87	0.83	0.85	–	–	–
Text_Flow[22]	2015	1	–	–	–	0.86	0.76	0.81	0.85	0.76	0.8	–	–	–
Zhang2015[21]	2015	1	–	–	–	0.84	0.76	0.8	0.88	0.74	0.8	–	–	–
Huang 方法[19]	2014	1	0.84	0.67	0.75	0.88	0.71	0.78	–	–	–	–	–	–

1:传统区域建议方法;2:文字建议网络方法;3:基于分割的方法;4:文字建议网络和分割混合的方法

表 6　算法在各数据集表现 2

方法	年份	分类	COCO – Text			SVT			MSRA – TD500		
			P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Xue 方法[44]	2018	3	–	–	–	–	–	–	0. 83	0. 77	0. 80
TextSnake[45]	2018	3	–	–	–	–	–	–	0. 83	0. 74	0. 78
RRD[40]	2018	2	–	–	–	–	–	–	0. 87	0. 73	0. 79
PixelLink[50]	2018	3	–	–	–	–	–	–	0. 83	0. 73	0. 78
TextBoxes + +[37]	2018	2	0. 61	0. 57	0. 59	–	–	–	–	–	–
Lyu 方法[55]	2018	4	0. 63	0. 62	0. 62	–	–	–	0. 88	0. 76	0. 82
RRPN[30]	2017	2	–	–	–	–	–	–	0. 82	0. 69	0. 75
Tang 方法[48]	2017	3	–	–	–	–	0. 76	–	–	–	–
He W 方法[53]	2017	4	–	–	–	–	–	–	0. 77	0. 7	0. 74
EAST[51]	2017	4	0. 41	0. 34	0. 37	–	–	–	0. 87	0. 67	0. 76
He D 方法[49]	2017	3	–	–	–	–	0. 78	–	–	–	–
SegLink[38]	2017	2	–	–	–	–	–	–	0. 86	0. 7	0. 77
WDN Recognition[43]	2017	3	–	–	–	0. 47	0. 63	0. 54	–	–	–
CCTN[47]	2016	3	–	–	–	–	–	–	0. 79	0. 65	0. 72
Text – CNN[24]	2016	1	–	–	–	–	–	–	0. 76	0. 61	0. 69
Zhu 方法[25]	2016	1	–	–	–	0. 41	0. 34	0. 37	–	–	–
Gupta 方法[35]	2016	2	–	–	–	0. 30	0. 41	0. 35	–	–	–
Yao 方法[42]	2016	3	0. 43	0. 27	0. 33	–	–	–	0. 77	0. 75	0. 76
Zhang2016[46]	2016	3	–	–	–	–	–	–	0. 83	0. 67	0. 74

1:传统区域建议方法;2:文字建议网络方法;3:基于分割的方法;4:文字建议网络和分割混合的方法

表 7　常用的场景文字检测数据集介绍

数据集	发布时间	语种	训练集图像个数	测试集图像个数	F-Measure 最大值	下载
ICDAR 2013	2013	英文	229	233	0. 92	[58][59]
ICDAR 2015	2015	英文	1000	500	0. 90	[60]
MSRA – TD500	2012	中英文	300	200	0. 79	[61]
COCO – Text	2016	英文	43686 张训练图像、及 10000 验证图像	10000	0. 59	[62]
ICDAR 2017 – MLT	2017	9 种语言	每种语言 2000 张图像,共计 18000 张图像	9000	0. 72	[63]
RCTW	2017	中英文	8034	4229	0. 67	[64]
CTW	2018	中文	共 32285 张图像,75 % 用于训练,10% 用于分类测试,10% 用于检测测试,5% 用于验证	–	–	[65]
MTWI	2018	中英文	10000	10000	0. 80	[66]

3　数据集

本领域有若干公开数据集,表 7 中列出认可度高且较为重要的数据集. 目前,学术界能够使用的有挑战性的场景文字数据集偏少,主要局限于表 7 中数据集,而以上数据集的测试集普遍较小,且多局限于英文. 而非拉丁文字(如中文等)与多方向的场景文字问题更有挑战,却少有针对性数据集.

4　总结和展望

本文旨在对基于深度学习的场景文字检测定位方法进行整理、分类、归纳与分析,并在此基础上对国内

外技术进行综合评述与展望,得出未来研究趋势如下.

(1)增加模型融合度与专业性

2014 ~2018 年间,本领域的发展得到了图像分类、目标检测与图像分割三个领域的启发(如表 8 所示),进而产生前文所述四类方法. 四类方法本身各有优劣,研究者必然会将不同模型结构相互融合,产生更适用于本领域的模型.

(2)提高算法的鲁棒性与性能

从图 5 中可以看出,在 2003 年 ~2018 年间,公开数据集的图像数量从最少的 252 幅发展到最多的 63686 幅,且难度增大,不断接近真实环境,预示了场景文字检测算法与模型需要更好地鲁棒性与性能.

(3)对场景非拉丁文字(特别是汉字)的检测定位

研究者对场景拉丁文字的检测研究相对充分,且日趋成熟;场景非拉丁文字(特别是汉字)的检测研究工作还比较有限,有很大提升空间,且场景中文数据集较少(如表 7).

表 8 相关三个领域对本领域的影响

领域名称	成果	本领域用途
图像分类	VGG,ResNet 等	网络主干
目标检测	Faster RCNN, YOLO, SSD 等	文字建议网络的框架与思想
图像分割	语义分割与实例分割	基于分割的方法的框架与思想

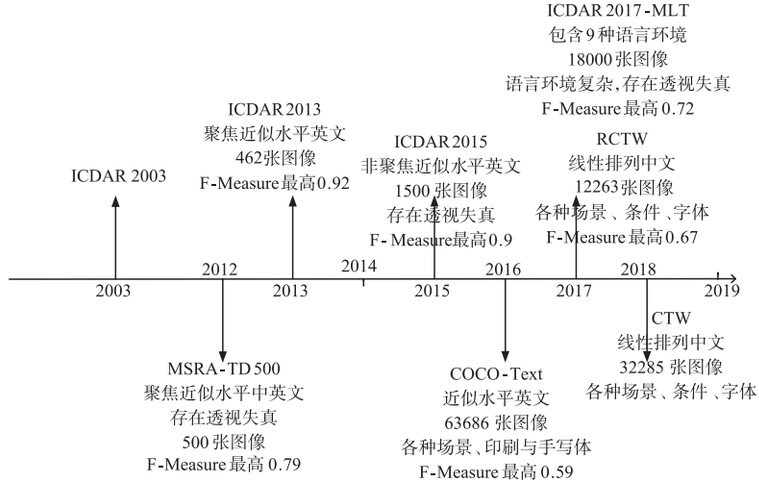


图5 公开数据集情况概览

(4)对多方向多形状排布的场景文字的检测

多方向多形状排布的场景文字是检测问题的难点. 研究所针对的数据集从水平排布的 ICDAR 2013 发展到多方向多形状排布的 SCUT-CTW1500 与 Total-Text,文字形状表示经历了水平矩形,旋转矩形,四边形与多边形或圆盘序列(如图 6)等变化,预示着对多方向多形状排布的场景文字的检测将是下个待解决的难点.

图6 文字外接框示意图

(5)基于深度学习的场景文字识别和端对端识别

场景文字识别相比普通印刷体文字识别,受文字旋转和形变影响很大. 在 SVT-Perspective 与 CUTE80 旋转与形变场景文字数据集中,没有字典支持的识别率在 0.6 ~0.75 左右,还有很大提升空间. 此外,场景中文的识别因为汉字上千个类别,仍是一个艰巨的挑战. 目前研究者更多是将检测与识别问题分开解决,端对端识别的工作还较少,也将会成为下一个研究热点.

(6)对无约束的场景文字的检测定位

遮挡、光照不均、透视变形、噪声、暗对比度、反光以及光照过曝等不利条件是本领域研究必须克服的问题. 除了透视变形,还没有研究者专门针对以上情况进行工作,也没有专门数据集用来评价不同算法分别在这些环境下的表现.

参考文献

[1] 丁晓青,王言伟,等. 文字识别原理方法和实践[M]. 北京:清华大学出版社,2017. 1.
Ding X Q,WANG Y W,et al. Character Recognition:Theories,Methods and Practice[M]. Beijing:Tsinghua University Press,2007. 1. (in Chinese)

[2] Zhang H,Zhao K,Song Y Z,et al. Text extraction from natural scene image: A survey[J]. Neuro Computing,2013,122(51):310 - 323.

[3] Ye Q,Doermann D. Text detection and recognition in imagery:A survey[J]. IEEE Transactions on Pattern Analysis

& Machine Intelligence,2015,37(7):1480 – 1500.

[4] Yin X C, et al. Text detection, tracking and recognition in video: A comprehensive survey[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society,2016,25(6):2752 – 2773.

[5] Zhu Y Y, Yao C, Bai X. Scene text detection and recognition: recent advances and future trends[J]. Frontiers of Computer Science,2016,10(1):19 – 36.

[6] Chen Xiangrong, Yuille A L. Detecting and reading text in natural scenes[A]. IEEE International Conference on Computer Vision and Pattern Recognition[C]. Washington DC: IEEE Computer Society,2004. 366 – 373.

[7] Shehzad Muhammad Hanif, Lionel Prevost, Pablo Augusto Negri. A cascade detector for text detection in natural scene images[A]. IEEE International Conference on Pattern Recognition[C]. Tampa: IEEE Computer Society,2008. 1 – 4.

[8] Shehzad Muhammad Hanif, Lionel Prevost. Text detection and localization in complex scene images using constrained adaBoost algorithm[A]. IEEE International Conference on Document Analysis and Recognition[C]. Barcelona: IEEE Computer Society,2009. 1 – 5.

[9] Boris Epshtein, Eyal Ofek, Yonatan Wexler. Detecting text in natural scenes with stroke width transform[A]. IEEE International Conference on Computer Vision and Pattern Recognition[C]. San Francisco: IEEE Computer Society, 2010. 2963 – 2970.

[10] Yao Cong, Bai Xiang, et al. Detecting texts of arbitrary orientations in natural images[A]. IEEE International Conference on Computer Vision and Pattern Recognition[C]. Providence: IEEE Computer Society,2012. 1083 – 1090.

[11] Neumann Lukas, Jiri Matas. Real-time scene text localization and recognition[A]. IEEE International Conference on Computer Vision and Pattern Recognition[C]. Providence: IEEE Computer Society,2012. 3538 – 3545.

[12] Yin X C, Yin Xuwang, Huang Kaizhu. Robust text detection in natural scene images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, (99): 2264 – 2268.

[13] Pan Y F, Hou X W, Liu C L. A robust system to detect and localize texts in natural scene images[A]. The Eighth IAPR International Workshop on Document Analysis Systems[C]. Nara: IEEE Computer Society,2008. 35 – 42.

[14] Pan Y F, et al. A hybrid approach to detect and localize texts in natural scene images[J]. IEEE Transactions On Image Processing,2011,20(3):800 – 813.

[15] Zhou Gang, Liu Yuehu. Scene text detection based on probability map and hierarchical model[J]. Optical Engineering,2012,51(6):1 – 10.

[16] Hosang J, Dollar P, Dollar P, et al. What makes for effective detection proposals[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2015,38(4):814.

[17] Coates Adam, Blake Carpenter, et al. Text detection and character recognition in scene images with unsupervised feature learning[A]. IEEE International Conference On Document Analysis And Recognition[C]. Beijing, IEEE Computer Society,2011. 440 – 445.

[18] Wang T, Wu D J, Coates A, et al. End-to-end text recognition with convolutional neural networks[A]. International Conference on Pattern Recognition[C]. Stockholm: IEEE Computer Society,2012. 3304 – 3308.

[19] Jaderberg M, Vedaldi A, Zisserman A. Deep features for text spotting[A]. European Conference on Computer Vision[C]. Cham: Springer,2014. 512 – 528.

[20] Jaderberg M, et al. Reading text in the wild with convolutional neural networks[J]. International Journal of Computer Vision,2016,116(1):1 – 20.

[21] Zhang Z, Shen W, Yao C, et al. Symmetry-based text line detection in natural scenes[A]. IEEE Conference on Computer Vision And Pattern Recognition[C]. Boston: IEEE Computer Society,2015. 2558 – 2567.

[22] Tian S, Pan Y, Huang C, et al. Text flow: A unified text detection system in natural scene images[A]. IEEE International Conference on Computer Vision[C]. Sydney: 2016. 4651 – 4659.

[23] Huang W, Qiao Y, Tang X. Robust scene text detection with convolution neural network induced MSER trees [A]. European Conference on Computer Vision[C]. Zurich: Springer,2014. 497 – 511.

[24] He T, et al. Text-attentional convolutional neural network for scene text detection[J]. IEEE Transactions on Image Processing,2016,25(6):2529 – 2541.

[25] Zhu A, Gao R, Uchida S. Could scene context be beneficial for scene text detection[J]. Pattern Recognition, 2016,58:204 – 215.

[26] Ma J, Wang W, Lu K, et al. Scene text detection based on pruning strategy of MSER-trees and Linkage-trees[A]. IEEE International Conference on Multimedia and Expo [C]. Hong Kong: IEEE Signal Processing Society,2017. 367 – 372.

[27] Ren S, Girshick R, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,39(6):1137.

[28] Zhong Z, Jin L, Huang S. DeepText: A new approach for text proposal generation and text detection in natural images[A]. IEEE International Conference on Acoustics, Speech and Signal Processing[C]. New Orleans: IEEE Signal Processing Society,2017. 1 – 18.

[29] Tian Z, Huang W, He T, et al. Detecting text in natural image with connectionist text proposal network[A]. European Conference on Computer Vision[C]. Cham: Springer, 2016. 56 – 72.

[30] Ma J, Shao W, Ye H, et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia, 2018, 20(11): 3111 – 3122.

[31] Liu Y, Jin L. Deep matching prior network: toward tighter multi-oriented text detection[A]. IEEE Conference on Computer Vision and Pattern Recognition[C]. Hawaii: IEEE Computer Society, 2017. 3454 – 3461

[32] Liu Y, Jin L. Detecting curve text in the wild: new dataset and new solution[DB/OL]. arXiv: 1712. 02170v1, 2017.

[33] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Las Vegas: IEEE Computer Society, 2016. 770 – 778.

[34] Liu X, et al. FOTS: Fast oriented text spotting with a unified network[A]. IEEE Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City: IEEE Computer Society, 2018. 5676 – 5685.

[35] Gupta A, Vedaldi A, Zisserman A. synthetic data for text localisation in natural images[A]. IEEE Conference on Computer Vision and Pattern Recognition[C]. Las Vegas: IEEE Computer Society, 2016. 2315 – 2324.

[36] Redmon J, et al. You only look once: unified, real-time object detection[A]. IEEE Conference on Computer Vision and Pattern Recognition[C]. Las Vegas: IEEE Computer Society, 2016. 779 – 788.

[37] Liao M, Shi B, Bai X. TextBoxes + +: A single-shot oriented scene text detector[J]. IEEE Transactions on Image Processing, 2018, 27(8): 3676 – 3690.

[38] Shi B, Bai X, Belonge S. Detecting oriented text in natural images by linking segments[A]. IEEE Conference on Computer Vision and Pattern Recognition[C]. Hawaii: IEEE Computer Society, 2017. 3482 – 3490.

[39] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multi box detector[A]. European Conference on Computer Vision[C]. Cham: Springer, 2016. 21 – 37.

[40] Liao M H, Zhu Z, Shi B G, Xia G S, Bai X. Rotation-sensitive regression for oriented scene text detection[A]. IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City: IEEE Computer Society, 2018. 5909 – 5918.

[41] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[A]. IEEE Computer Vision and Pattern Recognition[C]. Boston: IEEE Computer Society, 2015. 3431 – 3440.

[42] Yao C, Bai X, Sang N, et al. Scene Text Detection Via Holistic, Multi-Channel Prediction [DB/OL]. arXiv: 1606. 09002v2, 2016.

[43] Polzounov A, Ablavatski A, Escalera S. Wordfence: Text detection in natural images with border awareness[A]. IEEE International Conference on Image Processing[C]. Beijing: IEEE Computer Society, 2017. 1222 – 1226.

[44] Xue C, Lu S, Zhan F. Accurate scene text detection through border semantics awareness and bootstrapping [A]. European Conference on Computer Vision[C]. Cham: Springer, 2018. 370 – 387.

[45] Long S, Ruan J, Zhang W, et al. TextSnake: A flexible representation for detecting text of arbitrary shapes[A]. European Conference on Computer Vision[C]. Cham: Springer, 2018. 19 – 35.

[46] Zhang Z, et al. Multi-oriented text detection with fully convolutional networks[A]. IEEE Computer Vision and Pattern Recognition[C]. Las Vegas: IEEE computer Society, 2016. 4159 – 4167.

[47] He T, Huang W, Qiao Y. Accurate Text Localization in Natural Image with Cascaded Convolutional Text Network[DB/OL]. arXiv: 1603. 09423v1, 2016.

[48] Tang Y, Wu X. Scene text detection and segmentation based on cascaded convolution neural networks[J]. IEEE Transactions on Image Processing, 2017, 26(3): 1509 – 1520.

[49] He D, Yang X, Liang C, et al. Multi-scale FCN with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild[A]. IEEE Conference on Computer Vision and Pattern Recognition[C]. Hawaii: IEEE Computer Society, 2017. 474 – 483.

[50] Deng D, Liu H, Li X, et al. PixelLink: detecting scene text via instance segmentation[A]. AAAI Conference on Artificial Intelligence[C]. New Orleands: the Association for the Advance of Artificial Intelligence, 2018.

[51] Zhou X, Yao C, Wen H, et al. EAST: An efficient and accurate scene text detector[A]. IEEE Computer Vision and Pattern Recognition[C]. Hawaii: IEEE Computer Society, 2017. 2642 – 2651.

[52] Hong S, Roh B, Kim K H, et al. PVANet: lightweight Deep Neural Networks for Real-Time Object Detection [DB/OL]. arXiv: 1611. 08588v2, 2016.

[53] He W, Zhang X Y, Yin F, et al. Deep Direct Regression for Multi-Oriented Scene Text Detection[DB/OL]. arXiv: 1703. 08289v1, 2017.

[54] Qin S, Manduchi R. Cascaded segmentation-detection networks for word-level text spotting[A]. Proceed of International Conference Document Analysis Recognion[C]. Kyoto: IEEE Computer Society, 2017. 1275 – 1282.

[55] Lyu P, Yao C, Wu W, et al. Multi-oriented scene text de-

tection via corner localization and region segmentation [A]. IEEE Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City:IEEE Computer Society, 2018. 7553 – 7563.

[56] Lyu P, Liao M, Yao C, et al. Mask text spotter: An end-to-end trainable neural network for spotting text with arbitrary shapes[A]. European Conference on Computer Vision[C]. Cham: Springer, 2018. 71 – 88.

[57] Ch'ng C K, Chan C S. Total-Text: A comprehensive dataset for scene text detection and recognition[A]. IAPR International Conference on Document Analysis and Recognition[C], Kyoto: IEEE Computer Society, 2017. 935 – 942.

[58] Karatzas D, Shafait F, et al. ICDAR 2013 Robust reading competition[A]. IEEE International Conference on Document Analysis and Recognition[C]. Washington DC: IEEE Computer Society, 2013. 1484 – 1493.

[59] ICDAR2013 场景文字竞赛数据集[DB/OL]. http://rrc. cvc. uab. es/? ch = 2&com = downloads, 2013.

[60] ICDAR2015 场景文字竞赛数据集[DB/OL]. http://rrc. cvc. uab. es/? ch = 4&com = downloads, 2015.

[61] MSRA-TD500 场景文字数据集[DB/OL]. http://www. iapr-tc11. org/mediawiki/index. php/MSRA_Text_Detection_500_Database_(MSRA-TD500), 2012.

[62] COCO-Text 场景文字数据集[DB/OL]. http://rrc. cvc. uab. es/? ch = 5&com = tasks, 2016.

[63] ICDAR 2017-MLT 多语言文字数据集[DB/OL]. http://rrc. cvc. uab. es/? ch = 8&com = downloads, 2017.

[64] RCTW 场景中英文字数据集[DB/OL]. http://mclab. eic. hust. edu. cn/icdar2017chinese/, 2017.

[65] CTW 场景中文数据集[DB/OL]. https://ctwdataset. github. io/, 2018.

[66] MTWI 网络图像文字数据集[DB/OL]. https://tianchi. aliyun. com/competition/entrance/231685/information, 2018.

[67] IIIT5K 文字识别数据集[DB/OL]. http://cvit. iiit. ac. in/projects/SceneTextUnderstanding/IIIT5K. html, 2012.

[68] Neocr: Natural environment ocr dataset[DB/OL]. http://www6. cs. fau. de/research/projects/pixtract/neocr/, 2011.

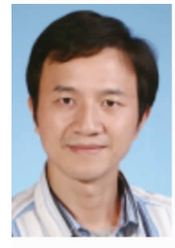
[69] Oriented Scene Text Dataset[DB/OL]. http://media-lab. engr. ccny. cuny. edu/ ~ cyi/, 2010.

[70] Multi-Orientation Scene Text Detection and USTB-SV1K Dataset[DB/OL], http://prir. ustb. edu. cn/TexStar/MOMV-text-detection/, 2014.

作者简介

姜　维　男,1981 年 12 月生,河南郑州人,现为华北水利水电大学讲师,主要研究领域为场景文字检测与识别.
E-mail:jiangwei@ ncwu. edu. cn

张重生(通讯作者)　男,1982 年 9 月生,河南南阳人,现为河南大学教授,主要研究领域为大数据分析、深度学习.
E-mail:chongsheng. zhang@ yahoo. com



殷绪成　男,1977 年生,湖南武冈人,现为北京科技大学计算机与通信工程学院教授,博导,副院长,主要研究领域为模式识别与计算机视觉、智能信息检索与数据挖掘、计算机网络及互联网信息智能处理、智能嵌入式软硬件系统及 AI 芯片.
E-mail:xuchengyin@ ustb. edu. cn