软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(5):1465−1496 [doi: 10.13328/j.cnki.jos.005988]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于深度学习的自然场景文本检测与识别综述*

王建新[1,2], 王子亚[1,2], 田 萱[1,2]

[1](北京林业大学 信息学院,北京 100083)

[2](国家林业草原林业智能信息处理工程技术研究中心(北京林业大学),北京 100083)

通讯作者: 田萱, E-mail: tianxuan@bjfu.edu.cn

摘 要: 自然场景文本检测与识别研究对于从场景中获取信息有重要意义,而深度学习技术有助于提高文本检测与识别的能力.主要对基于深度学习的自然场景文本检测与识别方法和其研究进展进行整理分类、分析和总结.首先论述自然场景文本检测与识别的相关研究背景及主要技术研究路线;然后,根据自然场景文本信息处理的不同阶段,进一步介绍文本检测模型、文本识别模型和端到端的文本识别模型,并阐述和分析每类模型方法的基本思路和优缺点;另外,列举了常见公共标准数据集以及性能评估指标和方法,并对不同模型相关实验结果进行了对比分析;最后总结基于深度学习的自然场景文本检测与识别技术面临的挑战和发展趋势.

关键词: 深度学习;自然场景;文本检测;文本识别;端到端

中图法分类号: TP391

中文引用格式: 王建新,王子亚,田萱.基于深度学习的自然场景文本检测与识别综述.软件学报,2020,31(5):1465−1496. http://www.jos.org.cn/1000-9825/5988.htm

英文引用格式: Wang JX, Wang ZY, Tian X. Review of natural scene text detection and recognition based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2020,31(5):1465−1496 (in Chinese). http://www.jos.org.cn/1000-9825/5988.htm

Review of Natural Scene Text Detection and Recognition Based on Deep Learning

WANG Jian-Xin[1,2], WANG Zi-Ya[1,2], TIAN Xuan[1,2]

[1](School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China)

[2](Engineering Research Center for Forestry-oriented Intelligent Information Processing of National Forestry and Grassland Administration (Beijing Forestry University), Beijing 100083, China)

Abstract: Natural scene text detection and recognition is important for obtaining information from scenes, and it can be improved by the help of deep learning. In this study, the deep learning-based methods of text detection and recognition in natural scenes are classified, analyzed, and summarized. Firstly, the research background of natural scene text detection and recognition and the main technical research routes are discussed. Then, according to different processing phases of natural scene text information processing, the text detection model, text recognition model and end-to-end text recognition model are further introduced, in which the basic ideas, advantages, and disadvantages of each method are also discussed and analyzed. Furthermore, the common standard datasets and performance evaluation indicators and functions are enumerated, and the experimental results of different models are compared and analyzed. Finally, the challenge and development trends of deep learning-based text detection and recognition in natural scenes are summarized.

Key words: deep learning; natural scene; text detection; text recognition; end-to-end

自然场景文本是指存在于任意自然情境下的文本内容,例如道路路牌、广告牌、商场指示牌、商品包装等.自然场景下的文本识别(scene text recognition,简称 STR)通常先利用文本检测技术得到文本位置信息,再使用文

* 基金项目: 国家重点研发计划(2018YFC1603302, 2018YFC1603305)
Foundation item: National Key Research and Development Program of China (2018YFC1603302, 2018YFC1603305)
收稿时间: 2019-06-09; 修改时间: 2019-07-28, 2019-11-08; 采用时间: 2019-12-03; jos 在线出版时间: 2020-04-07

本识别技术得到根据位置信息裁剪的图像中的文本内容.不同于文档图像中的文本规则性,自然场景文本通常在字体大小、字体类别、排列方向、字体颜色、文本稀疏程度就有很大的差异性,同时受到光照强度不同、复杂背景和拍照角度等因素的影响,自然场景文本检测与识别技术研究有很大的阻力.目前,传统的 OCR 技术无法适用于复杂自然场景图像中的文本识别.随着信息技术的发展和智能应用的需求不断增加,从自然场景图像中获取文本信息的技术研究具有广阔的应用前景,成为研究者关注的焦点.其中,文档分析和识别国际会议(Int'l Conf. on document analysis and recognition,简称 ICDAR)是推动该领域不断发展的重要国际会议,国内清华大学和中国科院自动化研究所曾在 2011 年共同举办了第 11 届文档分析和识别会议(ICDAR 2011).

随着深度学习技术的不断发展,基于深度学习的自然图像文本检测与识别已成为当前文档分析与识别领域的热点研究.深度神经网络本身所具备的很强的非线性拟合能力,理论上可以映射任意复杂的函数,具有很强的鲁棒性.因此,相对于传统的文本检测与识别方法,深度神经网络能够解决复杂自然场景下的文本图像到文本位置和文本内容的映射问题.虽然文献[1,2]对文本检测的研究与发展做了较为系统的阐述和总结,但没有对基于深度学习的文本检测与识别领域的相关技术进行全面的综述,因此,本文将系统地综述该领域目前的技术发展状况以及挑战,希望为研究者提供一定参考和帮助.本文在对相关文献进行了整理和总结后,将基于深度学习的自然场景文本检测与识别根据模型功能分为文本检测方法、文本识别方法和端到端的识别方法,如图 1 所示,每类方法根据实现技术特点的不同又分为多种类别,后续将详细介绍和分析这些模型方法.

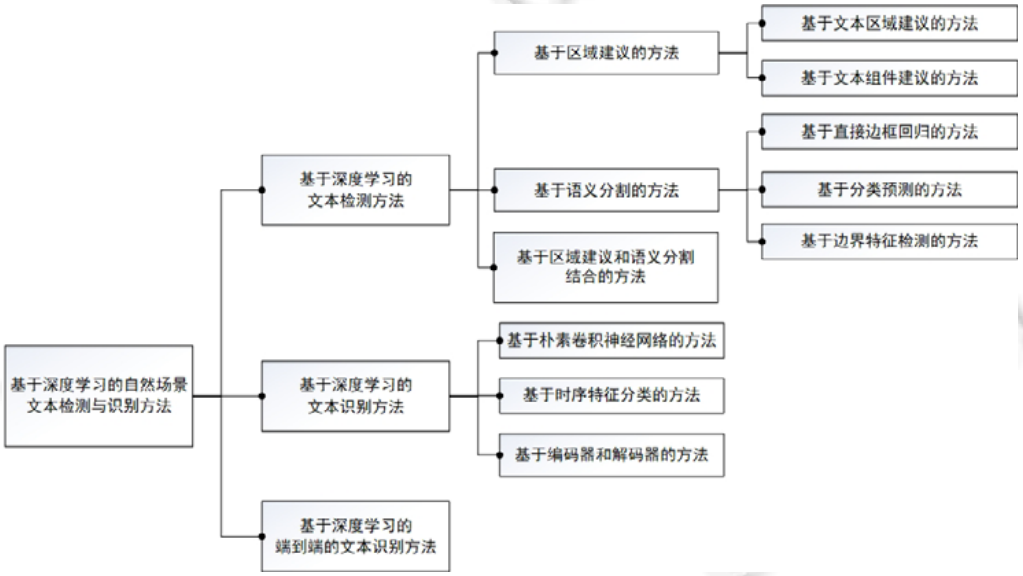


Fig.1 A taxonomy of natural scene text detection and recognition methods based on deep learning

图 1 基于深度学习的自然场景文本检测与识别方法分类

本文第 1 节介绍自然场景文本检测与识别相关背景和早期研究.第 2 节对基于深度学习的自然场景文本检测方法进行分析和总结.第 3 节对基于深度学习的自然场景文本识别方法进行分析和总结.第 4 节介绍基于深度学习的端到端的文本识别方法模型结构.第 5 节介绍该领域相关公共数据集和评估方法,并对基于深度学习的自然场景文本检测与识别方法的实验结果进行分析与比较.第 6 分析自然场景文本检测与识别的技术发展与挑战.第 7 节是全文的总结.

1 相关背景及早期研究介绍

自然场景的文本识别是指在无约束的环境中处理图像文本信息.当前研究通常将自然场景的文本识别分为两个步骤:文本检测(text detection)和文本识别(text recognition),即采用视觉处理技术来提取图像中文本实例

和借助自然语言处理技术来获得其中的文字内容.显然,这两个步骤紧密相关,文本检测结果的准确性直接影响最终的文本识别结果.自然场景图像的文本信息检测与识别技术有助于场景内容信息的获取、分析、理解,对于提高图像检索能力、工业自动化水平、场景理解能力等具有重要意义,可应用于自动驾驶、车牌票据识别、智能机器人、图片检索、大数据产业等场景.目前,自然场景文本检测与识别已成为计算机视觉与模式识别、文档分析与识别领域的研究热点.

2010 年,Neumann 等人[3]提出将 MSERs(maximally stable extremal regions)方法应用于自然场景文本检测,通过对图像中的一些最大稳定极值区域的检测来获得文本候选区域.2012 年,Wang 等人[4]借鉴传统的目标检测思想,提出了一个端到端文本识别模型.他们先通过滑动窗口检测字符区域,再利用字符置信度得到字符内容,最后利用字符之间的空间约束关系得到文本单词内容.传统的文本检测和识别通常需要手工设计复杂的特征和分类器以及后处理流程来检测和识别图像文本内容,这些技术难以满足复杂的自然场景文本识别需求.

随着深度学习(deep learning,简称 DL)[5]技术的不断发展和人工神经网络(artificial neural networks)研究的不断进步,越来越多的研究者使用深度学习模型替代传统的图像文本识别算法并取得了不错的成果.深度学习的优势在于其模型具有层次性并且参数比较多,通过组合低层特征形成更加抽象的高层表示属性类别或特征,因而具有更好的数据特征表示和更高的识别准确率.

深度神经网络主要分为卷积神经网络(convolutional neural networks,简称 CNN)和循环神经网络(recurrent neural networks,简称 RNN).CNN 利用卷积运算(convolutional layer)提取类似网格结构的数据特征,通过下采样(pooling layer)保留主要特征防止过拟合,并且减少参数和计算量,同时使得提取的特征具有旋转和平移不变性.由于 CNN 的特点,其在分析处理图形图像数据研究领域具有广泛的应用.RNN 最大的特点在于网络隐含层的输入包含当前时间点该隐含层的输入和上一时间点该隐含层的输出,利用在不同时间点共享参数使网络具有关于时间向后的连接,可以学习具有时序性的数据特征和规则.当前,主流的循环神经结构包含长短时记忆结构(long short-term memory,简称 LSTM)[6]、门循环单元(gated recurrent unit)[7]和双向长短时记忆(bidirectional long short-term memory,简称 BiLSTM).RNN 主要用于处理语音、视频、文本等具有时序性特征数据,广泛应用在视频分类、在线翻译、语音识别等领域.

深度学习技术的不断发展,使得越来越多的研究者将基于CNN的图像特征提取技术和基于RNN的自然语言处理技术应用到自然场景文本检测与识别领域,推动着该领域的不断前进和发展.本文将对基于深度学习的自然场景文本检测与识别方法和技术进行总结归类,分析和阐述当前深度学习技术在自然场景文本检测与识别中的应用,并概括深度学习技术在该领域研究面临的挑战,为未来研究者更好的解决自然场景中的文本检测与识别提供参考或帮助.

2 基于深度学习的自然场景文本检测方法

目前,自然场景文本检测是识别的必要步骤,先对图像中的文本进行定位检测获得文本区域的图像,便于后续的文本识别.自然场景文本检测是标注场景图像中每个文本实例区域坐标位置的过程,即获得图像中不同单词或者文本行区域的过程.20 世纪 90 年代起,研究者利用传统的计算机视觉技术在自然场景文本检测领域取得一定的成果,其中最具代表性方法包括:基于连通域分析的笔画宽度变换(stroke width transformation,简称 SWT)[8]以及最大稳定极值区域(MSREs)[9]算法和基于滑动窗口的区域特征分类方法[10,11].然而,受制于自然场景文本图像中文本和背景的复杂性以及图像噪声干扰等因素,基于这些算法仍然难以高效准确检测自然场景中的文本实例.

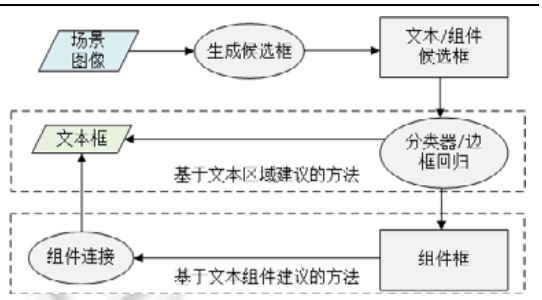
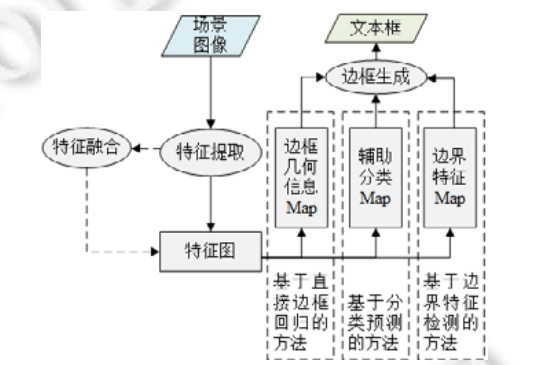
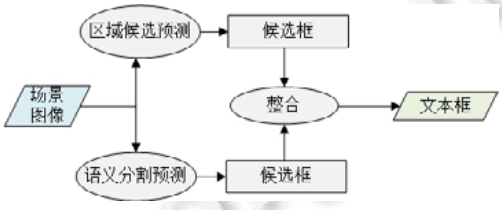
近 10 年来,计算机视觉领域的主要研究领域之一目标检测(object detection)迅猛发展,文本检测该领域目前已经涌现出一系列基于深度学习的目标检测算法.文本检测作为目标检测研究内容的特定领域,许多研究者借鉴了目标检测的技术思路,将深度学习技术和文本检测相结合,提取自然场景文本的复杂特征,通过深度神经网络模型检测自然场景文本实例.

根据文本检测模型的技术实现特点,本文将基于深度学习的自然场景文本检测方法分为 3 类(见表 1):基于

区域建议(region proposal)的自然场景文本检测方法、基于语义分割(semantic segment)的自然场景文本检测方法、基于区域建议和语义分割结合的自然场景文本检测方法.下面将对这 3 类算法的特点、关键技术和主要优缺点进行分析介绍.

Table 1 Natural scene text detection methods based on deep learning

表 1 基于深度学习的自然场景文本检测方法

方法类别	方法特点	代表算法	方法流程
基于区域建议的方法	基于文本区域建议的方法根据候选文本框的特征筛选文本候选框并微调文本候选框大小及位置	TextBoxes++[12] RRD[13]	
	基于文本组件建议的方法根据组件候选框的特征分类和回归得到文本组件框,再将组件框连接文本文本框	CTPN[14] SegLink[15]	
基于语义分割的方法	基于直接边框回归的方法利用图像全局特征直接预测每个像素或者栅格所在的文本框的几何信息	EAST[16] AF-RPN[17]	
	基于分类预测的方法利用图像全局特征预测每个像素或栅格位置的多任务分类结果	PixelLink[18] PSENet[19]	
	基于边界特征检测的方法利用边界和文本区域的特征关系检测文本框	TextField[20] TextMountain[21]	
基于区域建议和语义分割结合的方法	通过多模型集成提高检测准确率,结合区域建议预测和语义分割预测的结果得到更准确的文本检测结果	FTSN[22] PixelAnchor[23]	

2.1 基于区域建议的自然场景文本检测方法

该类检测算法主要通过区域建议网络(region proposal network,简称 RPN)、候选区域生成算法或其他方式预定义不同尺寸的多个候选框,并利用 CNN 网络提取的候选区域(region of interest,简称 ROI)图像特征判断该区域是否属于文本实例.另外,由于文本框检测精确性影响文本识别率,大多数该类方法使用边框回归(bounding box regression)校正原来的建议区域,生成更准确的文本框坐标.这类基于文本区域建议的方法通常建立在经典目标检测算法基础上,如 RCNN(region proposal cnn)[24]、Faster-RCNN[25]、SSD(single shot multibox detector)[26]、R-FCN(region-based fully convolutional networks)[27]和 YOLO(you only look once)[28]等,具有深度神经网络提取图像特征的优点,有较好的检测效果.根据该类算法候选框粒度的不同,我们将该类算法分为基于文本区域建议的方法和基于文本组件建议的方法.本节将对这两类算法特点、关键技术、主要优缺点进行分析介绍.

2.1.1 基于文本区域建议的方法

传统的文本检测算法通常利用边缘梯度、方向梯度直方图(histograms of oriented gradients)、局部二值化(local binary pattern)等手工设计的特征,将候选区域区分为文本区域和非文本区域.通常,该类手工设计特征无

法准确描述自然场景中的复杂文本域.受到目标检测算法的启发,许多研究者将基于深度学习模型的目标检测方法用到文本检测中,利用卷积神经网络生成文本候选区域,并根据候选区域的特征筛选文本候选区域,这类算法称为基于文本区域建议的方法.

文献[29]受到目标检测 YOLO[28]算法的启发,首次提出了一种基于区域建议的全卷积回归网络(fully convolutional regression network,简称 FCRN)模型,模型先根据全卷积网络提取图像的特征图,再对特征图进行卷积操作(不包含激活函数)回归预测每个栅格位置所属文本区域的中心坐标偏移、宽高和角度信息.与 YOLO 网络中最后使用全连接层提取包含全局信息特征的方法不同,为了满足文本分类检测特征的局部平移不变性,FCRN 通过卷积提取图像特征,每个栅格位置的特征仅包含栅格位置附近的信息.该文献另一个主要贡献是基于深度学习和分割技术将文本嵌入自然场景图像中生成合成数据,并用于文本检测模型的训练中,降低了收集数据集的工作量.

文献[30]针对不同尺寸比例的文本检测提出一种 TextBoxes 的网络结构,如图 2 所示.该网络结构基于 SSD[26]模型,可根据不同卷积层的多尺度特征有效检测不同尺寸文本,并可根据文本区域的纵横比特点设定 6 种不同纵横比默认的文本候选框.该文献还将多种比例缩放的图片作为检测的输入数据来提高不同大小文本的检测准确率.与文献[30]中使用规则的矩形表示文本框不同,文献[31]首次提出了基于四边形的文本实例检测模型.该模型在 SSD[26]的基础上增加了 6 种不同倾斜角度的候选文本框,利用四边形的 4 个顶点坐标来表示文本候选框.针对四边形 IoU(intersection over union)难以快速计算的难点,该文献提出了基于采样策略的共享蒙特卡洛方法,提高了 IoU 计算速度.

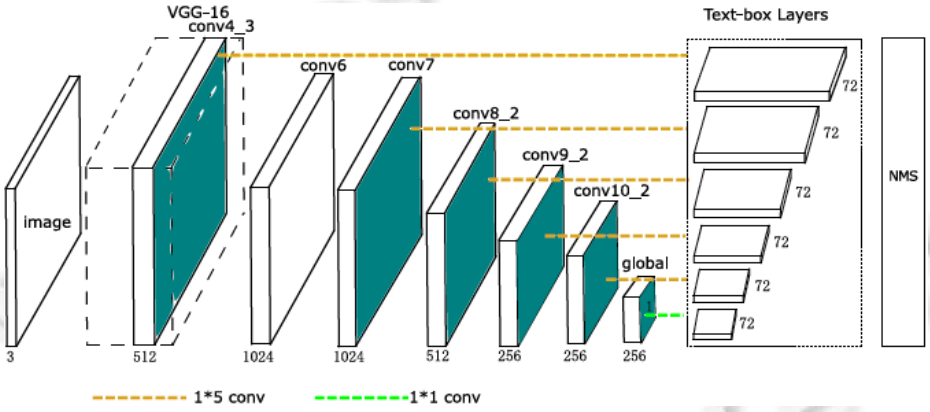


Fig.2 Architecture of TextBoxes[30]

图 2 TextBoxes 模型结构[30]

针对文献[30,31]无法很好地检测自然场景中倾斜率较大的文本实例的问题,文献[32]沿用 Faster-RCNN[25]中 RPN 生成候选区域的思想,提出旋转候选框(rotation region proposal)、旋转兴趣区域池化(rotation region-of-interest pooling)和倾斜候选框的 IoU 计算方法,通过增加旋转角度参数来控制候选框的方向,对旋转候选框的边框回归,增强对于倾斜文本的检测效果.文献[33]同样基于 Faster-RCNN[25]提出了任意方向的场景文本检测模型,与文献[32]不同的是,该模型不使用角度表示文本框的方向,而是使用顺时针方向的前两个点的坐标和边界框的高度来表示不同方向的文本框.文献[12]则在文献[30]的基础上提出了一种能够检测自然场景中任意方向文本的 TextBoxes++文本检测模型,该模型利用四边形文本框或者包含倾斜角度的文本框实现不规则形状的文本检测.

相较于其他文献中利用矩形或者四边形来表示文本框,文献[34]提出种基于滑动线点回归的文本检测模型,该模型利用多条水平或垂直等距线划分矩形文本候选区域,并且回归预测文本边界与水平等距线交点的 x 轴坐标以及文本区域与垂直等距线交点的 y 轴坐标,最后根据预测的矩形文本区域坐标和等距线与文本边界交

点坐标构建文本框.由于这种方法增加了多个文本边界坐标的回归预测,使得该方法能够更精确地检测文本.

文献[13,35]则主要通过改进网络结构优化文本检测模型,以此提高文本检测准确率.其中,文献[35]参考GoogleNet[36]网络中的 Inception 架构,利用分层 Inception 模块(hierarchical inception module,简称 HIM)来融合卷积特征.该模型还包括文本注意力模块(text attention module,简称 TAM),TAM 将学习得到的像素级文本概率特征图与不同层 HIM 模块提取的特征结合得到的 AIF(aggregated Inception features)进行数量积运算,以增强文本区域特征,降低图像背景对于文本检测的干扰.由于文本区域分类任务是旋转不敏感的,而边框回归任务却相反,文献[13]认为这两类不兼容的任务共享特征会导致检测性能的下降,提出了旋转敏感的回归检测模型(rotation-sensitive regression detector,简称 RRD).该模型包含边框回归分支和文本区域分类分支,其中,边框回归分支利用主动旋转卷积提取旋转敏感特征增强分支预测准确率,文本分支通过池化融合旋转敏感特征预测文本区域的分类.

场景文本检测方法往往依赖于手动设计不同比例尺寸的候选框,然后通过边框回归调整候选框.文献[37]中提出一种新颖的自适应候选文本区域建议模型,该模型通过角点检测获取文本区域不同位置的候选角点,并预测每个角点的匹配方向,通过匹配候选角点得到四边形文本区域候选框,最后利用边框回归得到更准确的文本框.同时,为了增强模型检测不同方向文本的鲁棒性且不增加网络模型和训练数据的规模,该模型在网络结构中增加了双兴趣区域池化模块(dual-RoI pooling module),通过融合不同方向的池化特征,提高模型的检测性能.

2.1.2 基于文本组件建议的方法

目前,基于文本区域建议的自然场景文本检测方法大多数源于对目标检测算法的改进,难以避免该类算法带来的缺点,即自然场景中的文本形状各异,尺寸和宽高比差异大,预先设定的候选框尺寸大小无法近似匹配真实文本实例,而边框回归只能对候选框的位置进行微小的调整,导致该类算法在自然场景文本检测中难以取得较好的效果.基于文本组件建议的方法思路是将文本区域视为一组连续的文本组件,其中组件为字符或者文本的一部分,再利用区域建议方法检测文本组件区域,最后将文本组件连接为文本区域以达到文本检测的目标.

文献[38]在卷积神经网络的基础上提出一种文本注意力卷积神经网络(text-attentional CNN),通过从候选文本组件区域提取卷积特征代替手工设计特征的方式来提高候选文本组件的分类准确率.该网络共享低层卷积特征,通过包括字符区域分割、字符分类、文本及非文本二分类的多任务监督学习训练字符检测模型,最后,将模型检测的文本字符按照字符的高度、水平位置、纵横比等特征将字符连接为单词.但由于这种滑动窗口的方式需要滑动扫描全部图像,通常检测速度较慢.

与文献[38]中使用滑动窗口和卷积网络检测字符区域不同,文献[14]利用微分的思想,将文本区域视为由多个字符或者字符的一部分构成的文本组件序列,再结合目标检测方法检测文本组件.该方法将候选框设为固定宽度、不同高度的垂直组件候选框,先通过预测候选框的垂直和水平位置偏移检测小的文本区域组件,再通过后处理将一系列文本组件连接起来.该方法最大的特点是将 RNN 引入到文本检测中,利用其处理序列数据的优点检测文本组件序列,该方法能够克服任意长度文本的检测难点,但是缺点也很明显:其检测水平文本组件序列的方法只能检测水平或者微斜的文本区域,并且由于模型嵌入 BiLSTM 结构导致网络参数大量增加,降低了检测效率.

文献[15]在 CTPN 模型[14]的基础上提出了面向任意文本方向的分段链接检测模型(segment linking,简称 SegLink),如图 3 所示,其主要思想是将文本检测分解为分段检测和分段链接两个部分.不同于文献[14]中利用 RNN 检测文本组件,该模型在 VGG16[39]网络结构的基础上提取不同的高维特征和低维特征,使用高维和低维特征同时检测不同尺寸的文本分段,对大尺度文本和小尺度文本检测有很好的鲁棒性.为了克服 CTPN 模型[14]无法检测倾斜文本的缺点,该模型通过预测分段 8 个方向是否有与其他分段连接,使预测分段可以链接生成任意方向的文本框.对于密集的文本区域,SegLink 预测文本分段是否链接的方式的仍然可能无法正确区分单独的文本实例,文献[40]在 SegLink 的基础上改进了文本组件的分组预测方式,通过学习文本分段之间的吸引与排斥系数来分组文本分段,使得文本分段之间的分组更加准确,易于区分排列紧凑的文本,并且能够很好地识别不规则曲线文本.

大多数基于文本区域建议或者文本组件建议的方法局限于其实现特点,往往只能检测固定矩形或者四边形区域的文本域.文献[41]认为,如果检测出每个字符并且可以找到将它们分组生成文本的方式,则可以很容易根据字符的边框得到文本的边框.因此,该文献将这种思路应用到自然场景中变形扭曲的不规则文本检测中,提出一种字符嵌入网络(character embedding network,简称 CENet),模型包含两个学习任务:基于字符候选的字符区域检测任务和基于字符嵌入的嵌入向量聚类任务.先由字符区域检测任务检测图像的字符组件,再由字符嵌入向量聚类任务通过模型的嵌入网络将字符特征嵌入到向量空间中,并学习每个嵌入向量之间的匹配关系,即两个字符是否属于同一单词文本.为了减少运算,该模型采用基于距离半径为 r 的 K 近邻算法计算得到字符聚合后的文本框.

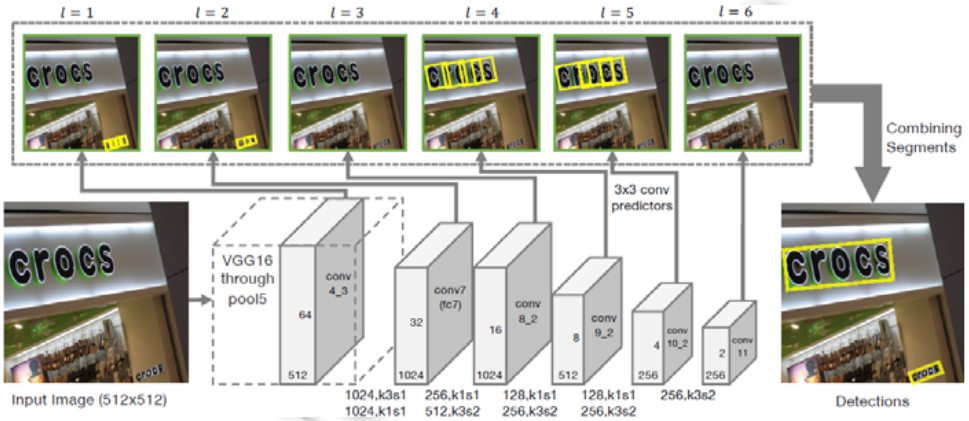


Fig.3 Architecture of SegLink[15]

图 3 SegLink 模型结构[15]

2.2 基于语义分割的自然场景文本检测方法

基于 Faster-RCNN[25]和 SSD[26]的方法通过预测不同候选框对自然场景文本检测不够有效,为了使得候选框能够很好地包含文本内容,从而需要手工设计各种尺寸、比例、倾斜度的候选框,很难使候选框与真值框之间具备较好的匹配关系,而且训练过程候选框的匹配预处理以及采样导致检测速度较慢.基于语义分割的自然场景文本检测方法通过借鉴经典语义分割算法的思路,例如 FCN(fully convolutional networks)[42],FPN(feature pyramid networks for object detection)[43]和 FCIS(fully convolutional instance-aware)[44]等,利用深度卷积和上采样提取多级融合特征,预测图像中的每个像素是否属于文本区域.为了准确区分不同的文本实例,该类算法通常需要其他辅助预测结果来分割文本实例.根据该类算法的辅助预测任务的不同,我们将该类算法分为 3 种:基于直接边框回归的方法、基于分类预测的方法和基于边界特征检测的方法,并根据算法的特点、关键技术、主要优缺点对这 3 种算法进行总结归纳.

2.2.1 基于直接边框回归的方法

图像语义分割技术的思路是:利用深度神经网络提取包含图像丰富信息的特征图,并根据特征图来预测目标分割结果[45].基于直接边框回归方法通过卷积网络提取图像特征图,根据特征预测每个像素或者栅格位置是否为文本和直接边框回归得到该位置所属文本框的参数信息,如文本框顶点坐标、方向角度或其他文本实例区域表征参数,而不是回归预测真值框与候选框表征参数的偏移量.

为了克服基于区域建议方法利用候选框间接回归检测文本框耗时的缺点,文献[16,46,47]根据卷积网络模型输出的特征图检测图像中的文本框,其中,文献[16,46]提出了端到端的任意四边形文本检测模型,如图 4 所示.这两种模型均利用上采样融合不同层的特征得到预测特征图,低层的特征语义信息比较少,但是目标位置准确,有利于检测小尺度的文本框,高层的特征语义信息比较丰富.但是目标位置比较粗略,有利于检测大尺度的文本框.文献[47]基于文本识别模型引入注意力机制,将直接边框回归检测得到的文本图像输入到基于时序注意力

机制的检测优化模块中,根据文本区域的序列特征剔除检测结果中假性文本区域,以此提高文本检测的准确率.

文献[17]将直接边框回归与 Faster-RCNN[25]目标检测网络结合,提出了无锚的区域建议网络(anchor free region proposal network,检测 AF-RPN),该模型使用直接边框回归代替 RPN 中手动设计文本候选框的方式,其网络结构借鉴 FPN[43]模型的思想,在提取的不同层级金字塔特征上分配单独的检测模块,而不是融合多层金字塔特征后再检测文本.通常,文本检测模型将文本框内的所有像素都标记为文本,当文本内容不够紧凑时,周围背景噪声会对文本检测产生影响,该方法通过收缩文本区域,生成文本核心区域,只有核心区域的文本才被标记为正样本.

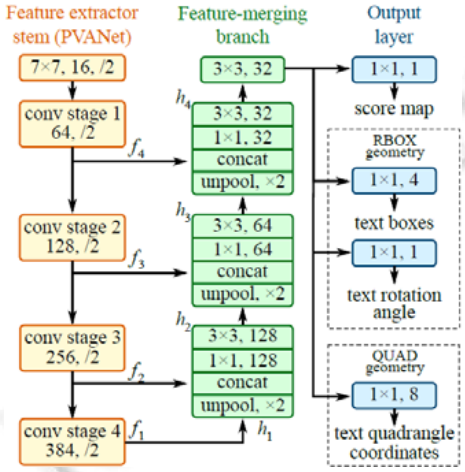


Fig.4 Structure of EAST[16]

图 4 EAST 模型结构[16]

针对直接边框回归的方法在检测长文单词或者文本行时由于其感受野的限制导致其定位准确率不高、文本框短边的回归误差较大的问题,文献[48]在直接边框回归任务的基础上增加了文本框边界的学习任务,通过网络模型输出矩形文本框的长短边 4 个边界概率图,再利用长边提取文本行,将文本行区域像素的直接边框回归结果得到文本框的短边,最后将 4 个边界合并得到文本框.

大多数文本检测模型都基于文本实例形状是线性的这个前提,采用简单的旋转矩形或者四边形来表示文本实例,这使得模型在检测不规则形状的文本时效果很差.文献[49]借鉴文本组件建议方法的先检测组件再连接组件的思路,提出一种非常灵活的文本实例表征方式.该方式利用一系列连接且重叠的圆盘来表示文本区域,每个圆盘的圆心在文本区域中心线上,圆盘由参数半径和方向角度表示,半径为文本实例局部宽度的一半,方向角度为圆心所在文本中心线位置的切线角度.该模型利用多级融合特征分类预测每个像素位置是否为文本区域和是否为文本中心线,并直接回归预测每个位置的圆盘半径和角度,根据这些预测结果,很容易重新构建得到文本实例预测结果.该方法能够实现对线性文本和不规则文本的检测,而且对于很长的文本实例也有较好的检测效果.

2.2.2 基于分类预测的方法

无论是基于文本区域建议还是基于直接边框回归的方法都是通过文本框分类和边框回归任务检测文本实例,基于分类预测的文本检测方法依据语义分割的核心思想,通过训练文本、非文本或其他辅助分类任务实现自然图像中的文本检测.该类方法的特点是仅包含分类任务,而分类任务相较于回归任务更易于训练学习.

文献[50]首次将基于语义分割的文本像素分类预测引入到文本检测中,并提出了基于全卷积网络的多方向文本检测方法,该方法首先利用 FCN 网络提取高低维度的融合特征预测得到文本区域分割图,再使用 MSERs 算法提取文本分割区域的候选字符,并根据后处理算法生成连接候选字符生成文本行.相较于传统方法直接在文本图像上通过 MSERs 提取候选字符的方式,先利用 FCN 分割文本区域能够提出大量的背景噪声.为了进一

步过滤错误预测结果,该方法增加了文本行质心的预测网络,利用角度以及文本行中质心平均得分筛选得到更加准确的文本行.

仅利用文本二分类预测任务很难将距离很小的多个文本实例分开,但每个文本区域都有不同的文本中心线特征,并且通常不同文本区域的文本中心线不会重叠.因此,文献[51]提出了基于文本中心线区域预测的文本检测模型.该方法将文本区域分为文本区域和文本中心线区域,文本中心线区域的宽度为文本宽度的一半.该模型分两个检测步骤:首先,利用多层卷积后的低分辨率高层特征粗略预测图像中的文本区域,由于高层特征的位置不敏感性,预测结果只是区分图像中的文本和背景,没有精确区分不同的文本实例;然后,再将检测的文本区域裁剪调整到固定大小输入精细检测网络中预测文本区域和文本中心线区域.最后,根据预测结果得到文本框.这种先粗略估计文本区域再精确检测分割区域文本的方式对不同尺度的文本检测有较好的鲁棒性.

文献[52,53]提出了基于文本边界检测的文本检测模型,模型通过引入边界类来区分不同的文本实例,模型将图像像素分为 3 类:文本、非文本、边界.为了增加边界检测的鲁棒性,文献[52,53]将文本实例边界一定宽度的带状区域均标记为文本边界,利用 FCN[42]网络以及上采样提取图像特征信息,预测每个像素位置的分类置信度,根据预测结果利用连通分量分析得到不同的文本区域.最后,通过后处理算法得到矩形文本框.文献[53]还通过基于文本实例像素数的权重归一化来优化损失函数来,使得小尺度的文本区域预测损失权重更大,有利于小尺度文本的检测.

文本实例区域像素在图像空间特征上具备连通的特点,通过将连通像素链接在一起来可以分割出文本实例,然后直接根据分割结果得到文本边界框而不再需要进行位置回归.文献[18]根据这种思路提出了基于像素链接的文本检测模型,如图 5 所示.此模型在预测像素文本分类的同时,还预测像素位置的 8 个方向是否连通,文本像素点在不同方向上的连通性可以区分不同的文本实例.根据预测结果,该模型利用并查集数据结构生成文本实例像素集,最后基于 OpenCV 的 minAreaRect 连通域算法生成最小外接矩形,从而得到文本框.

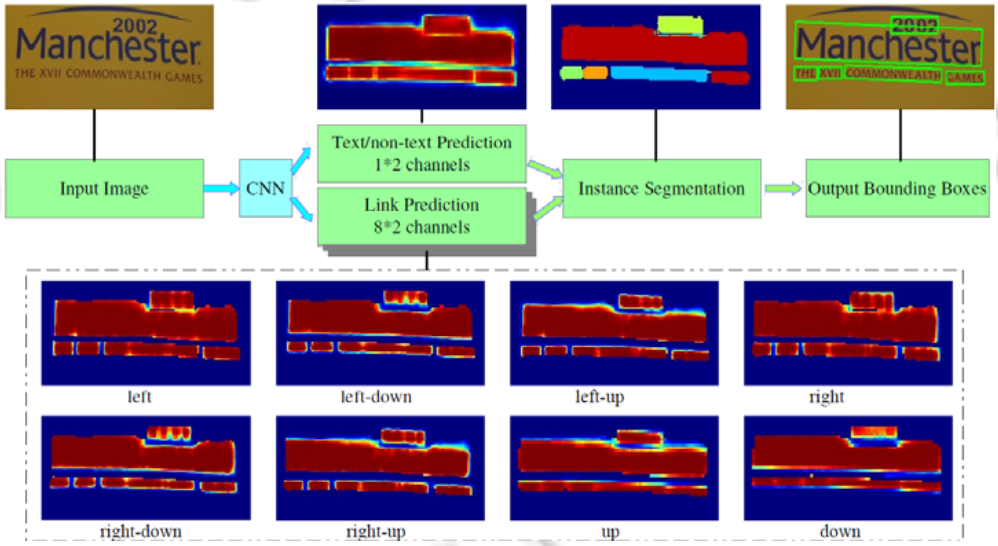


Fig.5 Architecture of PixelLink[18]

图 5 PixelLink 模型结构[18]

虽然不同的文本实例之间的距离可能很小,但是不同文本实例的中心区域通常有很明显的区分边界.文献[19]基于该思想提出一种对文本实例实现多级预测的渐进式扩展模型.对于每个文本实例,通过收缩法将文本缩小至不同的尺度生成多级中心区域训练标注,利用文本分类和多级中心区域的预测结果,采用基于广度优先搜索的尺度扩展算法逐渐生成不同的文本区域.相较于仅采用文本分类的预测方法,该渐进式扩展模型提升了

文本检测准确率和召回率.

2.2.3 基于边界特征检测的方法

由于图像中不同文本实例存在紧凑、距离较小的情况,仅预测文本分类可能会将两个不同的文本区域视为一个文本实例.由于不同文本实例的边界与其中心区域的关系特征可以区分不同的文本实例,基于边界特征检测的方法基于该思想将文本区域标记为中心区域和具有一定宽度的边界区域,利用文本实例的边界和中心区域像素之间的方向位置等关系特征训练文本检测模型.

为了充分挖掘利用文本实例与边界的关系特征,文献[20]利用图像中文本方向场来表示图像文本特征,提出了深度方向场不规则文本检测方法(如图 6 所示).文本区域的方向场由与每个像素距离最近的文本边界像素指向此像素的二维单位向量表示,非文本区域的方向场由二维零向量表示,整个图像向量场的大小可以区分背景区域和文本区域,这种方向场适合于描述任意形状的文本.该模型根据文本向量场预测并利用其后处理算法构造森林数据结构来得到每个文本实例的代表像素,最后根据这些代表像素扩展得到不同的文本实例.

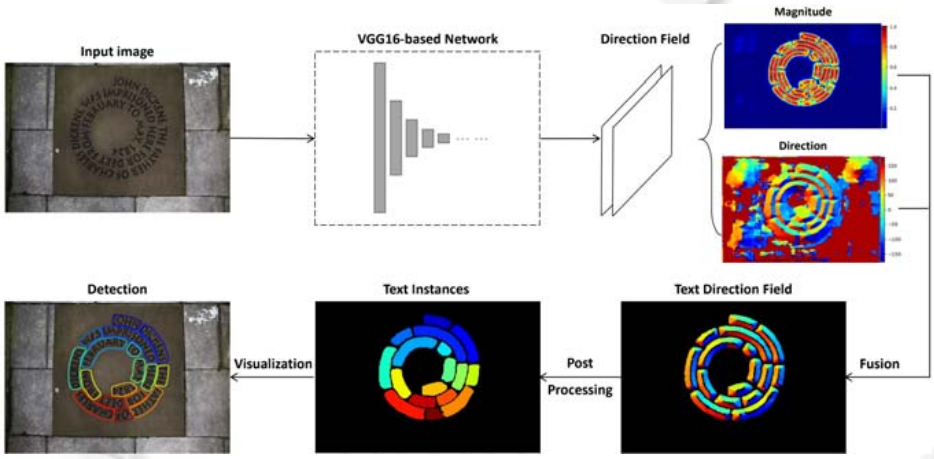


Fig.6 Text detection pipeline of TextField[20]

图 6 TextField 文本检测流程图[20]

对于图像中的文本,通常很难严格地区分文本边界和文本中心区域.文献[21]将文本中心和边界分类视为概率模型,像素越接近中心,则概率值越高;并且将文本区域每条边界上的一点到像素点的垂直的向量加权平均作为该像素点的中心方向向量,像素属于文本中心的概率越高,则向量的模越小.由于两个距离紧凑的文本实例相邻区域的文本中心边界具有相同的概率,预测文本中心方向可以很容易区分不同的文本实例.该方法的检测准确率优于大部分文本检测方法.

文献[54]设计了一种新颖的多尺度形状回归预测模型,通过预测结果得到密集的文本实例区域边界.该模型与其他方法直接边框回归预测文本框的顶点坐标或者宽高不同,模型增加了文本中心区域像素与距其最近的边界点的水平和垂直方向上的距离预测,即边界点的相对坐标预测,最后根据该预测得到用于生成文本框的文本实例区域边界点.该方法通过对密集边界点的检测,不仅避免了多边形稀疏顶点回归预测在处理长文本时遇到的回归不准确的问题,而且能够精确定位任意方法和形状的文本.

2.3 基于区域建议和语义分割结合的方法

基于区域建议和语义分割结合的方法将两类预测方式融合到统一的检测框架中,该类方法通过构建并结合多分类器来完文本预测,将多个检测任务进行结合,通常可以获得比单一分类器更优越的泛化性能和准确率.该类方法通常包含两个主要预测模块:基于语义分割的文本区域检测和基于区域建议的文本框预测,最后结合两类分支的预测结果得到文本框.

传统的基于连通域分析的文本检测方法需要检测整个图像上的候选字符区域,为了降低背景内容对于文

本检测影响,文献[55]先利用基于 MSERs[9]的算法检测得到文本框,再利用文本显著图筛选得到高得分的文本框.不同于文献[55]使用传统算法检测图像中文本候选区域,文献[56]提出了基于 Faster-RCNN 的字符检测和文本语义分割的双任务的检测模型 DSTD(deep scene text detection),如图 7 所示.该模型中,文本像素分割预测和字符候选框检测子网共享基础网络提取的卷积特征,文本像素分割分支通过训练预测图像中每个像素文本与非文本分类以及文本区域的中心线来生成文本框,再结合字符预测结果和文本框,仅保留含有字符的文本框.

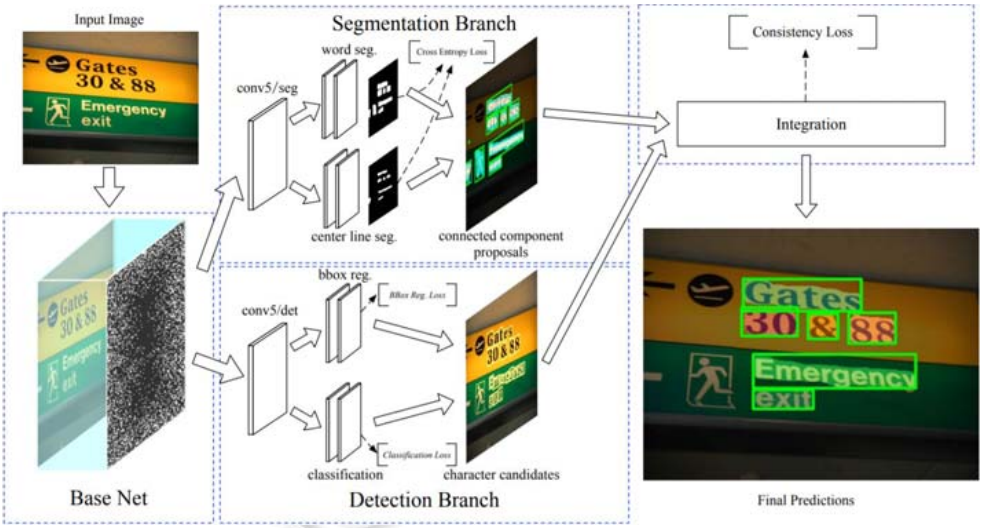


Fig.7 Architecture of DSTD[56]

图 7 DSTD 模型结构[56]

文献[23]将文献[16]和文献[12]的两种方法结合到统一的网络模型中,该模型包含基于像素预测的模块和基于候选框预测的模块.其中,基于像素预测的模块在文献[16]的基础上增加了空洞金字塔池化结构(atrous spatial pyramid pooling),输出像素点距离文本边框的 4 个距离值和文本注意热度图;基于候选框预测的模块将文本注意热度图馈送到模块中增强文本区域特征,并在文献[12]的基础上增加自适应预测层来更好检测尺度和宽高比差异大的文本.最后,该模型结合两个模块的预测结果得到检测的文本框.

通过预设候选框来回归预测文本框位置的方式具有一定的盲目性,候选框存在的大量负样本一定程度上降低了预测的准确率.与文献[37]类似,文献[57]从角点检测的角度考虑,提出了基于角点定位和区域分割的文本检测模型.该模型角点定位分支将候选框的中心视为文本区域的角点,通过分类和回归预测得到角点框以及其所属的位置(即左上、左下、右上、右下)类型,再将角点分组连接得到若干候选文本框;区域分割分支利用位置敏感分割网络得到文本区域像素属于不同位置区域的置信度,用于筛选候选文本框过滤噪声框.

为了解决卷积网络特征的位置不敏感性,文献[22]将实例分割模型 FCIS[44]的位置敏感网络模型和文本检测结合来提高文本检测的准确率,提出了一种融合文本分割网络(fused text segmentation networks,简称 FTSN).该模型由文本区域建议网络和位置敏感文本实例预测网络,其中,位置敏感网络提取文本分类特征和边框回归特征,文本分类特征包含文本像素属于前景和背景的得分以及像素属于文本框 k^2 块不同区域的得分,边框回归特征包含文本 k^2 块不同区域的坐标位置特征.最后,模型利用兴趣区域位置敏感池化(position-sensitive ROI-pooling,简称 PSROIPooling)特征预测得到文本框的分类得分和坐标以及文本区域分割图,并结合预测结果得到文本框.

文献[58]在文献[22]的基础上,将 GoogleNet[36]的 Inception 结构融合到网络中,以提取更丰富的图像特征.此外,为了克服兴趣区域位置敏感池化只能提取水平文本特征的缺点,该模型使用可变形卷积代替普通卷积,并使用可变形 PSROI 池化代替 PSROI 池化.由于可变形卷积、池化增加了方向参数,感受野能够自适应不规则兴

趣区域,使得该方法能够更好地提取不规则文本的特征,并进一步提高了文本检测性能.

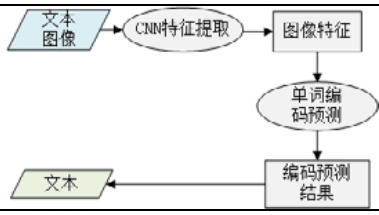
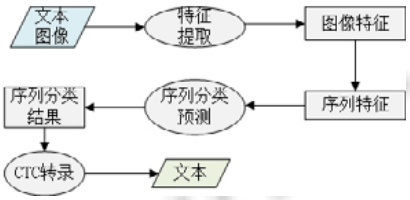
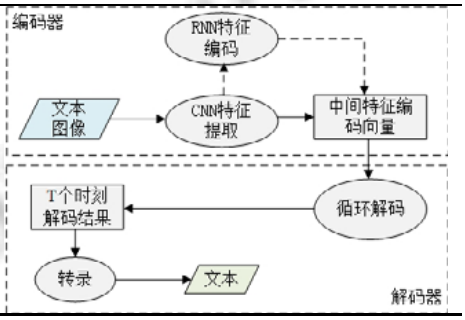
3 基于深度学习的自然场景文本识别方法

自然场景文本识别将上述文本检测步骤提取的文本图像识别为可编辑的计算机符号,本质上是将文本图像翻译为字符序列的过程.传统的文本识别方法[59,60]依赖多步骤的识别流程:首先,对平滑去燥、二值化、图像归一化等预处理;然后,利用文本内容笔画特征、形状特征、边缘特征训练的分类器,通过滑动窗口的方式识别字符内容.然而,利用滑动框提取人工设计的特征是极为耗时,又由于人工设计特征的简易性而缺乏较好的泛化性能,导致传统的方法在识别准确率上较差.

与传统的方法不同,基于深度学习的自然场景文本识别方法通过深层神经网络学习更抽象的高维特征和更复杂的映射关系.同时,深度卷积网络直接将图像像素矩阵作为模型输入,克服了手工设计特征的缺点,也减少了文本识别的预处理步骤.我们根据文本识别模型的技术实现特点将自然场景文本识别方法分为 3 类:基于朴素卷积神经网络的方法、基于时序特征分类的方法以及基于编码器和解码器的方法,概括见表 2.本节将对这 3 类算法的特点、关键技术和主要优缺点进行分析介绍.

Table 2 Natural scene text recognition methods based on deep learning

表 2 基于深度学习的自然场景文本识别方法

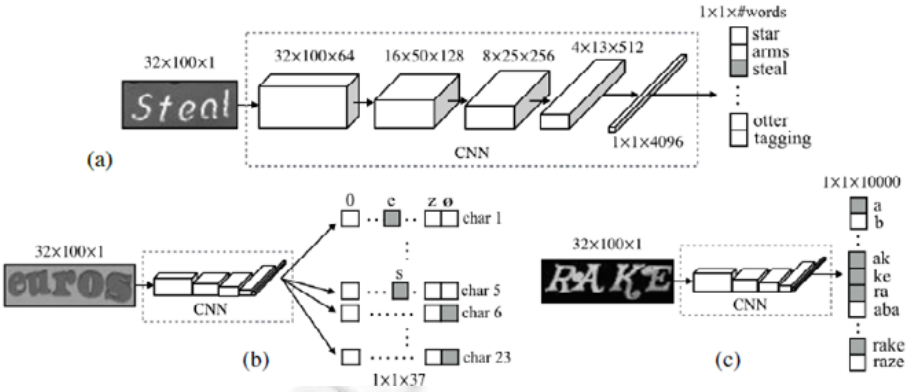
方法类别	方法特点	代表算法	方法流程
基于朴素卷积神经网络的方法	(1) 网络结构基于图像分类识别网络; (2) 需要设置文本最大长度以及编码格式,根据编码预测结果转录得到图像文本内容	DICT[61] CHAR[61] NGRAM[61] CHAR+NGRAM[62]	
基于时序特征分类的方法	(1) 利用基于滑动窗口的 CNN 提取图像的一组序列特征或者直接分割 CNN 提取的图像特征得到一组序列特征; (2) 通常序列预测结果与真值文本无法对齐,需要借助 CTC 算法对齐预测结果和真值	CRNN[63] DTRN[64]	
基于编码器和解码器的方法	(1) 方法包含两个模块:编码器模块和解码器模块; (2) 编码器通过网络得到文本图像的中间特征编码向量,解码器利用中间特征编码向量进行循环解码,直到输出结束标记.这种方式理论上可以预测任意长度文本内容	SCAN[65] AON[66] EP[67]	

3.1 基于朴素卷积神经网络的方法

基于朴素卷积神经网络的文本识别方法主要利用卷积神经网络预测不同编码格式的单词概率,该类方法的主要优点是利用卷积神经网络提取图像特征,利用训练学习得到卷积核提取图像高维特征代替传统方法手工设计的文本特征,并利用卷积参数共享降低网络参数数量级以提高识别效率.由于该类方法依赖于单词的编码格式,通常需要预先设计单词的编码格式来完成卷积神经网络的训练和学习任务.

传统的文本识别方法先需要将图像中的文本分割为字符识别再组合成单词,而文献[61]提出了基于卷积网

络的文本识别模型,直接通过卷积得到的文本图像全局特征预测文本单词.如图 8 所示,该模型设计了 3 种不同的单词编码方式识别图像文本:90k 字典编码、字符序列编码、*N*-gram 语言模型编码,其中,90k 字典编码将目标单词限定在预定义的词典中,词典中不同形式的高级英语词汇仅总数大约为 90k,这使得文本识别转化为分类问题,通过卷积神经网络预测单词分类概率识别单词;字符序列编码设定单词的最大长度为 23,模型通过预测每个位置的字符分类概率得到单词识别结果;*N*-gram 语言模型基于单词中第 n 字符出现与前 $n-1$ 个字符相关这一假设,最终根据联合概率规则得到概率最大的预测结果.相较于传统分割字符的文本识别方法,该 3 种识别方式显著提高了文本识别准确率.



(a) 字典编码 (b) 字符序列编码 (c) *N* 元语言模型编码

Fig.8 Text recognition models with different encoding types[61]

图 8 不同编码方式文本识别模型[61]

为了进一步提高文本识别率,文献[62]将文献[61]中的字符序列编码和 *N*-gram 结合到网络模型中,该模型同时预测单词的字符序列和 *N*-gram 概率分布,再根据两个预测结果利用 Beam 搜索算法得到使得字符序列与 *N*-gram 概率和最高的单词.这种根据特定编码格式预测字符的缺点在于:90k 编码方式对于字典中不存在的单词或者某些缩写文本无法识别;字符序列编码的先验条件是单词的长度固定,不适用于超过固定长度文本的识别;*N*-gram 语言模型极大地依赖预先统计得到的庞大语料库.

3.2 基于时序特征分类的方法

与基于朴素卷积神经网络的方法根据全局图像特征直接预测文本内容不同,基于时序特征分类的方法首先利用卷积网络将图像转换为图像特征序列,然后采用循环神经网络或者卷积网络将图像特征序列识别为字符概率预测序列.鉴于图像特征序列长度和字符概率预测序列长度相同,预测结果和真值可能无法对齐导致模型无法计算损失函数和训练,该类方法中往往引入连接时序分类(connectionist temporal classification,简称 CTC)算法[68].CTC 算法首先定义预测结果到真值序列之间的转换方式,利用动态规划的思想从预测概率矩阵中得到多条状态转移路径,并将最大化所有路径概率和作为优化目标.

文献[63]基于卷积网络和循环网络提出一种新颖的识别图像中序列对象的卷积循环神经网络模型(convolutional recurrent neural network,简称 CRNN),该模型结构如图 9 所示,包含卷积层、循环层和转录层:卷积层从产生的图像特征中提取特性向量序列,循环层采用多层 BiLSTM 结构学习特征序列的双向依赖关系,并预测得到文本字符序列概率;转录层根据 CTC 算法定义的预测结果转换方式将预测的字符概率序列转录为文本.由于 RNN 能获得文本序列的上下文关系特征,使得该方法的识别性能优于基于朴素卷积神经网络的方法.

不同于文献[63]先利用卷积神经网络提取文本图像特征再将特征分割为特征序列的方式,文献[64]提出一种深度文本循环网络(deep-text recurrent network,简称 DTRN),通过宽高与图像的高度相同的滑动窗口将图像剪裁为一组子图像,并将子图像依次输入卷积网络中提取特征向量,然后得到与其对应的一组特征序列.DTRN

的卷积网络采用 MaxOut 激活方法,即将特征图分为固定组数,并将每组特征图中每个位置的最大值作为激活特征.最后,该网络模型将特征序列输入到循环网络中得到序列分类预测结果.该激活方法增强了模型的拟合能力,但增加了模型的参数量.

文献[69]在文献[63]的基础上,针对图书馆库存管理应用场景提出了端到端的图书馆书籍识别模型.该模型首先利用基于 Hough 变换和场景文本显著图的分割方法得到图书的书脊,再利用 CRNN[63]时序分类文本识别方法识别书脊内容.一般的 CTC 损失函数将最大化输出完全正确的预测序列概率作为优化目标,这导致不同程度错误预测的损失值是相同的.该模型将预测结果与真值之间的编辑距离作为惩罚权重项优化 CTC 损失函数,以此提高模型的识别性能.

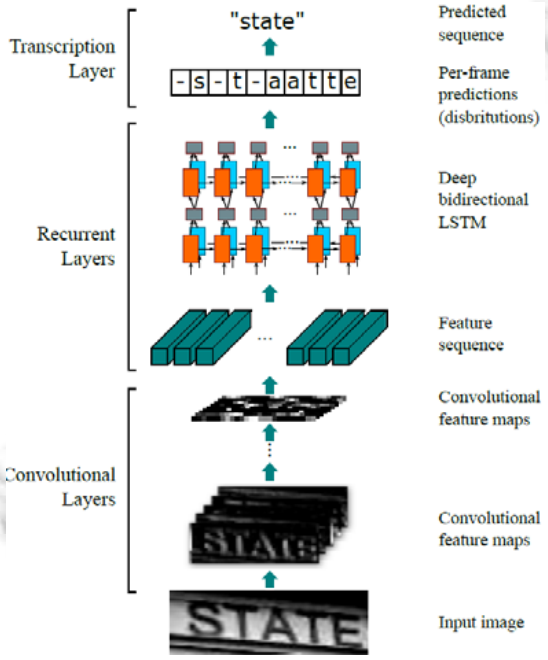


Fig.9 Structure of CRNN[63]

图 9 CRNN 模型结构[63]

文献[70]从模型集成的角度提高文本识别模型的鲁棒性,提出了一种简单但有效的自适应集成深度神经网络.该网络将不同轮次迭代训练得到的模型快照保存到模型集合中,并自适应组合来自不同迭代次数的模型.在模型训练阶段,根据快照模型预测结果是否正确、预测结果是否属于给定字典、预测结果与真值的编辑距离计算得到模型的损失并优化网络模型,在模型测试阶段利用自适应集成算法得到预测结果.相比于其他训练多个不同的网络模型再进行简单的结果投票或平均的集成模型,该文献提出的快照集成方式不需要构建多个不同的网络模型即可生成不同的预测模型.

通常,基于时序特征分类的方法大都首先利用卷积网络提取图像序列特征.文献[71]认为,这些特征的空间不变性不利于预测具有时序依赖的文本序列,因此提出了基于注意力的文本识别网络模型,并将像素的坐标添加到卷积特征中增强特征的空间相关性,其注意力机制将图像特征和循环网络隐含层输出的时变偏移特征相结合得到空间注意力矩阵,循环网络当前时刻分类预测输入特征包含前一时刻的预测输出和前一时刻得到的空间注意力权重矩阵与图像特征的加权.该方法在具有挑战性的法国街道名称标志数据集上达到了较好的识别准确率.

3.3 基于编码器和解码器的方法

基于编码器和解码器的方法是一类将序转化为另一序列的算法框架,该类方法通过编码器将图像特征转换为固定长度的中间语义编码特征,解码器将中间语义编码特征解码为文本序列.这种识别方法可以训练预测任意的两个序列之间的对应关系,而且避免了时序特征分类识别方法中的序列对齐问题.但是由于其解码器的输入特征仅依赖于的固定长度的中间语义编码特征向量,当输入序列较长时,编码器编码过程存在信息丢失的问题;其次,在解码器每个时刻的解码过程中,使用的中间语义特征是相同的,这都会给解码识别目标序列带来一定的困难.为了解决上述问题,该类方法通常引入注意力机制,使得编码器每个中间特征向量的权重不同.这样,每个时刻的输入特征与当前预测输出具有时序上下文关系,更有利于得到准确的预测结果.

受到文献[62]的启发,文献[72]提出了包含注意力模型的递归循环神经网络.相较于文献[62]中使用多分支CNN 预测人工设计的文本编码模型,文献[72]利用递归卷积网络提取文本图像特征,并使用 RNN 预测文本语言模型.该模型在不增加参数总数的条件下,使用卷积参数共享的递归卷积网络提取图像全局特征.这种方式的缺点是,仅使用单个卷积权重层来学习像素之间的长期依赖关系非常困难.该模型的解码模块包含两层 RNN 结构:第 1 层 RNN 根据前一时刻的字符解码输出学习字符序列特征,第 2 层 RNN 根据第 1 层的字符序列特征和图像注意力特征得到当前时刻的字符解码结果.这种递归循环神经网络的识别率相比文献[62]有较大的提升.

文献[73]借鉴文献[72]的注意力模型提出了基于 LSTM[6]的视觉注意力文本识别模型,如图 10 所示,该模型的编码器仅使用卷积神经网络从图像中提取一组特征作为中间特征向量.不同于文献[72]中使用 RNN 学习注意力模型,文献[73]视觉注意力模块利用多层感知机学习得到当前时刻的注意力权重矩阵,通过注意力权重使得解码器更关注与解码结果相关的图像区域内容.该模型结合 *N*-gram 模型和基于字典的预测优化损失函数提高模型的准确率.

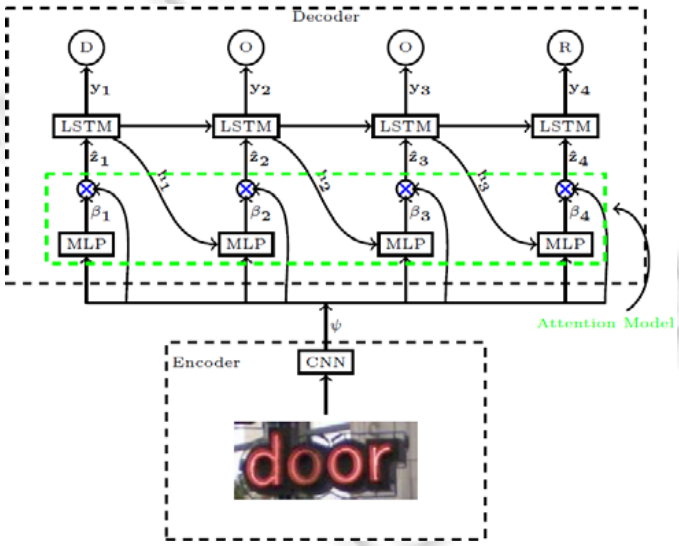


Fig.10 LSTM-based recognition framework with visual attention model[73]

图 10 基于 LSTM 的视觉注意力模型识别框架[73]

对于单层且卷积核宽度为 k 的卷积网络而言,每次卷积的输出包含该卷积位置附近宽度 k 的区域信息,多层卷积网络可以学习得到更宽的区域特征.文献[65]根据该思想提出了基于滑动卷积网络的文本识别模型,模型不依赖于循环神经网络来计算中间编码特征向量和解码结果,而是使用完全卷积结构来进行序列到序列建模.模型通过滑动卷积网络提取图像序列特征,然后使用一维卷积编码网络得到中间特征编码向量,最后利用由多层卷积网络解码器将注意力权重与中间编码特征结合的特征序列解码为文本序列.

文献[74]将多尺度网络思想应用到文本识别模型中,并提出了尺度感知特征文本识别模型.模型的编码器使用卷积神经网络从不同尺度的图像中提取不同的图像特征,尺度注意力网络根据空间注意力权重矩阵和多尺度特征得到更准确识别输入特征.通常,场景图像中字符的比例变化很大,具有不同大小感受野的多尺度编码器能够更准确地提取文本区域的上下文信息.因此,该方法对于不同比例大小的文本图像有很好的识别率.

目前,大部分的文本识别网络都假设图像中的文字平直的,然而自然场景图像中的文本通常是不规则的,例如弯曲、变形等,这使得自然场景文本识别仍然具有挑战性.文献[75]提出了能够检测不规则文本的识别模型,该模型的特点是:在识别模块的基础上引入了空间变换网络(spatial transformer network,简称 STN)模块,STN 模块通过定位网络预测得到图像中文本边缘区域的 k 个基准点,其网格生成网络再利用基准点计算空间转换参数并得到一个矫正图像的采样网格.最后,再根据采样网格将不规则输入文本图像修正为规则图像,提高识别模型的精确率.为了进一步提高 STN 网络矫正不规则文本图像的能力,文献[76]在文献[75]的基础上使用不同分辨率大小的原始图像作为定位网络以及采样网格的输入,定位网络通过低分辨率的图像预测基准点以减少网络的参数数量,而修正图像由采样器在原始图像上采样得到.由于修正模块通常会裁剪输入图像,因此这种改进能够有效地保持修正图像的质量.该文献的识别模块还增加了反向的解码预测分支,以进一步利用文本序列的前后依赖关系提高识别的准确率.

文献[66]提出了任意方向的文本特征提取网络,并将文本特征馈送到基于注意力的解码器中以生成字符序列.为了表示任意方向的字符特征,该模型使用卷积网络提取图像及旋转图像的两组特征,再将特征向量逆转得到图像的 4 组不同方向的特征,利用过滤门的加权参数,将不同方向的 4 个特征序列组合起来得到图像编码特征向量,最后,解码器根据方向加权编码特征向量预测结果序列.

文献[77]为了能够识别变形的图像文本,提出了迭代修正的文本识别模型.该模型通过图像修正模块将文本弯曲变形的图像转换为文本平直的图像,再将经过 N 次迭代,修正的图像输入到识别模块识别图像文本内容.其中,图像修正模块根据预测的文本区域 L 段线段端点的坐标以及修正后的线段在图像中的坐标得到薄板样条变化矩阵,再根据薄板样条变化矩阵修正图像内容.文献[78]认为,仿射变换无法满足多种结构变形的复杂文本图形的修正要求,因此,该文献提出了基于像素位置修正的文本识别模型.模型将图像分割为多个部分,通过图像修正模型直接预测每个图像部分的像素与修正值的偏移,然后得到整个图像的像素偏移矩阵,根据像素偏移矩阵采样得到修正后的图像.这些算法[66,77,78]的优点在于识别过程不需要对不规则的文本进行整形预处理.

由于在真实场景文本识别任务中,图像内容复杂性和质量差的特点导致现有的包含注意力机制的文本识别方法通常表现不佳,文献[79]根据可视化实验分析结果认为,注意力模型不能准确地将解码器的输入特征与图像中对应的目标区域相关联,是导致识别准确率不高的主要原因.为了解决该问题以提高文本识别准确率,该文献提出了聚焦注意力网络来识别自然图像中的文本,通过增加聚焦网络模块来检测注意力网络的注意力关注区域是否和图像中的目标字符位置对应,然后调整注意力关注区域,使得注意力网络准确聚焦相关区域.

文献[67]在文献[79]的基础上提出了一种基于编辑概率的文本识别方法来解决注意力偏移导致预测结果字符缺失或者重复的问题.该方法在注意力模块计算当前时刻字符预测正确、当前字符重复冗余、前一时刻字符丢失这 3 种情况的概率,根据编辑距离计算这 3 种情况下将预测字符序列调整为真值序列的代价,并将最小化编辑代价作为模型优化目标训练网络.相较于文献[79]提出的方法,基于编辑概率的文本识别方法不需要额外的图像像素文本标注.

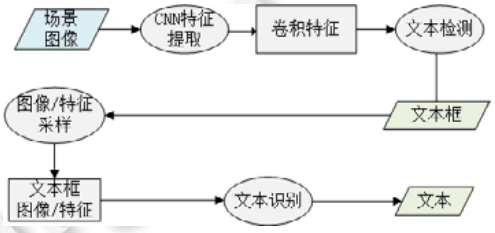
4 基于深度学习的端到端的自然场景文本识别方法

目前,大部分研究者将自然场景文本检测和识别分割为两个独立的任务,即首先利用检测网络得到图像中文本框,再将根据文本框得到剪裁的文本实例图像输入到文本识别网络识别文本内容.文献[80]尝试将文本检测和识别结合起来,利用基于滑动窗口的文本检测和字符识别模型构建一个端到端的文本识别系统,但本质上,该系统依然将文本检测和识别分割为两个单独的模型,其中,滑动窗口的方法需要进行大量的计算,且复杂的后处理在实际应用效率并不高.与此不同的是,基于深度学习的端到端的自然场景文本识别方法将文本检测任务

和文本识别任务结合在统一的网络模型中.该类方法通常共享底层卷积特征,根据共享特征检测文本区域,再将文本区域共享特征馈送到识别模块中识别文本内容.相较于将文本检测和识别分割为不同任务的方法,端到端的识别方法更具有挑战性,其优点在于,共享底层特征的方式降低了文本检测到识别过程的运算参数,并且其文本识别损失根据反向传播算法能够优化底层特征的提取和文本检测.本节将对端到端的自然场景文本识别方法(见表 3)的特点、关键技术和主要优缺点进行分析介绍.

Table 3 End-to-end natural scene text recognition method based on deep learning

表 3 基于深度学习的端到端的自然场景文本识别方法

方法类别	方法特点	代表算法	方法流程
基于深度学习的端到端的文本识别的方法	(1) 模型将文本检测模块和文本识别模块结合到统一的网络中; (2) 文本识别的结果可以反馈至文本检测网络并提高文本检测准确率	TextSpotter[81] Deep TextSpotter[82] TE-CRNN[83] FOTS[84] Mask TextSpotter[85]	

通常,自然场景文本端到端的识别方法通常都需要先进行定位再识别文本.文献[86]提出了基于深度卷积网络的街景图像多位数字识别模型,该模型通过直接在图像像素矩阵上进行卷积操作获取图像特征,并通过全连接层输出预测值.该模型根据现实场景特点将图像中的数字长度分为 0~5 以及大于 5 共 7 种情况,并将图像中的数字长度和对应长度的多位数字作为预测结果,最终选取数字长度概率和对应长度数字概率和最大的为预测结果表示图像中的数字.由于该方法对数字长度设定的先验条件,导致其只能识别出场景图像中数字长度低于 5 的多位数字,使得其应用场景很受限.

文献[82]将文本检测和识别整合到端到端的场景文本识别网络中,如图 11 所示.该网络利用统一的神经网络训练学习文本检测和识别模块,其中,文本检测模块基于通用目标检测框架 YOLOv2[87],RPN 网络的预设文本候选框采用 *k*-means 算法从训练集上得到的不同比例大小矩形框;识别模块使用基于时序特征分类文本识别方法对识别文本框内容,并且根据识别结果过滤错误的文本检测结果.为了提取不同倾斜角度的文本卷积特征,该模型用仿射变换和双线性采样代替 YOLOv2[87]的 ROI 池化方法.

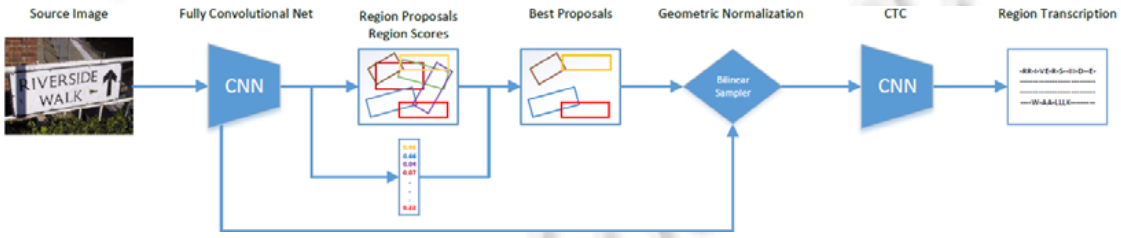


Fig.11 Structure of Deep TextSpotter[82]

图 11 Deep TextSpotter 模型结构[82]

相较于文献[82]中仅使用 RPN 模块提取候选文本框,文献[83]为了提高文本框的检测准确率,增加了对文本框的筛选和位置微调,提出了端到端的文本识别模型.该模型的文本检测网络根据文本框特征编码序列得到文本框的置信度的位置偏移,再由 LSTM 编码得到细化后的文本框特征.最后,基于编码器和解码器的文本识别模块,利用特征编码序列识别文本内容.

为了提高端到端的识别网络运算效率,文献[88]提出了全卷积端到端的文本识别网络,该网络的文本检测模块使用目标检测算法 R-FCN[27]检测图像中的文本框.R-FCN[27]的优点在于增加了目标位置敏感特征信息,并

根据位置敏感特征得到的文本框得分筛选检测结果,这使得模型能够检测到更准确的文本框从而提高其性能.受到文献[89]的启发,该模型的文本识别模块同样使用卷积网络代替文本识别模型 CRNN[63]中的循环网络结构,在不降低性能的前提下,提高模型的运算速度.

文献[81]基于文本检测模型 EAST[16]提出了端到端的文本识别模型,该模型在文本识别模块增加了注意力对齐学习,通过引入额外的聚焦损失来监督学习,得到更准确的编码字符空间信息,以此提高文本识别准确率.该聚焦损失用于优化当前时刻的解码器聚焦区域的中心坐标与该时刻解码字符在图像中的距离,其中,聚焦区域的中心坐标由每个解码时刻的注意力矩阵和文本池化层采样网格每列的中心点坐标得到.文献[84]与文献[81]类似,将文本检测模型 EAST[16]与基于时序特征分类的文本识别模型相结合,得到统一的端到端的文本识别模型.该模型使用 RoIRotate 连接文本检测模块和识别模块,RoIRotate 根据预测的角度,通过放射变化将检测的文本框特征区域变换为固定宽度和高度的特征,再使用基于时序特征分类的方法识别文本框内容.

为了识别不规则形状文本的内容,文献[85]提出一种能够检测识别任意形状文本实例的模型.该模型的文本检测模块采用目标检测模型 Faster-RCNN[25]检测文本水平矩形区域,模型的文本识别模块根据文本区域特征输出文本实例概率图、字符(包含大小写字符和数字)实例概率图和字符背景概率图,其中,文本概率图用于预测矩形区域中的文本实例区域,不同的字符实例概率图用于预测矩形区域中不同字符区域;字符背景概率图用于预测矩形区域中非文本区域.然后,该模型使用像素投票算法从左至右构造字符序列,再使用编辑距离查找给定字典中与预测字符序列最佳匹配的单词.该模型独特的文本识别方式在不同的数据集上都有较好的性能,但缺点在于只能识别英文和数字文本,对于中文这种字符数量很大的文本识别并不合适.

大多数端到端的文本识别模型的训练数据标签包含文本框的位置信息和文本框的内容信息,对于需要大量训练数据的文本识别模型来说增加了手工标注的工作量.文献[90]提出一种半监督的端到端的文本识别模型,模型的训练数据标签仅包含一组文本内容标签,不使用任何标签来训练文本检测模块.该模型的文本检测模块使用卷积网络和 BiLSTM 预测一组空间变换参数,根据变换参数和图像大小生成采样网格,再根据采样网格从原图中提取文本区域,文本检测模块使用采样得到的文本区域图像识别图像中的文本.虽然该方法的识别率不及其他方法,但为端到端的文本识别技术发展提供了新的思路.

5 自然场景文本检测与识别方法分析对比

由于单独的文本识别模型以文本检测结果作为输入,因此文本检测和文本识别模型分别间接和直接影响最终的场景图像中的文本识别结果.为了便于对自然场景文本检测与识别方法进行对比分析,本节首先介绍自然场景文本检测与识别领域的主要公共数据集,再对上述章节中介绍的一些经典自然场景文本检测与识别方法的实验结果进行对比分析.

5.1 常用公共数据集

目前常用的文本检测与识别公共数据集见表 4,其中,ICDAR2013 和 ICDAR2015 是主要的线性文本检测与识别数据集.由于非线性文本的检测与识别技术的研究需要,CTW-1500 和 Total-text 已经成为弯曲文本检测与识别的重要数据集.

(1) ICDAR2003[91]:随着自然场景中的文本检测与识别研究的快速发展,研究者迫切需要建立一些公开标准数据集.ICDAR 会议于 2003 年提出了基于 robust reading 的竞赛数据集,其该竞赛包含文本定位和文本识别任务.

(2) ICDAR2013[92]:该数据集是 2013 年 ICDAR 举行的稳健阅读竞赛(robust reading competition,简称 RRC)所提供的公共数据集.数据集的图片包含路标、书籍封面和广告牌等清晰的场景文本(focused scene text)图片.

(3) ICDAR2015[93]:该数据集是 2015 年 ICDAR 在 RRC 中增加的偶然场景文本(incidental scene text)阅读竞赛提供的公共数据集.数据集是由 Google Glass 在未聚焦的情况下随机拍摄的街头或者商场图片,旨在帮助文本检测和识别模型提高泛化性能.

Table 4 Common datasets of natural scene text detection and recognition

表 4 常用自然场景文本检测与识别数据集

数据集	文献	图片数			语言	文本形状	功能
		训练集	测试集	总计			
ICDAR2003	[91]	258	251	509	英文	线性	检测+识别
ICDAR2013	[92]	229	233	462	英文	线性	检测+识别
ICDAR2015	[93]	1 000	500	1 500	英文	线性	检测+识别
MSRA-TD500	[94]	200	300	500	中+英文	线性	检测
ICDARMLT	[95]	248	239	487	多种语言	线性	检测+识别
COCO-Text	[96]	43 686	20 000	63 686	英文	线性	检测+识别
SVT	[10]	100	250	350	英文	线性	检测+识别
RCTW-2017	[97]	8 034	4 229	12 263	中文	线性	检测+识别
CTW	[98]	25 887	6 398	32 285	中文	线性	检测+识别
CTW-1500	[99]	1 000	500	1 500	中+英文	线性+弯曲	检测+识别
Total-Text	[100]	1 255	300	1 555	英文	线性+弯曲	检测+识别

(4) MSRA-TD500[94]:MSRA-TD500 是由华中科技大学于 2012 年提供的文本检测数据集.数据集的图像包含办公室、商场和街道等场景中拍摄的图片,图片的文本由不同方向的中文和英文组成.

(5) ICDARMLT[95]:该数据集是 ICDAR 于 2017 年提供的多语言场景文本(multi-lingual scene text)图像数据集,数据集图像中的文本包含中文、英文、日文、韩文、法文和德文等主要语言类别,该数据集旨在帮助模型提高检测和识别模型多语言文本的能力.

(6) COCO-Text[96]:COCO-Text 数据集基于微软提供的目标识别数据集 MS COCO.COCO-Text 数据集图片包含背景复杂自然图片和生活场景图片,由于图片是在不关注文本的情况下收集的,因此大部分图片的中文本目标尺度小甚至不清晰,图片中也可能不包含文本内容.

(7) SVT[10]:Street View Text(简称 SVT)数据集源自 Google 街景图像.该数据集中的图像文本通常来自商业标牌,图片的文字差异较大且大部分图片具有较低的分辨率.SVT 主要应用于在街景图像中识别附近的企业的名称.

(8) RCTW-2017[97]:该数据集是由 Reading Chinese Text in the Wild 竞赛提供的针对中文检测和识别的数据集.数据集包含手机拍摄的街景、海报、菜单和室内场景图片以及屏幕截图,并且加入了合成的文本图像.

(9) CTW[85]:CTW 是由清华大学提供的中文自然场景文本图片数据集.该数据集采集于腾讯街景,具有高度多样性,图片包含了平面文本、城市街景文本、乡镇街景文本、弱照明条件下的文本、远距离文本、部分显示文本等.

(10) CTW-1500[99]:其他数据集很少包含曲线文本,为了使模型具有更优的检测性能,华南理工大学针对曲线文本检测提供了 Curve Text in the Wild 1500(简称 CTW-1500)数据集.该数据集每张图片至少包含一个曲线文本,还包含大量水平和多方向的文本.

(11) Total-Text[100]:弯曲的文字是一个很容易被忽视的问题,Total-text 是另一个针对曲线文本检测的公开数据集.数据集图片中包含商业标识、标志入口等现实生活场景中的弯曲文本.

5.2 文本检测性能评估

5.2.1 文本检测性能评估指标

目前,检测性能主要包含 3 个评价指标:召回率(*recall*)、准确率(*precision*)和调和平均(*F-measure*),其中,调和平均的计算方法如公式(1)所示.

$$F\text{-}measure=2\times Recall\times Precision/(Recall+Precision) \tag{1}$$

为了客观评估不同文本检测方法的性能,目前召回率和准确率已有不同的评估方式,下面将介绍召回率和准确率的 3 种计算方法.

(1) *Recall* 和 *Precision* 的第 1 种计算方法主要考虑真值框(ground truth)与检测框之间的 3 种匹配情况[101],

如图 12 所示,包括一对一、一对多和多对一(由于多对多的匹配情况较为少见,因此忽略).

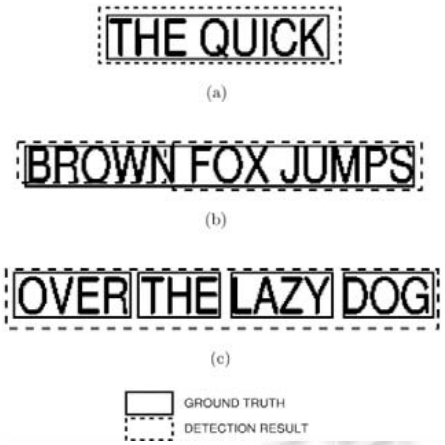


Fig.12 Different match types between ground truth rectangles and detected rectangles[101]

图 12 真值框与检测框的不同匹配类型[101]

评估方法的性能指标 *Recall* 和 *Precision* 的计算方法如公式(2)和公式(3)所示.

$$R(G,D,t_r,t_p)=\frac{\sum_i Match_G(G_i,D,t_r,t_p)}{|G|} \tag{2}$$

$$P(D,G,t_r,t_p)=\frac{\sum_j Match_D(D_j,G,t_r,t_p)}{|D|} \tag{3}$$

其中,G 表示真值框集合;D 表示检测框集合;$t_r\in[0,1]$和 $t_p\in[0,1]$分别为召回率和准确率的约束项,用于约束真值框和预测框的重叠比例(overlap),通常,t_r 和 t_p 分别取值 0.8 和 0.4;$Match_G$ 和 $Match_D$ 是用于评估不同匹配情况的函数,其定义如公式(4)和公式(5)所示.

$$Match_G(G_i,D,t_r,t_p)=\begin{cases}1, & 若G_i与一个检测框匹配\\ 0, & 若G_i不与任何检测框匹配\\ f_{sc}(k), & 若G_i与k个检测框匹配\end{cases} \tag{4}$$

$$Match_D(D_j,G,t_r,t_p)=\begin{cases}1, & 若D_j与一个真值框匹配\\ 0, & 若D_j不与任何真值框匹配\\ f_{sc}(k), & 若D_j与k个真值框匹配\end{cases} \tag{5}$$

其中,$f_{sc}(k)$表示该匹配惩罚值,通常取值为 0.8.

(2) *Recall* 和 *Precision* 的第 2 种计算方法方直接利用真值框和检测框的 Intersection over Union(简称 IoU)来评估模型的检测性能,而不考虑第 1 种方法的复杂匹配情况.*Recall* 和 *Precision* 具体计算方法如公式(6)和公式(7)所示.

$$Recall(G,D)=\frac{\sum_i IF(\max(IouMat_{i,j})>t)}{|G|} \tag{6}$$

$$Precision(D,G)=\frac{\sum_j IF(\max(IouMat_{i,j})>t)}{|D|} \tag{7}$$

其中,G 表示真值框集合;D 表示检测框集合;t 为阈值,通常取值 0.5;IF 函数为逻辑函数.公式(8)计算真值框与检测框的 *IoU* 矩阵,即 *IoUMat*.

$$IoUMat_{i,j}=\frac{area(intersection(G_i,G_j))}{area(union(G_i,G_j))} \tag{8}$$

(3) *Recall* 和 *Precision* 的第 3 种计算方法是数据集 MSAR-TD500 提供性能指标计算方法,G'和 D'分别表示将真值框和检测框旋转至水平位置的矩形框,真值框 G 和检测框 D 的重叠比例 *IoU* 定义如公式(9)所示.

$$IoU(G,D)=\frac{area(intersection(G',D'))}{area(union(G',D'))} \tag{9}$$

由于 MSAR-TD500 数据集真值框标注包含角度,因此,当真值框 G 和检测框 D 的倾斜角之差小于$\pi/8$ 且 *Overlap* 大于 0.5,则认为 D 为正确的检测框.*Recall* 和 *Precision* 具体计算方法如公式(10)和公式(11)所示.

$$Recall=\frac{|TP|}{|T|} \tag{10}$$

$$Precision=\frac{|TP|}{|E|} \tag{11}$$

其中,*TP*,*T* 和 *E* 分别表示正确的检测框集合、检测框集合和真值框集合.

5.2.2 文本检测方法分析对比

大多数文本检测方法采用 ICDAR2013、ICDAR2015 和 Total-Text 作为训练测试数据集,本节对比介绍主要文本检测方法在这些数据集上的实验结果.

(1) 文本检测方法在数据集 ICDAR2013 和 ICDAR2015 上的性能见表 5,其中,基于区域建议的方法通常借鉴 RCNN、Faster-RCNN、SSD、YOLO 等经典目标检测模型,而这类目标检测模型具有优秀的泛化性能,因此,此类文本检测方法大多数具有较好的准确率和召回率,其 *F-measure* 均超过 0.70.FCRN[29]基于 YOLO 的改进模型在 ICDAR2013 数据集上的 *F-measure* 达到了 0.83.TextBoxes[30]在 SSD 的基础上改进了候选框和卷积核的比例,使得其在 ICDAR2013 数据集上的 *F-measure* 达到了 0.85.DMPNet[31]基于 SSD 模型并使用四边形表示文本框,使得其能够在 ICDAR2015 数据集上取得 0.706 4 的 *F-measure*.RRPN[32]、R2CNN[33]、SSTD[35]、RRD[13]和 TextBoxes++[12]针对图像中文本倾斜的特点增加了倾斜文本框的检测方式,检测性能取得了不同程度的提升,其中,R2CNN[33]使用顺时针前两个坐标和高度作为文本框的表示方式,克服了角度定义的不确定性,ICDAR2013 和 ICDAR2015 数据集上的召回率分别达到了 0.935 5 和 0.856 2.SLPR[34]通过检测多条水平或者垂直等距线与文本边界的交点得到文本框,召回率达到 0.836,优于大部分文本检测方法.CRPN[37]使用基于文本区域特征的角点检测提取候选文本框,检测准确率达到 0.887,比其他直接预设大量文本框的方式在准确率上至少提高了 0.1.

基于文本组件候选的方法中,CTPN[14]在 ICDAR2013 数据集上的 *F-measure* 达到 0.88,但由于其技术特点,CTPN[14]对倾斜文本的检测性能不佳,在 ICDAR2015 数据集上的 *F-measure* 只有 0.61.SegLink[15]在网络模型中增强了对不同尺度文本的检测鲁棒性,克服了 CTPN[14]只能检测水平文本的缺点,使得其在数据集上 ICDAR2015 的 *F-measure* 相较于 CTPN[14]提高了 0.14,而 SegLink++[40]针对分组预测的改进相较于 SegLink[15]在数据集 ICDAR2015 的 *F-measure* 提高了 0.07.CENets[41]通过直接检测不同的字符是否属于同一个文本的方法将检测字符组合为文本,与 SegLink[15]相比减少了错误分组结果,其在数据集 ICDAR2013 和 ICDAR2015 上的准确率比 SegLink[15]分别提高了 0.05 和 0.13.

基于直接边框回归的方法中,DDR[46]在数据集 ICDAR2013 上达到了 0.92 的准确率,仅次于 AF-RPN[17].EAST[16]和 DDR[46]利用多尺度融合特征直接边框回归检测文本框,在数据集 ICDAR2015 上的 *F-measure* 分别达到了 0.807 2 和 0.81.DFAR[47]在直接边框回归检测的基础上增加了基于文本识别的 Refinement 筛选处理,减少了错误的检测结果,在数据集 ICDAR2015 上的准确率和 *F-measure* 与 DDR[46]相比分别提高了 0.04 和 0.02.ASTD[48]方法增加文本框长短边界的检测,克服了直接边框回归方法很难检测较长文本的缺点,其召回率相对于 DDR[46]提高了 0.061.AF-RPN[17]在数据集上 ICDAR2013 上的 *F-measure* 达到了 0.92,该模型中,3 种不同尺度的文本检测模块使得其在检测准确率和召回率上优于其他基于直接边框回归的方法,相较于 ASTD[48],AF-RPN[17]在数据集 ICDAR2013 上 *F-measure* 提高了 0.028.

Tabel 5 Performance comparison of text detection methods on ICDAR2013 and ICDAR2015

表 5 文本检测方法在 ICDAR2013 和 ICDAR2015 上的性能对比

方法分类		方法名称	ICDAR2013			ICDAR2015		
			Recall	*Precision*	*F-measure*	*Recall*	*Precision*	*F-measure*
基于区域建议的方法	基于文本区域建议的方法	FCRN[29]	0.755	0.92	0.83	–	–	–
		TextBoxes[30]	0.83	0.88	0.85	–	–	–
		DMPNet[31]	–	–	–	0.682 2	0.732 3	0.706 4
		RRPN[32]	–	–	–	0.732 3	0.821 7	0.774 4
		R2CNN[33]	0.935 5	0.83	0.877	0.856 2	0.796 8	0.825 4
		SSTD[35]	0.86	0.88	0.87	0.73	0.8	0.77
		RRD[13]	0.86	0.92	0.89	0.8	0.88	0.838
		TextBoxes++[12]	–	–	–	0.785	0.878	0.829
		SLPR[34]	–	–	–	0.836	0.855	0.845
		CRPN[37]	0.839	0.92	0.876	0.807	0.887	0.845
	基于文本组件建议的方法	CTPN[14]	0.83	0.93	0.88	0.52	0.74	0.61
		SegLink[15]	0.83	0.88	0.853	0.768	0.731	0.75
		SegLink++[40]	–	–	–	0.803	0.837	0.820
		CENet[41]	0.859 4	0.93	0.894	0.792	0.861	0.825
基于语义分割的方法	基于直接边框回归的方法	EAST[16]	–	–	–	0.783 3	0.832 7	0.807 2
		DDR[46]	0.81	0.92	0.86	0.8	0.82	0.81
		DFAR[47]	–	–	–	0.8	0.86	0.83
		ASTD[48]	0.871	0.915	0.892	–	–	–
		AF-RPN[17]	0.9	0.94	0.92	0.83	0.89	0.86
	基于分类预测的方法	CCTN[51]	0.83	0.9	0.86	–	–	–
		STD[52]	0.78	0.91	0.84	–	–	–
		PixelLink[18]	0.875	0.89	0.881	0.82	0.855	0.837
		PSENet[19]	–	–	–	0.852 2	0.893	0.872 1
	基于边界特征检测的方法	TextField[20]	–	–	–	0.805	0.843	0.824
		TextMountain[21]	–	–	–	0.841 6	0.885 1	0.862 8
		MSR[54]	0.885	0.92	0.901	–	–	–
基于区域建议和语义分割结合的方法		DSTD[56]	0.915	0.92	0.919	–	–	–
		CLRS[57]	0.920	0.844	0.880	0.895	0.797	0.843
		FTSN[22]	–	–	–	0.8	0.886	0.841
		PixelAnchor[23]	–	–	–	0.870 5	0.883 2	0.876 8
		IncepText[58]	–	–	–	0.873	0.938	0.905

基于分类预测的方法利用图像全局特征预测像素分类图,其中,CCTN[51]、STD[52]、PixelLink[18]和 PSENet[19]在数据集 ICDAR2013 和数据集 ICDAR2015 上的 *F-measure* 均超过 0.8.CCTN[51]将文本检测区域粗略裁剪放大后再检测文本中心线和文本区域的方法相较于 STD[52]提高了 0.2 的 *F-measure*,但由于边界类在分类定义上具有不确定性(即对于某像素无法直接区分其属于边界或者文本区域),其召回率相比 CCTN[51]降低了 0.05. PixelLink[18]在数据集 ICDAR 上的 *F-measure* 达到了 0.881,该方法在文本二分类预测的基础上增加了像素连通性检测,能够更准确地区分不同的文本实例.由于 ICDAR2015 数据集的文本存在模糊的情况,导致 PixelLink[18]存在错误像素连通性预测,使其检测准确率和召回率下降明显.PSENet[19]采用多级文本区域检测的方法克服了图像质量不佳情况下的不同文本实例难以区分的缺点,在数据集 ICDAR2015 上的 *F-measure* 达到了 0.872 1,优于其他基于候选区域以及基于语义分割的方法.

基于边界特征检测的方法中,TextField[20]在文本区域分类预测的基础上增加了文本区域的边界中心向量场检测,在数据集 ICDAR2015 上的 *F-measure* 达到了 0.824.TextMountain[21]使用基于边界概率图和文本中心方向预测的方法,在数据集 ICDAR2015 上的 *F-measure* 提高了 0.022 8.MSR[54]通过检测边界点与文本区域像素的坐标偏移,直接得到文本边界点集合,相较于直接边框回归方法仅检测文本框顶点偏移的方式,在数据集 ICDAR2013 上的 *F-measure* 仅次于 AF-RPN[17],达到了 0.901.

对基于区域建议和语义结合的方法而言,由于在使用区域建议的方法的基础上增加了利用文本分割图检测结果对文本框的筛选处理,检测性能优于大部分基于区域建议方法和基于语义分割的方法.DSTN[56]在数据集 ICDAR2013 上的 *F-measure* 达到了 0.919,与 AF-RPN[17]相当.CLRS[57]通过检测角点来获取文本框能够较少

错误预测的文本框,该方法在数据集 ICDAR2015 和 ICDAR2013 上的召回率分别达到了 0.920 和 0.895,优于其他方法.FTSN[22]方法增加了位置敏感概率图预测,克服了卷积网络提取的全局特征的空间位置不敏感性,在数据集 ICDAR2015 上的准确率达到了 0.886.PixelAnchor[23]将 TextBoxes++[12]和 EAST[16]结合,在数据集 ICDAR2015 上的 *F-measure* 相较于 TextBoxes++[12]和 EAST[16]分别提高了 0.047 8 和 0.069 6.IncepText[58]在 FTSN 的基础上增加了可变形卷积和池化,并引入了 Inception 结构,在数据集 ICDAR2015 上的 *F-measure* 提高了 0.028 2.

(2) 文本检测方法在曲线文本检测数据集 Total-Text 和 CTW-1500 上的性能见表 6.目前,很多文本检测模型将文本区域的线性特征作为先验条件检测水平或者倾斜的文本框,导致无法较好的检测曲线文本,其中,DMPNet[31]、CTPN[14]、SegLink[15]、CENet[41]和 PixelLink[18]在数据集 Total-text 上的 *F-measure* 不超过 0.6,在数据集 CTW-1500 上的 *F-measure* 最高只有 0.604.SegLink++[40]的改进,使其在 Total-Text 数据集上的 *F-measure* 达到了 0.815,在数据集 CTW-1500 上的 *F-measure* 相较于 SegLink[15]提高了 0.405,仅略次于 TextField[20]. TextSnake[49]采用不同大小和连接角度的圆盘表示文本区域,从而能够检测直线和曲线文本,在数据集 Total-Text 和 CTW-1500 上的 *F-measure* 分别达到了 0.784 和 0.756.PSENet[19]凭借其简单独特的多级文本区域分类预测,在数据集 CWT-1500 上取得了极佳的性能,其 *F-measure* 达到了 0.811 7.TextField[20]和 MSR[54]基于边界特征检测的方法在数据集 Total-Text 和 CTW-1500 上的准确率都超过了 0.8,其中,TextField[20]使用方向向量场检测方式在数据集上 Total-Text 和 CTW-1500 上分别取得了 0.806 和 0.814 的 *F-measure*;而 MSR[54]方法的检测准确率则优于其他方法,在数据集 Total-Text 上的准确率相对 TextField[20]提高了 0.04.

Tabel 6 Performance comparison of text detection methods on Total-Text and CTW-1500

表 6 文本检测方法在 Total-Text 和 CTW-1500 上的性能对比

方法名称	Total-Text			CTW-1500		
	Recall	*Precision*	*F-measure*	*Recall*	*Precision*	*F-measure*
DMPNet[31]	–	–	–	0.56	0.699	0.622
CTPN[14]	–	–	–	0.538	0.604	0.569
SegLink[15]	–	–	–	0.4	0.423	0.408
SegLink++[40]	0.809	0.821	0.815	0.798	0.828	0.813
CENet[41]	0.544 1	0.598 9	0.570 2	–	–	–
EAST[16]	–	–	–	0.491	0.787	0.604
TextSnake[49]	0.745	0.827	0.784	0.853	0.679	0.756
PSENet[19]	–	–	–	0.798 9	0.825	0.811 7
PixelLink[18]	0.354	0.54	0.424	–	–	–
TextField[20]	0.799	0.812	0.806	0.798	0.83	0.814
MSR[54]	0.73	0.852	0.786	0.778	0.838	0.807

5.3 文本识别性能评估

5.3.1 文本识别性能评估指标

文本识别模型的性能评估指标有两个:基于标准编辑距离度量和单词识别率.标准编辑距离定义为一个序列通过编辑(插入、删除和修改)操作转换为另一个单词所需要的最小次数,基于标准编辑距离度量为识别结果与真值之间的归一化编辑距离之和,字符串 S_i 和 S_j 的归一化编辑距离如公式(12)所示.

$$NED(S_i,S_j)=\frac{edit\_dist(S_i,S_j)}{\max(l_i,l_j)} \tag{12}$$

其中,l_i 和 l_j 分别表示字符串 S_i 和 S_j 的长度.

单词识别率是另一种定性分析文本识别模型的性能评价指标,单词识别率即为正确识别的单词总数与所有待识别单词的比例.通常,识别方法在评估模型识别性能时使用两种转录方式:无词典转录方式和基于词典的转录方式.不同的转录方式评价结果不同,其中,词典(例如,50-lexicon 包含 50 个图像中的所有单词以及从训练或测试集的其余部分选择的干扰词,full-lexicon 表示训练和测试集中所有单词组成的词汇表)用于预测结果的拼写检查约束.有词典约束的转录从词典中寻找与原始输出具有最小编辑距离的单词,无词典的转录方式直接将 t 时刻预测的最大概率标签值作为结果.

5.3.2 文本识别方法分析对比

本节对比介绍数据集 ICDAR2003、ICDAR2013 和 SVT 上各文本识别方法的性能测试结果,对比结果如表 7 所示.在数据集 ICDAR2003 上:大多数文本识别方法在 50-lexicon 和 full-lexicon 约束条件下的识别准确率高于 0.9,仅有 CHAR[61]识别模型在 50k-lexicon 约束条件下略低于 0.9;而 full-lexicon 约束条件下的识别准确率只有 DTRN[64]识别准确率低于 0.95.

Tabel 7 Performance comparison of text detection methods on ICDAR2003, ICDAR2013 and SVT

表 7 文本识别方法在 ICDAR2003、ICDAR2013 和 SVT 上的性能对比

方法分类	方法名称	ICDAR2003				ICDAR2013	SVT	
		50	50k	Full	None	None	None	50
基于朴素卷积网络的方法	DICT+IC03[61]	0.992	–	–	0.981	–	0.87	0.961
	DICT+90k[61]	0.987	0.933	–	0.986	0.908	0.807	0.954
	CHAR[61]	0.967	0.895	–	0.94	0.795	0.68	0.926
	NGRAM[61]	0.965	–	–	0.94	–	–	–
	CHAR+NGRAM[62]	0.978	0.934	0.967	0.896	0.818	0.717	0.932
基于时序分类的方法	CRNN[63]	0.978	0.955	0.976	0.894	0.867	0.808	0.964
	DTRN[64]	0.97	–	0.938	–	–	–	0.935
	RSD[69]	0.982	–	0.958	–	–	–	0.946
基于解码器和编码器的方法	VAM[73]	0.962	–	0.957	–	–	–	0.954
	SCAN[65]	0.983	–	0.972	0.921	0.904	0.85	0.957
	S-SAN[74]	0.985	–	0.977	0.929	–	0.855	0.971
	AON[66]	0.985	–	0.971	0.915	–	0.828	0.96
	RARE[75]	0.983	0.948	0.962	0.901	0.886	–	–
	ASTER[76]	0.988	–	0.980	0.945	0.918	0.936	0.992
	ESIR[77]	–	–	–	–	0.913	0.902	0.974
	FAN[79]	0.992	–	0.973	0.942	0.933	0.859	0.971
	EP[67]	0.987	–	0.979	0.946	0.944	0.875	0.966

DICT[61]识别模型使用 90k 单词合成的图像作为训练数据集,使得其在 ICDAR2003 上无约束识别率达到了 0.986,但基于词典单词编码的方式在识别中文序列或者词典中不存在单词时并不适用.CHAR[61]和 NGRAM[61]在数据集 ICDAR2003 上的准确率相当,其中,无约束识别准确率均达到了 0.94.CHAR+NGRAM[62]方法在数据集 ICDAR2013 和 SVT 上识别性能优于 CHAR[61],其中,无约束识别准确率分别提高了 0.023 和 0.037,50-lexicon 约束条件识别准确率提高了 0.006.

基于时序特征分类的方法在数据集 ICDAR2003 上 50-lexicon 约束条件识别准确率超过了 0.97,在数据集 ICDAR2003 和 SVT 上无约束条件识别准确率高于 CHAR[61],NGRAM[61]和 CHAR+NGRAM[62],仅次于 DICT[61].CRNN[63]在数据集 ICDAR2003 上 50k-lexicon 约束条件识别准确率达到了 0.955,高于基于朴素卷积神经网络的方法.DTRN[64]和 RSD[69]在数据集 ICDAR2003 上 full-lexicon 约束条件和数据集 SVT 上 50-lexicon 约束条件的准确率低于 CRNN[63],仅 RSD[69]在数据集 ICDAR2003 上 50-lexicon 的识别准确率比 CRNN[63]提高了 0.004.

基于编码器和解码器的方法利用基于注意力机制解码神经网络提取的中间编码特征,使得不同解码时刻关注与该时刻相关的上下文时序特征,该类方法识别准确率通常高于基于时序特征分类的方法.SCAN[65]、S-SAN[74]、ASTER[76]、ESIR[77]、FAN[79]和 EP[67]在数据集 ICDAR2013 上无约束条件的识别率均高于 0.9,在数据集 SVT 上无条件识别准确率高于 0.8,其中,ASTER[76]在数据集 SVT 上无约束和 50-lexicon 约束识别准确率分别达到了 0.936 和 0.992,高于其他文本识别方法.基于编辑距离优化损失函数的方法 EP[67]在数据集 ICDAR2013 上无约束条件的识别准确率达到了 0.944,在数据集 ICDAR2003 上 full-lexicon 约束条件识别准确率达到了 0.979,与 ASTER[76]相当.

5.4 端到端的文本识别性能评估

5.4.1 端到端的文本识别性能评估方式

通常,ICDAR 端到端的文本识别任务采用文献[4]中的评估方式,如果图像中某个检测框与真值框重叠超过阈值(一般为 50%)并且检测框中的单词识别正确,则该检测框文本识别成功;否则为识别失败.评估方式分为两

类:end-to-end 和 word spotting,其中,end-to-end 表示检测并识别图像中的文本,word spotting 表示检测并识别词汇表单词(即将包含不合法字符的标注的真值单词视为无关项,该项识别结果不影响评估结果).与文本识别类似,端到端的文本识别任务提供 3 种不同的约束词汇表.

(1) Strong:每张图像的强语境词汇表(100 个单词),包括图像中的所有单词以及从训练或测试集的其余部分选择的干扰词(参见文献[4]).

(2) Weakly:包括训练和测试集中所有单词的弱语境词汇表.

(3) Generic:源自 Jaderberg 等人[102]的数据集,大约 90K 单词的通用词汇表.

目前,大部分文本识别研究主要面向单独的文本检测和单独文本识别模型构建,涉及端到端的文本识别网络模型的构建研究并不多,本节介绍几种主要的端到端的文本识别方法在数据集 ICDAR2013 和 ICDAR2015 上的识别性能.

5.4.2 端到端的文本识别方法分析对比

端到端的文本识别方法在数据集 ICDAR2013 上的性能见表 8,通用词典、弱语境词典和强语境词典约束条件的文本识别 *F-measure* 逐渐提高.由于 word spotting 评估方式将图像中包含非法字符的文本视为无关文本,使得 word spotting 评估方式的 *F-measure* 高于 end-to-end.Deep TextSpotter[82]采用 RPN 和卷积特征时序分类识别图像文本内容在 end-to-end 和 word spotting 评估方式的强语境约束条件下的 *F-measure* 分别达到了 0.89 和 0.92.TE-CRNN[83]采用 Faster-RCNN 以及基于 LSTM 的编码器和解码器识别文本内容,该方法在数据集 ICDAR2013 上的不同评估方法和约束条件的识别 *F-measure* 均高于 Deep TextSpotter[82].TextSpotter[81]使用 ESAT[16]模块检测文本框并在解码器增加了注意力对齐和增强,其 *F-measure* 在数据集 ICDAR2013 上与 TE-CRNN[83]相近.Mask TextSpotter[85]方法在数据集 ICDAR2013 上 end-to-end 评估方式的 *F-measure* 达到了 0.922,0.911 和 0.865,该方法基于字符实例概率图的识别方式在数据集 ICDAR2013 上仅 word spotting 评估方式的强语境约束的 *F-measure* 低于其他方法.FOTS[84]基于 ESAT[16]和时序特征分类的识别模型在 ICDAR2013 上的 *F-measure* 与 TE-CRNN[83]相当,其中 word spotting 评估方式的强语境和弱语境约束的 *F-measure* 高于其他方法.

Tabel 8 *F-measure* comparison of end-to-end text recognition methods on ICDAR2013

表 8 端到端的文本识别方法在 ICDAR2013 上的 *F-measure* 对比

方法名称	ICDAR2013 End-to-end			ICDAR2013 Word spotting		
	Strong	Weakly	Generic	Strong	Weakly	Generic
DeepTextSpotter[82]	0.89	0.86	0.77	0.92	0.89	0.81
TE-CRNN[83]	0.910 8	0.898 1	0.845 9	0.941 6	0.924 2	0.882
TextSpotter[81]	0.91	0.89	0.86	0.93	0.92	0.87
MaskTextSpotter[85]	0.922	0.911	0.865	0.925	0.92	0.882
FOTS[84]	0.919 9	0.901 1	0.847 7	0.959 4	0.939	0.877 6

端到端的文本识别方法在数据集 ICDAR2015 上的性能见表 9.

Tabel 9 *F-measure* comparison of end-to-end text recognition methods on ICDAR2015

表 9 端到端的文本识别方法在 ICDAR2015 上的 *F-measure* 对比

方法名称	ICDAR2015 End-to-End			ICDAR2015 Word spotting		
	Strong	Weakly	Generic	Strong	Weakly	Generic
DeepTextSpotter[82]	0.54	0.51	0.47	0.58	0.53	0.51
TextSpotter[81]	0.82	0.77	0.63	0.85	0.8	0.65
MaskTextSpotter[85]	0.793	0.73	0.624	0.793	0.745	0.642
FOTS[84]	0.835 5	0.791 1	0.653 3	0.870 1	0.823 9	0.676 9

因为数据集 ICDAR2015 的图像文本质量低于 ICDAR2013,所以 Deep TextSpotter[82]、TextSpotter[81]和 Mask TextSpotter[85]的 *F-measure* 均有所下降.Mask TextSpotter[85]基于字符实例概率图的识别方法没有充分利用单词字符之间的上下文本特征,使得该方法在文本内容质量不佳的情况下 *F-measure* 下降更明显,甚至低于 TextSpotter[81].FOTS[84]在数据集 ICDAR2015 上仍然具有较好的性能,其 *F-measure* 高于其他方法.

6 发展与挑战

目前,基于卷积和循环网络的深度学习技术已经成为自然场景文本检测和识别领域的研究热点,下面将介绍该领域的技术发展挑战和展望.

(1) 文本检测技术的发展和挑战

文本检测作为目标检测研究领域的子问题,随着目标检测技术的发展,诸如 Faster-RCNN[25]、YOLO[28]、SSD[26]、FPN[43]、FCIS[44]和 Mask-RCNN[103]等网络模型为场景文本检测提供了技术思路.基于区域建议的方法由于其包含候选框提取网络导致模型大小和运算量增大,文本检测速度受到极大的影响.如何减少文本检测网络模块或者提高每个模块的效率,是未来的一个研究方向.基于文本区域建议的方法对于与预设候选框尺寸比例差别较大的文本框的检测召回率不高,如何提高该类方法对尺寸差异较大文本的检测鲁棒性是一个挑战.基于文本组件建议的方法在一定程度上避免了基于文本区域建议方法的问题,但是在图像质量不佳情况下的分组连接错误预测是亟需解决的问题.基于语义分割的方法由于其摒弃了提取候选框的步骤,在一定程度上可以提高文本检测速度,但是受到感受野大小的影响,该方法对长文本的短边定位回归误差较大.无论是基于文本区域建议的间接边框回归还是基于语义分割的直接边框回归的方法基本上都以文本框的线性形状为先验条件,导致这些方法无法检测弯曲不规则的文本,利用直接边框回归检测文本组件的方式[49]为解决该问题提供了一种解决思路.因此,如何设计合适的文本区域描述方式是提高文本检测性能的关键之一.事实上,目前基于语义分割的方法都将文本区域内的像素标注为文本类型,如果在定位的同时能够区分文本区域内的文本像素和背景像素[104],则有助于提高后续的文本识别准确率.基于分类预测的方法不直接检测文本框信息,而是利用像素分类预测区分图像中的文本和非文本区域,如何定义分类任务区分不同的文本实例,是研究者需要关注的重点.基于边界特征检测的方法利用文本区域边界与中心区域的特征关系区分不同的文本实例,如何构建简单且有效的边界特征是该类方法难点.

目前,大多数文本检测技术来源于基于深度学习的目标检测模型,往往忽略了文本与其他目标物体的特征差异性.对于检测尺寸差异性大的场景文本,可以从文本由笔画或者部分笔画组成的角度考虑,设置尺寸近似的微候选框以检测文本的任意部分(称为微文本框).这相较于基于文本组件建议的方法将进一步缩小检测的粒度,降低候选框和真值框的差异.此外,为了提高检测速度,可以借鉴基于直接边框回归的思路直接检测微文本框,再利用嵌入特征网络预测微文本框的分组.

(2) 文本识别技术的发展和挑战

深度学习技术中,基于时序特征分类的方法以及基于解码器和编码器的方法都使用循环网络学习并利用文本的时序上下文特征来识别图像文本,其中,注意力机制是如何更有效地利用时序上下文特征的一个研究重点,利用编辑概率损失[67]和聚焦注意力网络[79]解决注意力对齐问题,为研究者提供了参考.文本识别模型的输入通常来自文本检测模型的输出,这种相关性使得从检测到识别的端到端的文本识别方法是未来重要的研究方向.目前,端到端的文本识别统一模型并不多,该类方法[80−85,88]通常只是将文本检测和文本识别监督学习任务简单的堆叠在一个网络中,忽略了自然场景文本与其空间上下文可能存在的关联性,例如路牌上的文本内容为数字和地点信息以及车辆底部的文本内容可能是车牌号等.因此,可以尝试在端到端的文本识别网络模型的检测模块中,通过空间注意力机制提取文本区域的上下文特征,并将其作为识别模块输入的一部分,以提高模型的识别效果,这是值得关注的研究方向之一.

(3) 文本检测与识别的技术展望

目前,文本检测和识别技术主要采用包含卷积和循环网络的监督学习方法,但目前适用于监督学习的公开数据集中的图片数量较小且很难包含不同场景下的复杂情况,并且当训练数据与应用数据来自不同的领域或分布时经常导致模型性能下降,这无疑限制了文本识别技术的发展.目前,对抗生成网络(generative adversarial networks,简称 GAN)已经能够将图像和特征从源领域转换至目标领域,能够解决数据量不足的情况.然而领域之间的几何空间特征偏移经常会被忽略,导致数据质量不稳定.文献[105]提出了一种新颖的包含外观空间和几何空间特征转移的循环对抗生成网络,能够模拟图像外观空间和几何空间的特征变化,生成包含更加丰富的领

域特征自然场景文本图像作为数据集,从而间接提高文本检测与识别模型的性能.因此,对抗生成网络以及半监督学习甚至无监督学习等其他理论是基于深度学习的文本检测与识别领域研究需要关注的研究方向.事实上,当前的文本检测和识别仅仅达到了识别和感知程度,对于场景文本识别而言,对图像文本内容进行排版、存储和分析才是基于深度学习的文本检测与识别方法最终要要实现的目标.

7 总 结

自然场景文本检测与识别目前是计算机视觉和模式识别领域的研究热点之一,越来越多的研究者在CVPR、ECCV 和 ICDAR 等重要国际会议上提交了该领域的研究成果.本文对自然场景文本检测与识别的相关背景技术进行了阐述,分别对基于深度学的自然场景文本检测、文本识别、端到端的文本识别方法进行了分类介绍和技术特点分析对比,列举了部分方法在主要公开数据集上的测试性能,并对文本检测和识别的技术与挑战进行了总结,最后在表 10 中列出主流模型的源代码链接.人工智能深度学习技术的不断发展将为自然场景文本检测与识别应用提供更加优秀的技术解决方案.

Table 10 Source code links for mainstream detection and recognition models

表 10 主流检测与识别模型源代码链接

方法名称	链接
TextBoxes[30]	https://github.com/MhLiao/TextBoxes
TextBoxes++[12]	https://github.com/MhLiao/TextBoxes_plusplus
CRPN[37]	https://github.com/xhzdeng/crpn
RRD[13]	https://github.com/MhLiao/RRD
CTPN[14]	https://github.com/tianzhi0549/CTPN
SegLink[15]	https://github.com/dengdan/seglink
FCRN[29]	https://github.com/ankush-me/SynthText
EAST[16]	https://github.com/argman/EAST
TextSnake[49]	https://github.com/princewang1994/TextSnake.pytorch
PixelLink[18]	https://github.com/ZJULearning/pixel_link
PSENet[19]	https://github.com/whai362/PSENet
TextField[20]	https://github.com/YukangWang/TextField
CLRS[57]	https://github.com/lvpengyuan/corner
CRNN[63]	https://github.com/bgshih/crnn
AON[66]	https://github.com/huizhang0110/AON
MORAN[76,78]	https://github.com/Canjie-Luo/MORAN_v2
ASTER[76]	https://github.com/bgshih/aster
DeepTextSpotter[82]	https://github.com/VeitL/OCR
MaskTextSpotter[85]	https://github.com/lvpengyuan/masktextspotter.caffe2
SEE[90]	https://github.com/Bartzi/see

References:

[1] Li YX, Ma JW. The developments and challenges of text detection algorithms. Journal of Signal Processing, 2017,33(4):558−571. (in Chinese with English abstract). [doi: 10.16798/j.issn.1003-0530.2017.04.016]

[2] Wang RM, Sang N, Ding D, Chen J, Ye QX, Gao CX, Liu L. Text detection in natural scene image: A survey. Acta Automatica Sinica, 2018,44(12):2113−2141 (in Chinese with English abstract). http://kns.cnki.net/kcms/detail/11.2109.TP.20181010.1713.003.html [doi: 10.16383/j.aas.2018.c170572]

[3] Neumann L, Matas J. A method for text localization and recognition in real-world images. In: Proc. of the Asian Conf. on Computer Vision. 2010. 770−783. [doi: 10.1007/978-3-642-19318-7_60]

[4] Wang K, Babenko B, Belongie SJ. End-to-end scene text recognition. In: Proc. of the Int'l Conf. on Computer Vision. 2011. 1457−1464. [doi: 10.1109/ICCV.2011.6126402]

[5] Hinton GE, Salakhutdinov R. Reducing the dimensionality of data with neural networks. Science, 2006,313(5786):504−507. [doi: 10.1126/science.1127647]

[6] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997,9(8):1735−1780.

[7] Cho K, Van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078v3, 2014. [doi: 10.3115/v1/D14-1179]

[8] Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. 2010. 2963−2970. [doi: 10.1109/CVPR.2010.5540041]

[9] Matas J, Chum O, Urban M, Pajdla T. Robust wide-baseline stereo from maximally stable extremal regions. Image Vision Computing, 2004,22(10):761−767. [doi: 10.1016/j.imavis.2004.02.006]

[10] Wang K, Belongie SJ. Word spotting in the wild. In: Proc. of the European Conf. on Computer Vision. 2010. 591−604. [doi: 10.1007/978-3-642-15549-9_43]

[11] Tian S, Pan Y, Huang C, Lu S, Yu K, Tan CL. Text flow: A unified text detection system in natural scene images. In: Proc. of the Int'l Conf. on Computer Vision. 2015. 4651−4659. [doi: 10.1109/ICCV.2015.528]

[12] Liao M, Shi B, Bai X. TextBoxes++: A single-shot oriented scene text detector. IEEE Trans. on Image Processing, 2018,27(8): 3676−3690. [doi: 10.1109/TIP.2018.2825107]

[13] Liao M, Zhu Z, Shi B, Xia G, Bai X. Rotation-sensitive regression for oriented scene text detection. arXiv:1803.05265, 2018.

[14] Tian Z, Huang W, He T, He P, Qiao Y. Detecting text in natural image with connectionist text proposal network. In: Proc. of the European Conf. on Computer Vision. 2016. 56−72. [doi: 10.1007/978-3-319-46484-8_4]

[15] Shi B, Bai X, Belongie SJ. Detecting oriented text in natural images by linking segments. arXiv:1703.06520v3, 2017. [doi: 10.1109/CVPR.2017.371]

[16] Zhou X, Yao C, Wen H, Wang Y, Zhou S, He W, Liang J. EAST: An efficient and accurate scene text detector. arXiv:1704.03155v2, 2017. [doi: 10.1109/CVPR.2017.283]

[17] Zhong Z, Sun L, Huo Q. An anchor-free region proposal network for faster R-CNN based text detection approaches. Int'l Journal on Document Analysis and Recognition, 2019,22(3):315−327. [doi: 10.1007/s10032-019-00335-y]

[18] Deng D, Liu H, Cai D, Li X. PixelLink: Detecting scene text via instance segmentation. In: Proc. of the National Conf. on Artificial Intelligence. 2018. 6773−6780.

[19] Li X, Wang W, Hou W, Liu R, Lu T, Yang J. Shape robust text detection with progressive scale expansion network. arXiv:1903.12473v2, 2018.

[20] Xu Y, Wang Y, Zhou W, Wang Y, Yang Z, Bai X. TextField: Learning a deep direction field for irregular scene text detection. IEEE Trans. on Image Processing, 2018,28(11):5566−5579. [doi: 10.1109/TIP.2019.2900589]

[21] Zhu Y, Du J. TextMountain: Accurate scene text detection via instance segmentation. arXiv:1811.12786, 2018.

[22] Dai Y, Huang Z, Gao Y, Xu Y, Chen K, Guo J, Qiu W. Fused text segmentation networks for multi-oriented scene text detection. In: Proc. of the Int'l Conf. on Pattern Recognition. 2018. 3604−3609. [doi: 10.1109/ICPR.2018.8546066]

[23] Li Y, Yu Y, Li Z, Lin Y, Xu M, Li J, Zhou X. Pixel-anchor: A fast oriented scene text detector with combined networks. arXiv:1811.07432v1, 2018.

[24] Girshick RB, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv:1311.2524v3, 2013. [doi: 10.1109/CVPR.2014.81]

[25] Ren S, He K, Girshick RB, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis Machine Intelligence, 2017,39(6):1137−1149. [doi: 10.1109/TPAMI.2016.2577031]

[26] Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu C, Berg AC. SSD: Single shot MultiBox detector. In: Proc. of the European Conf. on Computer Cision. 2016. 21−37. [doi: 10.1007/978-3-319-46448-0_2]

[27] Dai J, Li Y, He K, Sun J. R-FCN: Object detection via region-based fully convolutional networks. arXiv:1605.06409v2, 2016.

[28] Redmon J, Divvala SK, Girshick RB, Farhadi A. You only look once: Unified, real-time object detection. arXiv:1506.02640v5, 2016. [doi: 10.1109/CVPR.2016.91]

[29] Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images. arXiv:1604.06646, 2016. [doi: 10.1109/CVPR.2016.254]

[30] Liao M, Shi B, Bai X, Wang X, Liu W. TextBoxes: A fast text detector with a single deep neural network. In: Proc. of the National Conf. on Artificial Intelligence. 2016. 4161−4167.

[31] Liu Y, Jin L. Deep matching prior network: Toward tighter multi-oriented text detection. arXiv:1703.01425, 2017. [doi: 10.1109/CVPR.2017.368]

[32] Ma J, Shao W, Hao Y, Li W, Hong W, Zheng Y, Xue X. Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans. on Multimedia, 2018,20(11):3111−3122. [doi: 10.1109/TMM.2018.2818020]

[33] Jiang Y, Zhu X, Wang X, Yang S, Luo Z. R2CNN: Rotational region CNN for arbitrarily-oriented scene text detection. In: Proc. of the Int'l Conf. on Pattern Recognition. 2018. [doi: 10.1109/ICPR.2018.8545598]

[34] Zhu Y, Du J. Sliding line point regression for shape robust scene text detection. In: Proc. of the Int'l Conf. on Pattern Recognition. 2018. 3735−3740. [doi: 10.1109/icpr.2018.8545067]

[35] He P, Huang W, He T, Zhu Q, Qiao Y, Li X. Single shot text detector with regional attention. In: Proc. of the Int'l Conf. on Computer Vision. 2017. 3066−3074. [doi: 10.1109/iccv.2017.331]

[36] Szegedy C, Liu W, Jia Y, Sermanet P, Reed SE, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. arXiv:1409.4842, 2014. [doi: 10.1109/CVPR.2015.7298594]

[37] Deng L, Gong Y, Lin Y, Shuai J, Tu X, Zhang Y, Ma Z, Xie M. Detecting multi-oriented text with corner-based region proposals. Neurocomputing, 2019,334:134−142. [doi: 10.1016/j.neucom.2019.01.013]

[38] He T, Huang W, Qiao Y, Yao J. Text-attentional convolutional neural network for scene text detection. IEEE Trans. on Image Processing, 2016,25(6):2529−2541. [doi: 10.1109/TIP.2016.2547588]

[39] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the Int'l Conf. on Learning Representations. 2015.

[40] Tang J, Yang Z, Wang Y, Zheng Q, Xu Y, Bai X. SegLink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. Pattern Recognition, 2019,96:Article No.106954. [doi: 10.1016/j.patcog.2019.06.020]

[41] Liu J, Zhang C, Sun Y, Han J, Ding E. Detecting text in the wild with deep character embedding network. arXiv:1901.00363, 2019.

[42] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017,39(4):640−651 [doi: 10.1109/TPAMI.2016.2572683]

[43] Lin T, Dollar P, Girshick RB, He K, Hariharan B, Belongie SJ. Feature pyramid networks for object detection. arXiv:1612.03144v2, 2017. [doi: 10.1109/CVPR.2017.106]

[44] Li Y, Qi H, Dai J, Ji X, Wei Y. Fully convolutional instance-aware semantic segmentation. arXiv:1611.07709v2, 2017. [doi: 10.1109/CVPR.2017.472]

[45] Tian X, Wang L, Ding Q. Review of image semantic segmentation based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2019,30(2):440−468 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5659.htm [doi: 10.13328/j.cnki.jos.005659]

[46] He W, Zhang X, Yin F, Liu C. Deep direct regression for multi-oriented scene text detection. In: Proc. of the Int'l Conf. on Computer Vision. 2017. 745−753. [doi: 10.1109/ICCV.2017.87]

[47] Song Y, Cui Y, Han H, Shan S, Chen X. Scene text detection via deep semantic feature fusion and attention-based refinement. In: Proc. of the Int'l Conf. on Pattern Recognition. 2018. 3747−3752. [doi: 10.1109/icpr.2018.8546050]

[48] Xue C, Lu S, Zhan F. Accurate scene text detection through border semantics awareness and bootstrapping. In: Proc. of the European Conf. on Computer Vision. 2018. 370−387.

[49] Long S, Ruan J, Zhang W, He X, Wu W, Yao C. TextSnake: A flexible representation for detecting text of arbitrary shapes. In: Proc. of the European Conf. on Computer Vision. 2018. 19−35. [doi: 10.1007/978-3-030-01216-8_2]

[50] Zhang Z, Zhang C, Shen W, Yao C, Liu W, Bai X. Multi-oriented text detection with fully convolutional networks. arXiv:1604.04018, 2016. [doi: 10.1109/CVPR.2016.451]

[51] He T, Huang W, Qiao Y, Yao J. Accurate text localization in natural image with cascaded convolutional text network. arXiv:1603.09423, 2016.

[52] Wu Y, Natarajan P. Self-organized text detection with minimal post-processing via border learning. In: Proc. of the Int'l Conf. on Computer Vision. 2017. 5010−5019. [doi: 10.1109/iccv.2017.535]

[53] Polzounov A, Ablavatski A, Escalera S, Lu S, Cai J. Wordfence: Text detection in natural images with border awareness. In: Proc. of the IEEE Int'l Conf. on Image Processing. 2017. 1222−1226. [doi: 10.1109/icip.2017.8296476]

[54] Xue C, Lu S, Zhang W. MSR: Multi-scale shape regression for scene text detection. arXiv:1901.02596v2, 2019.

[55] Bazazian D, Gomez R, Nicolaou A, Bigorda LGI, Karatzas D, Bagdanov AD. Improving text proposals for scene images with fully convolutional networks. arXiv:1702.05089, 2017.

[56] Jiang F, Hao Z, Liu X. Deep scene text detection with connected component proposals. arXiv:1708.05133, 2017.

[57] Lyu P, Yao C, Wu W, Yan S, Bai X. Multi-oriented scene text detection via corner localization and region segmentation. arXiv: 1802.08948v2, 2018. [doi: 10.1109/CVPR.2018.00788]

[58] Yang Q, Cheng M, Zhou W, Chen Y, Qiu M, Lin W. IncepText: A new inception-text module with deformable PSROI pooling for multi-oriented scene text detection. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence, 2018. 1071−1077. [doi: 10.24963/ijcai.2018/149]

[59] Bissacco A, Cummins MJ, Netzer Y, Neven H. PhotoOCR: Reading text in uncontrolled conditions. In: Proc. of the Int'l Conf. on Computer Vision. 2013. 785−792.

[60] Goel V, Mishra A, Alahari K, Jawahar CV. Whole is greater than sum of parts: Recognizing scene text words. In: Proc. of the Int'l Conf. on Document Analysis and Recognition. 2013. 398−402. [doi: 10.1109/ICDAR.2013.87]

[61] Jaderberg M, Simonyan K, Vedaldi A, Zisserman A. Synthetic data and artificial neural networks for natural scene text recognition. arXiv:1406.2227v4, 2014.

[62] Jaderberg M, Simonyan K, Vedaldi A, Zisserman A. Deep structured output learning for unconstrained text recognition. arXiv: 1412.5903v5, 2014.

[63] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. on Pattern Analysis Machine Intelligence, 2016,39(11):2298-2304. [doi: 10.1109/TPAMI.2016.2646371]

[64] He P, Huang W, Qiao Y, Loy CC, Tang X. Reading scene text in deep convolutional sequences. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2016. 3501−3508.

[65] Wu Y, Yin F, Zhang X, Liu L, Liu C. SCAN: Sliding convolutional attention network for scene text recognition. arXiv:1603.09423, 2018.

[66] Cheng Z, Xu Y, Bai F, Niu Y, Pu S, Zhou S. AON: Towards arbitrarily-oriented text recognition. arXiv:1711.04226v2. 2018. [doi: 10.1109/CVPR.2018.00584]

[67] Bai F, Cheng Z, Niu Y, Pu S, Zhou S. Edit probability for scene text recognition. arXiv:1805.03384v1, 2018. [doi: 10.1109/CVPR.2018.00163]

[68] Graves A, Fernandez S, Gomez FJ, Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proc. of the Int'l Conf. on Machine Learning. 2006. 369−376. [doi: 10.1145/1143844.1143891]

[69] Yang X, He D, Huang W, Zhou Z, Ororbia AG, Kifer D, Giles CL. Smart library: Identifying books in a library using richly supervised deep scene text reading. In: Proc. of the Joint Conf. on Digital Libraries. 2016. [doi: 10.1109/JCDL.2017.7991581]

[70] Yang C, Yin X, Li Z, Wu J, Guo C, Wang H, Xiao L. AdaDNNs: Adaptive ensemble of deep neural networks for scene text recognition. arXiv:1710.03425, 2017.

[71] Wojna Z, Gorban AN, Lee D, Murphy KP, Yu Q, Li Y, Ibarz J. Attention-based extraction of structured information from street view imagery. In: Proc. of the Int'l Conf. on Document Analysis Recognition. 2017. 844−850. [doi: 10.1109/ICDAR.2017.143]

[72] Lee C, Osindero S. Recursive recurrent nets with attention modeling for OCR in the wild. arXiv:1603.03101, 2016. [doi: 10.1109/CVPR.2016.245]

[73] Ghosh SK, Valveny E, Bagdanov AD. Visual attention models for scene text recognition. In: Proc. of the Int'l Conf. on Document Analysis and Recognition. 2017. 943−948. [doi: 10.1109/icdar.2017.158]

[74] Liu W, Chen C, Wong KK. SAFE: Scale aware feature encoder for scene text recognition. In: Proc. of the Asian Conf. on Computer Vision. 2019. 196−211. [doi: 10.1007/978-3-030-20890-5_13]

[75] Shi B, Wang X, Lyu P, Yao C, Bai X. Robust scene text recognition with automatic rectification. arXiv:1603.03915v2, 2016. [doi: 10.1109/CVPR.2016.452]

[76] Shi B, Yang M, Wang X, Lyu P, Yao C, Bai X. ASTER: An attentional scene text recognizer with flexible rectification. IEEE Trans. on Pattern Analysis Machine Intelligence, 2019,41(9):,2035−2048. [doi: 10.1109/TPAMI.2018.2848939]

[77] Zhan F, Lu S. ESIR: End-to-end scene text recognition via iterative image rectification. arXiv:1812.05824v3, 2018.

[78] Luo C, Jin L, Sun Z. MORAN: A multi-object rectified attention network for scene text recognition. Pattern Recognition, 2019, 90(12):109−118. [doi: 10.1016/j.patcog.2019.01.020]

[79] Cheng Z, Bai F, Xu Y, Zheng G, Pu S, Zhou S. Focusing attention: Towards accurate text recognition in natural images. In: Proc. of the Int'l Conf. on Computer Vision. 2017. 5086−5094. [doi: 10.1109/ICCV.2017.543]

[80] Wang T, Wu DJ, Coates A, Ng AY. End-to-end text recognition with convolutional neural networks. In: Proc. of the Int'l Conf. on Pattern Recognition. 2012. 3304−3308.

[81] He T, Tian Z, Huang W, Shen C, Qiao Y, Sun C. An end-to-end textspotter with explicit alignment and attention. arXiv:1803.03474v3, 2018. [doi: 10.1109/CVPR.2018.00527]

[82] Busta M, Neumann L, Matas J. Deep TextSpotter: An end-to-end trainable scene text localization and recognition framework. In: Proc. of the Int'l Conf. on Computer Vision. 2017. 2223−2231. [doi: 10.1109/ICCV.2017.242]

[83] Li H, Wang P, Shen C. Towards end-to-end text spotting with convolutional recurrent neural networks. In: Proc. of the Int'l Conf. on Computer Vision. 2017. 5248−5256. [doi: 10.1109/iccv.2017.560]

[84] Liu X, Liang D, Yan S, Chen D, Qiao Y, Yan J. FOTS: Fast oriented text spotting with a unified network. arXiv:1801.01671v2, 2018. [doi: 10.1109/CVPR.2018.00595]

[85] Lyu P, Liao M, Yao C, Wu W, Bai X. Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. arXiv:1807.02242v2, 2018. [doi: 10.1109/TPAMI.2019.2937086]

[86] Goodfellow IJ, Bulatov Y, Ibarz J, Arnoud SC, Shet VD. Multi-digit number recognition from street view imagery using deep convolutional neural networks. In: Proc. of the Int'l Conf. on Learning Representations. 2014.

[87] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. arXiv:1612.08242, 2016. [doi: 10.1109/CVPR.2017.690]

[88] Sui W, Zhang Q, Yang J, Chu W. A novel integrated framework for learning both text detection and recognition. In: Proc. of the Int'l Conf. on Pattern Recognition. 2018. 2233−2238. [doi: 10.1109/icpr.2018.8545047]

[89] Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN. Convolutional sequence to sequence learning. In: Proc. of the Int'l Conf. on Machine Learning. 2017. 1243−1252.

[90] Bartz C, Yang H, Meinel C. SEE: Towards semi-supervised end-to-end scene text recognition. In: Proc. of the National Conf. on Artificial Intelligence. 2018. 6674−6681.

[91] Lucas SM, Panaretos A, Sosa L, Tang A, Wong S, Young R. ICDAR 2003 robust reading competitions. In: Proc. of the Int'l Conf. on Document Analysis and Recognition. 2003. 105−122. [doi: 10.1109/ICDAR.2003.1227749]

[92] Karatzas D, Shafait F, Uchida S, Iwamura M, Bigorda LGI, Mestre SR, Mas J, Mota DF, Almazan J, Heras LDL. ICDAR 2013 robust reading competition. In: Proc. of the Int'l Conf. on Document Analysis and Recognition. 2013. 1484−1493. [doi: 10.1109/ICDAR.2013.221]

[93] Karatzas D, Gomezbigorda L, Nicolaou A, Ghosh SK, Bagdanov AD, Iwamura M, Matas J, Neumann L, Chandrasekhar VR, Lu S. ICDAR 2015 competition on robust reading. In: Proc. of the Int'l Conf. on Document Analysis and Recognition. 2015. 1156−1160. [doi: 10.1109/ICDAR.2015.7333942]

[94] Yao C, Bai X, Liu W, Ma Y, Tu Z. Detecting texts of arbitrary orientations in natural images. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2012. 1083−1090. [doi: 10.1109/CVPR.2012.6247787]

[95] Nayef N, Fei Y, Bizid I, Choi H, Ogier JM. ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification—RRC-MLT. In: Proc. of the Int'l Conf. on Document Analysis and Recognition. 2018. [doi: 10.1109/ICDAR.2017.237]

[96] Gomez R, Shi B, Gomez L, Numann L, Veit A, Matas J, Belongie SJ, Karatzas D. ICDAR2017 robust reading challenge on COCO-text. In: Proc. of the Int'l Conf. on Document Analysis and Recognition. 2017. 1435−1443. [doi: 10.1109/ICDAR.2017.234]

[97] Shi B, Cong Y, Liao M, Yang M, Xiang B. ICDAR2017 competition on reading chinese text in the wild (RCTW-17). In: Proc. of the Int'l Conf. on Document Analysis and Recognition. 2017. [doi: 10.1109/ICDAR.2017.233]

[98] Yuan T, Zhu Z, Xu K, Li C, Hu S. Chinese text in the wild. arXiv:1803.00085v1, 2018.

[99] Liu Y, Jin L, Zhang S, Zhang S. Detecting curve text in the wild: New dataset and new solution. arXiv:1803.00085, 2017.

[100] Chng CK, Chan CS. Total-text: A comprehensive dataset for scene text detection and recognition. In: Proc. of the Int'l Conf. on Document Analysis and Recognition. 2017. 935−942. [doi: 10.1109/ICDAR.2017.157]

[101] Wolf C, Jolion JM. Object count/area graphs for the evaluation of object detection and segmentation algorithms. Int'l Journal on Document Analysis, 2006,8(4):280−296. [doi: 10.1007/s10032-006-0014-0]

[102] Jaderberg M, Simonyan K, Vedaldi A, Zisserman A. Reading text in the wild with convolutional neural networks. Int'l Journal of Computer Vision, 2016,116(1):1−20. [doi: 10.1007/s11263-015-0823-z]

[103] He K, Gkioxari G, Dollar P, Girshick RB. Mask R-CNN. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020,42(2): 386−397. [doi: 10.1109/TPAMI.2018.2844175]

[104] Wang T, Jiang JH. Text recognition in any direction based on semantic segmentation. Applied Science and Technology, 2018, 45(3):59−64 (in Chinese with English abstract). http://kns.cnki.net/kcms/detail/23.1191.U.20170704.1807.006.html [doi: 10.11991/yykj.201705006]

[105] Zhan F, Xue C, Lu S. GA-DAN: Geometry-aware domain adaptation network for scene text detection and recognition. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019.

附中文参考文献:

[1] 李翌昕,马尽文.文本检测算法的发展与挑战.信号处理,2017,33(4):558−571. [doi: 10.16798/j.issn.1003-0530.2017.04.016]

[2] 王润民,桑农,丁丁,陈杰,叶齐祥,高常鑫,刘丽.自然场景图像中的文本检测综.自动化学报,2018,44(12):2113−2141. http://kns.cnki.net/kcms/detail/11.2109.TP.20181010.1713.003.html [doi: 10.16383/j.aas.2018.c170572]

[45] 田萱,王亮,丁琪.基于深度学习的图像语义分割方法综述.软件学报,2019,30(2):440−468. http://www.jos.org.cn/1000-9825/5659.htm [doi: 10.13328/j.cnki.jos.005659]

[104] 王涛,江加和.基于语义分割技术的任意方向文字识别.应用科技,2018,45(3):59−64. http://kns.cnki.net/kcms/detail/23.1191.U.20170704.1807.006.html [doi: 10.11991/yykj.201705006]

王建新(1972－),男,山东青岛人,博士,教授,博士生导师,主要研究领域为数据挖掘.

王子亚(1992－),男,硕士生,CCF 学生会员,主要研究领域为数据挖掘和机器学习.

田萱(1976－),女,博士,副教授,CCF 高级会员,主要研究领域为数据挖掘与智能信息处理.