工程科学学报
Chinese Journal of Engineering

自然场景文本检测技术研究综述

白志程 李擎 陈鹏 郭立晴

Text detection in natural scenes: a literature review

BAI Zhi-cheng, LI Qing, CHEN Peng, GUO Li-qing

引用本文:
白志程, 李擎, 陈鹏, 郭立晴. 自然场景文本检测技术研究综述[J]. 工程科学学报, 2020, 42(11): 1433-1448. doi: 10.13374/j.issn2095-9389.2020.03.24.002
BAI Zhi-cheng, LI Qing, CHEN Peng, GUO Li-qing. Text detection in natural scenes: a literature review[J]. *Chinese Journal of Engineering*, 2020, 42(11): 1433-1448. doi: 10.13374/j.issn2095-9389.2020.03.24.002

在线阅读 View online: https://doi.org/10.13374/j.issn2095-9389.2020.03.24.002

工程科学学报, 第 42 卷, 第 11 期: 1433−1448, 2020 年 11 月
Chinese Journal of Engineering, Vol. 42, No. 11: 1433−1448, November 2020
https://doi.org/10.13374/j.issn2095-9389.2020.03.24.002; http://cje.ustb.edu.cn

自然场景文本检测技术研究综述

白志程[1,2)], 李　擎[1,2)✉], 陈　鹏[3)], 郭立晴[1)]

1) 北京科技大学自动化学院, 北京 100083　2) 工业过程知识自动化教育部重点实验室, 北京 100083　3) 中国邮政储蓄银行金融科技创新部, 北京 100808

✉通信作者, E-mail: liqing@ies.ustb.edu.cn

摘　要　文本检测在自动驾驶和跨模态图像检索中具有极为广泛的应用. 该技术也是基于光学字符的文本识别任务中重要的前置环节. 目前, 复杂场景下的文本检测仍极具挑战性. 本文对自然场景文本检测进行综述, 回顾了针对该问题的主要技术和相关研究进展, 并对研究现状进行分析. 首先对问题进行概述, 分析了自然场景中文本检测的主要特点; 接着, 介绍了经典的基于连通域分析、基于滑动检测窗的自然场景文本检测技术; 在此基础上, 综述了近年来较为常用的深度学习文本检测技术; 最后, 对自然场景文本检测未来可能的研究方向进行展望.

关键词　文本检测; 场景文本; 连通域分析; 图像处理; 统计学习; 深度学习

分类号　TP18

Text detection in natural scenes: a literature review

BAI Zhi-cheng[1,2)], *LI Qing*[1,2)✉], *CHEN Peng*[3)], *GUO Li-qing*[1)]

1) School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

2) Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, Beijing 100083, China

3) FINTECH Innovation Division, Postal Savings Bank of China, Beijing 100808, China

✉ Corresponding author, E-mail: liqing@ies.ustb.edu.cn

ABSTRACT　Text detection is widely applied in the automatic driving and cross-modal image retrieval fields. This technique is also an important pre-procedure in optical character-based text recognition tasks. At present, text detection in complex natural scenes remains a challenging topic. Because text distribution and orientation are varied in different scenes and domains, there is still room for improvement in existing computer vision-based text detection methods. To complicate matters, natural scene texts, such as those in guideposts and shop signs, always contain words in different languages. Even characters are missing from some natural scene texts. These circumstances present more difficulties for feature extraction and feature description, thereby weakening the detectability of existing computer vision and image processing methods. In this context, text detection applications in natural scenes were summarized in this paper, the classical and newly presented techniques were reviewed, and the research progress and status were analyzed. First, the definitions of natural scene text detection and associated concepts were provided based on an analysis of the main characteristics of this problem. In addition, the classic natural scene text detection technologies, such as connected component analysis-based methods and sliding detection window-based methods, were introduced comprehensively. These methods were also compared and discussed. Furthermore, common deep learning models for scene text detection of the past decade were also reviewed. We divided these models into two main categories: region proposal-based models and segmentation-based models. Accordingly, the typical detection and semantic segmentation frameworks, including Faster R-CNN, SSD, Mask R-CNN, FCN, and FCIS, were integrated in the deep learning methods reviewed in this section. Moreover, hybrid algorithms that use region proposal ideas and segmentation strategies were also analyzed. As

收稿日期: 2020−03−24

基金项目: 国家自然科学基金资助项目(11296089)

a supplement, several end-to-end text recognition strategies that can automatically identify characters in natural scenes were elucidated. Finally, possible research directions and prospects in this field were analyzed and discussed.

KEY WORDS text detection; scene text; connected domain analysis; image processing; statistical learning; deep learning

文字是承载语言、记录思想、传递文明的图像或符号. 当今社会, 我们的生活场景中充满了各种各样的文本信息. 具有特定而且明确语义的文本是对自然场景极为重要的概括、说明和表达. 自然场景文本检测是实现智能场景感知的关键技术, 具有重要研究意义. 但由于自然场景中的文本存在背景复杂多样、文本字体不统一、大小不一致、方向不确定等问题, 目前对该任务的处理还未达到理想的效果. 本文首先简述了文本检测问题, 分析了自然场景文本检测的研究进展和现状. 接着, 从经典文本检测方法与深度学习文本检测方法两个方面, 分析并比较了各类自然场景文本检测技术的优缺点. 最后, 展望了自然场景文本检测未来可能的研究方向.

1 问题概述: 自然场景中的文本检测

1.1 问题定义与基本概念

文本检测(Text detection)可被视为计算机视觉目标检测(Object detection)任务的一种特殊形式. 该任务的输入为包含文本的图片, 输出为以边界框为主要形式的预测信息. 一般目标检测任务的输出为图片中动物、家具、汽车等对象的位置和区域, 而文本检测则主要关注图片中文本的精准定位. 相较于一般的目标检测, 自然场景中的文字具有多方向、不规则形状、极端长宽比和字体、颜色、背景多样等特点, 因此, 在一般目标检测上较为成功的算法往往无法直接迁移到文字检测中.

与文本检测相关联的概念是文本识别, 如光学字符识别(Optical character recognition, OCR). 该任务的输入为包含文本光学字符信号的图片或视频, 输出为对应的文字信息. 目前, OCR 技术可有效、准确地对 PDF、图片文档等形式的资料进行识别和分析, 获取文字. 然而, 对于自然场景中的路标、车牌号、建筑标识等对象, 现有 OCR 技术仍有较大的进步空间. 可大致将 OCR 分为识别特定场景的专用 OCR 和识别多种场景的通用 OCR. 比如车牌识别是对特定场景的 OCR, 而对自然场景中的文字识别则为通用场景 OCR.

与自然场景文本相关联的概念为文档图像文本、图片文档覆盖文本. 文档图像一般为二值化图像, 如文字、资料的照片和 PDF 文件, 其黑色为前景文字, 背景为白色, 便于文字的检测识别. 图片文档覆盖文本则以视频字幕、图片中经人工植入的说明性文本为主要形式. 文档图像文本、图片覆盖文本的布局相对固定, 文本区域分割相对容易. 而自然场景中的文本出现形式多变, 位置、对齐方式不统一. 自然场景图片大多为彩色, 文字区域往往产生强烈的亮度变化, 使得单从像素上区分文字和背景变得困难. 而同一文本块内, 文字的字体和字号、高度和宽度以及粗细往往保持一致, 同一文字块当中往往具有相同的颜色, 这给单词、单字等字符单元的切分带来新的困难[1]. 图 1 为自然场景中文本的示例图片.

图 1 自然场景示例图片

Fig.1 Sample images of nature scenes

一般来说, 自然场景的文本识别由于环境更加复杂多样, 其识别难度相对困难, 通常通过文本检测和文本识别两个步骤来完成. 文本检测作为 OCR 的重要技术手段之一, 也是文本识别的前提.

在文本检测任务中, 文本行(Text lines)检测是一个重要的环节. 文本行是由字符、部分字符或多字符组成的条状、不规则形状的区域. 文本检测在获取文本行后针对字符进行进一步切分.

1.2 研究进展与现状分析

文本检测与识别工作最初用于对文档图像进行分析. 由于文档图像的背景简单、文字排列整齐, 其检测识别难度较小. 经过几十年的发展, 基于文档图像的检测识别技术已经趋于成熟. 近年来, 高像素智能手机等设备的出现使越来越多的人们开始拍摄周围的事物, 积累了海量的自然场景图像. 有关自然场景图像中文本检测与识别技术的研究逐渐成为计算机视觉领域的热点问题[2].

外在、内在两方面的因素制约了对自然场景中文本的检测效果. 外在因素是指自然场景中常包含不同种类的对象如建筑、墙壁、动物、植物、行人等, 这些噪声信号会影响文本检测器的性能[3]. 在用手机拍摄图片时, 过强或偏暗的光照强度影响着对图片中文本的感知能力. 内在因素是指自然场景中文本可以是任意方向的, 所以需要检测的边界框通常为旋转的矩形或四边形; 场景文本边界框的长宽比变化很大, 且通常会存在极端的长宽比; 场景文本有字符、单词或者文本行等多种形式. 这些因素使算法在定位边界框时会难以判定文本实例.

相对人脸检测等问题, 自然场景文本检测研究相对滞后, 相关研究工作始于 20 世纪 90 年代[4]. 早期的自然场景文本检测算法利用初级、直观的图像特征; 近年来, 深度学习方法兴起, 通过深度神经网络表示图像信号[5], 可以避免繁琐低效的人工特征工程[6], 同时有效提高了场景文本检测的效果.

2 经典自然场景文本检测方法

经典的文本检测方法可分为两大类: 基于连通域分析的文本检测方法和基于滑动窗口的文本检测方法. 连通域方法首先利用边缘提取等数字图像处理技术对输入图片进行预处理, 获取文本候选区域, 进而采用不同的连通域分析方法对该区域进行细化加工, 实现字符和文本的联通和定位. 根据区域生成和特征表示方法的不同, 本文将基于连通域的方法进一步划分为基于边缘的方法、基于笔划宽度变换的方法和基于最大稳定极值区域的方法并分别进行介绍. 基于滑动窗口的方法则采用人工特征对候选区域进行表示, 并利用该特征训练分类器, 对候选区域进行预测和验证. 这两类方法在实际应用中可以互为补充.

2.1 基于连通域的方法

2.1.1 基于边缘的方法

自然场景中的文本往往具有丰富的边缘和角点信息, 基于边缘的文本检测方法通过 Canny[7] 边缘检测算子提取图片边缘和角点来获取文本的候选区域, 进而使用规则或分类器对文本候选区域进行定位预测.

文献 [8] 首先应用 Sobel 边缘检测算子[9] 获得水平, 垂直, 右上和左上方向的四个边缘图, 然后从四个边缘图中提取特征以表示文本的纹理属性, 进而应用 K 均值 (K-means) 聚类算法检测初始文本候选, 最后通过经验规则分析来识别文本区域, 并通过项目概况分析来完善文本区域. 文献 [10] 使用傅立叶–拉普拉斯滤波器过滤输入图像, 同样采用 K 均值聚类方法基于最大差异来识别候选文本区域, 随后采用文本字符串的直线度和边缘密度判断文本候选区域, 去除背景区域. 文献 [11] 通过候选边缘重组和边缘分类两个步骤优化笔划宽度变换方法. 边缘重组步骤利用分割、区域合并等手段, 将输入图像中的边缘信号处理为一组小单元 (边缘片段), 利用宽度、颜色等指导信息合并这些小单元, 从而区分文本边缘和背景; 在边界分类的步骤中, 首先将候选边界聚合到文本行中, 然后使用基于字符和基于链的特征对文本行进行分类. 文献 [12] 基于与周围像素的有效像素强度比较, 提出一种易于实现的笔划检测器, 首先检测特定的笔划关键点, 通过由关键点属性指导的局部阈值提取文本片段, 进而通过特征分析实现分类, 从而消除非文本区域. 基于边缘的文本检测方法适用于背景简单的图片, 在背景比较复杂时, 边缘检测算子极易受到干扰, 无法获取有效边缘轮廓.

2.1.2 基于笔划宽度变换的方法

笔划宽度变换 (Stroke width transform, SWT) 是一种有效的文本区域检测算法. 不同于基于边缘的方法从像素梯度、角点等方面获取图片级的特征信息, SWT 方法更关注于字符级的笔划特征. 如第 1 节中分析, OCR 技术在有噪声的图像上效果较差. SWT 通过提取出具有一致宽度的带状目标来检测文本, 有效消除了大部分噪声, 得到更可靠的光学字符识别结果.

笔划宽度变换算法由 Epshtein 等在文献 [13] 中首次提出, 该算法从高对比度边缘上的一点开始, 在垂直于边缘的方向上逐像素进行分析, 找到另一条与之平行的边缘上的一点, 由这两点构成一个笔划横截面. 许多宽度相似的笔划横截面连接构成一个完整的笔划. 笔划宽度的确定过程如图 2 所示, 其中 p 是笔划边界上的一个像素, 沿 p 点梯度方向搜索, 就可以找到笔划另一侧对应的像素 q, w 为对应笔划的宽度. 在此基础上, 笔划组成字符, 字符组成词汇和文本区域. SWT 算法的一个好处是不需要知道文本的语言和字体类型即可实现文本定位.

笔划宽度变换算法提出后, 文献 [14]、文献 [15] 对其进行了发展和改进. 文献 [14] 通过笔划宽度变换处理获得文本候选区域, 使用文本级分类器过滤非文本区域; 用文本之间的相似性连接文本

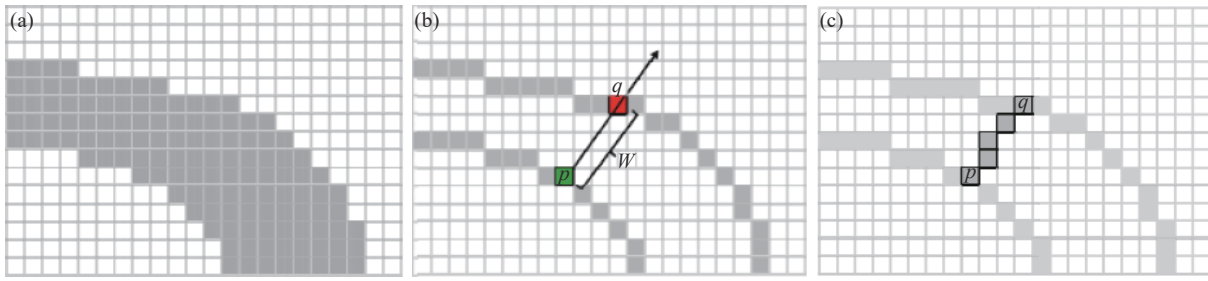

图 2　笔划宽度的定义[13]. (a)一种典型的笔划; (b)笔划边界像素; (c)笔划束上的每个像素

Fig.2　Definition of the stroke width[13]: (a) a typical stroke; (b) a pixel on the boundary of the stroke; (c) each pixel along the ray

行, 然后使用文本行级分类器进一步过滤背景区域. 尽管使用笔划宽度变换处理可以提取不同比例和方向的文本候选区域, 当图像中包含一些具有不规则梯度方向的边缘时, 受其干扰, 传统的笔划宽度变换方法往往不能准确地计算出笔划宽度, 因此文献 [15] 提出了笔划特征变换(Stroke feature transform, SFT)算子, 通过合并文本像素的颜色信息扩展笔划宽度计算, 有效分割字符中的不相关组件, 连接相关组件. 此外, 文献 [15] 依次采用文本组件分类器和文本行分类器提取文本区域, 对文本笔划的启发式属性和统计特征进行编码, 通过文本行置信度图进行阈值确定, 进而定位文本区域.

2.1.3　基于最大稳定极值区域的方法

最大稳定极值区域(Maximally stable extremal regions, MSER)是最为经典的文本检测算法之一[16]. 其主要思想源于分水岭算法, 由于文本区域往往具有相似的不连通“稳定极值”, 对于这些具有稳定极值的区域进行定位和分割即可获得字符笔划的边缘信息.

具体而言, MSER 对灰度图像进行二值化处理, 在 [0, 255] 区间内, 逐步提高阈值. 类似于分水岭算法中水平面的上升过程, 部分“山谷”和“较矮的丘陵”会被淹没, 如果从天空往下看, 则整个区域被分为陆地和水域两个部分, 即对应于切分字符和背景的二值图像. 每个阈值都会生成一个二值图. MSER 方法可以很好地描述文本内部颜色的一致性, 并且克服噪声和仿射变换的影响, 一些文献采用 MSER 方法在复杂的自然场景图像上获得出色的文本检测性能. 文献 [17] 提出将 MSER 方法应用于自然场景文本检测, 通过检测图像中的一些最大稳定极值区域来获得文本候选区域. 文献 [18] 用 MSER 算法初始化区域, 然后用自定义的距离公式合并初始区域生成一个区域集合, 最后对集合排序, 选出前几个作为文本区域. 在阈值变化过程中, MSER 的尺寸长时间保持不变. 在处理模糊、低对比度的图片时, 往往存在定位不精确、误差较大的问题. 因此, 文献 [19] 提出直接用极值区域(Extremal regions, ER)作为文本候选区域. 该方法检测图片中所有的极值区域 ER, 而不仅仅是 MSER 的子集, 并把文字检测问题处理为从 ER 集合中进行有效序列选择的问题, 达到实时检测效果. 考虑到获得的极值区域的数量过大会对后续的文本分类精度产生影响, 文献 [20] 提出了对比极值区域(Contrasting extremal region, CER)方法. CER 选取具有高对比度的极值区域, 获得的候选连通区域数量远小于 ER, 候选范围大大缩小, 提高了算法的效率. 文献 [21] 提出颜色增强的对比极值区域(Color-enhanced CER)方法, 进一步利用颜色空间中的信息滤除 CER 中的冗余像素和噪声. Color-enhanced CER 具有视觉感知一致性且对光照不敏感, 更接近人眼对颜色的判断. 文献 [22] 提出了一种基于多通道光照均衡化的 MSER 算法, 解决了传统 MSER 算法在光照不均匀图片上的文本漏检问题, 同时该文献还提出了伪字符区域过滤算法进行多特征融合, 解决了传统 MSER 算法在复杂背景图片上的漏检问题.

随着自然场景图片内容的日趋复杂, 往往出现文本目标不属于 MSER 的情况, 这限制了 MSER 方法的应用场景. 尽管 MSER 的检测准确率低于深度学习方法, 由于其具有较强的鲁棒性, 且计算成本低, 该方法常被应用于其它复杂文本检测方法的前期阶段, 产生尽可能多的候选区域.

2.2　基于滑动检测窗的方法

该类方法设计滑动检测窗, 利用窗格自上而下扫描图像, 并将每个窗格覆盖的图像区域视作文本候选区域. 通过对该区域提取特征, 分类器可得出置信度值, 通过阈值比较可实现定位和背景区域分割. 考虑到文本大小和文本行长度多变的情况, 还可以用多尺度滑动窗口进行候选区域的扫描.

文献 [23] 首先利用基础特征子对文本区域进行建模, 进而根据特征响应构建弱分类器. 这些弱

分类器被进一步集成为强分类器，在 Adaboost 算法框架下，该方法在提出时取得了具有竞争力的文本检测性能. 此外，文献 [23] 率先将纹理特征用于自然场景文本检测. 文献 [24] 进一步扩充了文献 [23] 中的特征提取方法，提取了 6 种特征并分别构建分类器，大幅度的改善了检测性能. 文献 [25] 首次提出采用多边形滑动窗口进行文本检测，该方法设计了四边形滑动窗口，在中间卷积层中对特征信号进行分析，选取重合度高的文本候选框. 在此基础上，使用基于像素点采样的 Monte-Carlo 方法快速计算多边形区域，最后使用顺序协议进行回归，实现对多边形文本的精准预测. 该文提出的多边形滑窗更加契合场景中的不规则文字（如图 3 所示），大幅度提升了召回率. 由于该方法采用了卷积特征，因此也可被看作基于深度学习的方法.

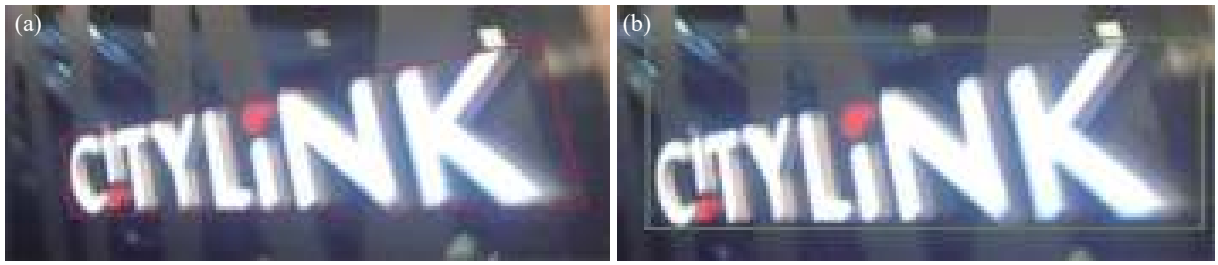

图 3 多边形滑动窗口和矩形滑动窗口检测结果比较[25]. (a) 多边形滑窗检测结果; (b) 矩形滑窗检测结果

Fig.3 Comparison of the detection results between polygon sliding windows and rectangular sliding windows[25]: (a) detection results of polygon sliding window; (b) detection result of rectangular sliding window

2.3 比较与分析

基于连通域的方法采用自底向上的策略检测文本，先检测得到单个文本，然后将相邻文本进行关联形成文本行. 这种方法利用笔化宽度的一致性和颜色的一致性等启发式规则构建文本候选区，即进行文本粗检测，然后利用分类器进一步过滤背景像素. 基于连通域的算法的一方面降低了计算的复杂度，另一方面由于检测到的连通域可以对文本直接进行分割，这有利于后续文本的识别. 然而基于连通域的算法常常面临着三个问题：第一，由于该算法对噪声的包容性差，因此非常容易形成不正确的连通域；第二，在利用启发式规则过滤连通域和文本行的噪声因素时，在不同的数据集上的检测结果具有较大差异性；第三，启发式滤除规则并不能百分百有效地区分文本区域与背景，从而造成误检.

基于滑动检测窗的方法通过“检测窗”界定文本框，因此无需像基于连通域的方法一样通过文本边缘、角点的提取来获取候选区域，可以有效避免粘连字符对候选区域提取的影响. 该类方法的主要缺陷在于对滑窗依赖极大，而窗口形状、大小、滑窗步长设置较为困难，通用性较差.

3 基于深度学习的自然场景文本检测方法

深度学习文本检测方法是一种特殊的基于学习的文本检测方法. 在经典的基于学习的文本检测方法中，多采用“人工特征子特征提取”和“分类器预测”两个步骤，受到人工特征子特征表示能力的制约. 深度神经网络具有在数据中自动学习特征表示的能力，而稠密的特征向量形式有效避免了稀疏特征向量可能造成的“维数灾难”，极大推动了机器学习技术的发展.

目前已经出现了大量的基于深度学习的自然场景文本检测方法，并取得了优于经典文本检测方法的效果[26]. 一般而言，基于深度学习的自然场景文本检测方法多采用 2 种深度学习图像处理策略：1）目标检测算法中的“区域建议”的策略；2）图像语义分割策略. 多数方法在这两种策略中有所侧重，也有很多方法既采用了基于区域建议的思想回归边界框，又用到了图像分割策略学习像素级的语义信息. 因此，本文分别介绍了基于区域建议的方法和基于分割的方法，同时在后文的“混合方法”一节中对综合采用两种策略的方法进行分析.

3.1 基于区域建议的方法

3.1.1 基本思想

该类方法以通用目标检测网络为基本模型，并在其基础上结合文本检测的实际应用对算法进行改良，如将通用的多类目标检测模型调整为单类（文本）检测模型. 以常见目标检测模型 Faster R-CNN（Faster region-based convolutional network）[27] 为例，其基本流程为：1）CNN 图片特征提取；2）候

选区域 RoI(Region of interest)与候选框生成; 3) 通过分类器生成候选框得分; 4)通过非最大值抑制方法(Non-maximum suppression, NMS)排除多余候选框, 得到最终检测结果. 被用于文本检测的常见目标检测模型还有 SSD(Single shot multi-box detector)[28]、R-FCN(Region-based fully convolutional networks)[29] 等.

3.1.2　基于 Faster R-CNN 的方法

Faster R-CNN[27] 由卷积层、区域建议网络(Region proposal network, RPN)、RoI 池化层(RoI Pooling layer)、分类回归层 4 个子模块构成. 卷积层用于提取图片特征, 其输入为整张图片, 输出为图片的特征图; RPN 用于生成与文本对象相关的多个候选框; RoI 池化层将不同尺寸的候选框转化为固定尺寸; 分类和回归层对候选区域进行预测, 同时获得候选区域在图像中的精确位置.

针对使用原生 Fast R-CNN 完成文本检测任务时有可能忽略文本行尺度的问题, 文献 [30] 提出一种基于特征融合的深度神经网络, 该网络将常用深度神经网络中的高层特征与低层特征相融合, 构建“高级语义”神经网络模型. 该文中设计了特征融合模块, 利用高层网络所抽取的高度抽象、具有强语义信息的特征信号来提高网络的整体性能. 此外, 通过多个输出层对不同尺度的文本直接进行预测. 在 ICDAR2011[31] 和 ICDAR2013[32] 数据集上的实验中, 该方法对小尺度文本的定位效果更为突出. 文献 [33] 沿用了 Faster-RCNN 中 RPN 的思想, 并在此基础上进行了改进, 提出一种旋转候选区域网络(Rotation 对齐方式 RPN). 整个网络结构和 Faster R-CNN 非常相似, 分成并行两路: 一路用于预测类别, 另一路用于回归旋转候选框. 首先由旋转候选区域网络生成带倾斜角的候选框, 同时输出候选框的类别; 接着通过 RoI 层将生成的候选框映射到特征图上. 文献 [34] 提出 R^2CNN (Rotational region CNN)算法来检测自然场景下任意角度旋转的文本. 该算法在原有 Faster R-CNN 的基础上使用 RPN 网络在文本区域坐标轴方向上产生不同方向的轴对称的候选框, 之后对每个方向的候选区域特征使用不同的池化尺寸进行特征融合. 该融合特征被用于预测文本/非文置信度, 确定轴对称候选框、倾斜候选框. 该算法取得了 F 值 82.54 的预测结果. 文献 [35] 提出一种无锚区域建议网络(Anchor-free region proposal network, AF-RPN)替代 Faster R-CNN 中的基于参考框的区域建议方法. 该方法能够摆脱复杂的参考框设计, 在水平和多方向文本检测任务中均取得了更高的召回率.

3.1.3　基于 SSD 的方法

SSD[28] 将图片输入到修改后的 VGG16[36] 得不同大小的特征映射, 然后抽取 Conv4_3、Conv7、Conv8_2、Conv9_2、Conv10_2、Conv11_2 这 6 个卷积层的特征图, 并分别在这些特征图上面的每一个点构造不同尺度大小的参考框, 最后利用 NMS 对所有的参考框进行处理, 抑制非最优参考框, 输出最终检测结果.

文献 [37] 提出 SegLink 算法, 将图片输入到 SSD 网络中, 分别提取文本框和不同框之间的连接信息. 该模型的输出不针对整个文本行或单词, 而是文本行或单词的一个部分, 被称为“片段(Segment)”. 该片段可以是 1 个或多个字符, 或 1 个单词. 通过对文本框连接信息的挖掘, 该算法以不同 Segment 的组合为最终输出, 避免了连接 Segment 构建文本行的后处理过程. 值得注意的是, SegLink 输出的参考文本框带有角度信息, 同时针对特征图上每个点仅输出一个框, 大幅度降低了计算复杂度. 文献 [38] 中的 TextBoxes 也是典型的基于 SSD 的算法. TextBoxes 修改了原始 SSD 中卷积核的大小, 同时调整了参考框的形状和长宽比, 使其更适用于文本检测. 文中还提出端到端的训练框架, 采用文本识别任务的结果进一步优化文本行检测模型, 在保证效率的情况下取得了良好的结果. 文献 [39] 提出的 TextBoxes++是 TextBoxes 的扩展版, 同样基于 SSD 网络. 该方法设计了一种文本框层(Textbox layer)结构, 解决了 SSD 无法有效检测极端长宽比文本的问题, 进一步提升检测性能. 此外, SSD 和 TextBoxes 仅支持水平方向的检测, 而 TextBoxes++可以产生有旋转角度的矩形文本检测框, 能够有效检测旋转文本. 文献 [40] 对 SSD 进行改良, 增加角度信息来检测多方向文字. 这一方法采用 Inception[41] 结构优化特征, 并在 SSD 的特征融合层增加 Attention 机制, 进一步强化文字特征. 文献 [42] 摒弃了 SSD 中分类和回归共享特征图的方式, 提出使用两个独立的网络分支分别进行分类和回归. 旋转不变特征用于分类, 方向敏感特征用于回归. 该方法可以嵌入到任何已存在的目标检测框架中, 并可以在提升精度的前提下大大减少运算时间, 对多方向文本进行检测. 文献 [43] 综合了特征金字塔网络(Feature pyramid networks, FPN)和 SegLink 模型, 提出一种高效场景文本检测模型 Seg−FPN. Seg−FPN 一方面将特

征金字塔机制与 SSD 框架相结合，对不同尺度的文本进行特征提取；另一方面通过 SegLink 链接可检测元素，实现对不同方向、长宽比的文本进行高效检测．FPN 的引入扩展了 SSD 中特征图的尺度，能够更好地定位大文本，准确识别小文本．

3.1.4　其他基于区域建议的方法

文献 [44] 以区域全卷积网络（Region based fully convolutional network, R-FCN）为基本结构，在其基础上提出了特征强化网络（Feature enhance network, FEN）．FEN 融合了高低两个维度的图像语义特征，仅采用固定尺度（3 像素×3 像素）的滑窗也可有效监测小文本，提高模型准确率、召回率．该文中还提出一种自适应权重的位置敏感 RoI 池化层，提高特征融合能力．

针对文本对象长度不统一，长短差异大的情况，文献 [45] 提出“垂直参考框”策略，仅预测文本垂直方向上的位置信息．这些参考框与 Faster-RCNN 生成的参考框类似，其主要不同在于采用了固定的 16 像素宽度，和 11 像素到 273 像素范围内的高度尺寸．这些固定宽度的小尺度文本经由循环神经网络（RNN）进一步加工、连接，得到最终文本行．文献 [46] 提出一种基于自适应区域表示的检测方法，在采用区域提取网络（Text region proposal network, Text-RPN）提取 RoI（Region of interest）时，通过基于 RNN 的修正网络（Refinement network）对 RoI 进行验证和改进．该 RNN 每次预测一对边界点，直至没有新的边界点出现为止．这一过程有效调整了文本区域的生成．

3.2　基于分割的方法

3.2.1　基本思想

该类方法以语义分割为基本技术手段，通过深度学习语义分割网络对自然场景图片进行处理，获取像素级别的标签预测．这些像素级的输出是文本行构建的基础．常被用于文本检测的分割网络有 Mask R-CNN[47]、全卷积网络（Fully convolutional network, FCN）[48]、FCIS（Fully convolutional instance-aware semantic segmentation）[49] 等．

3.2.2　基于 Mask R-CNN 的方法

Mask R-CNN[47] 扩展自 Faster R-CNN 与 Fast R-CNN，除原检测网络的的两个分支（分类、边界框回归）外，增加了用于语义分割的、具有像素级预测功能的 Mask 分支．该 Mask 分支采用平均二值交叉熵损失，与分类损失、边界框回归损失一同组成网络的损失函数．Mask R-CNN 的处理流程与 Faster R-CNN 类似，包括：1）CNN 图片特征提取；2）由 RPN 生成候选区域（ROI）和候选框；3）通过 ROI Align 层进行尺度转换；4）采用 Fast R-CNN 回归最终边界框；5）采用 Mask 分支进行像素级的语义预测或实例预测．

文献 [50] 于 ECCV（European conference on computer vision）会议提出一种基于 Mask R-CNN 的 Mask TextSpotter 网络，其主要创新点在于修改了 Mask 分支的输出结构，使其包含全局文本实例分割和字符分割功能．该版本的 Mask TextSpotter 采用字符级的分割与识别，因而可以对任意不规则形状的文本（如曲线文本）进行处理，其局限性在于需要字符级的标注来完成模型训练．针对该问题，文献 [51] 进一步改进了 Mask TextSpotter 网络（为文献 [50] 的期刊版本），在 Mask 分支中增加了空间注意力模块（Spatial attentional module, SAM）支路，有效利用空间信息和图像上下文语义，降低网络对字符级监督信息的依赖，可实现缺省字符级标注情况下的文本行识别与预测．

文献 [52] 中提出的 SPC Net（Supervised pyramid context network）也采用了实例分割方法，该模型在 Mask R-CNN 的基础上，针对曲型文本特点，添加改进的全局文本分割分支，还针对误检问题提出文本上下文模块和二次打分机制，使算法能够处理各种形状的文本．

3.2.3　基于 FCN 的方法

全卷积网络 FCN[48] 是一种端到端的语义分割方法，不同于 Mask R-CNN 等算法中带有 R-CNN 中的区域分类模块与边界框回归模块，在 FCN 中，网络输出是对整个图片的像素级预测．

文献 [53] 先利用 FCN 对图像进行处理，得到文本区域的显著图（Salient map），并对该显著图进行连通分量分析以得到文本块；在此基础上，利用 MSER 方法提取文本块中的候选字符区域，并结合候选字符的边界框生成每条文本行；该文献设计了质心 FCN 对每条文本行中字符的质心进行预测，利用质心信号过滤非文本行．文献 [54] 提出了一种灵活的文本行表征方式 Text Snake，这种“Snake”结构主要由多个有序重叠的“圆盘（disk）”串联组成，每个圆盘由文本行区域的中心 c、半径 r、方向 θ 来表征，这些表征属性借助 FCN 来预测，如图 4 所示．通过对圆盘参数的准确预测及一条分割出来的中心线，Text Snake 可以有效检测曲形文本，并得到精确的分割区域，还可以有效避免字符重叠的情况．Text Snake 是一种具有较高借鉴价值的文本行表示方法．文献 [55] 对 VGG16 网络进

行修改, 引入 2 个 1×1 的全卷积层替换原来的全连接层, 实现了从 CNN 到 FCN 的修改, 从而可以处理多尺度的输入图片. 修改后的网络可概括为文本块级 CNN 和文本行级 CNN. 面向文本块提取的 CNN 模型可有效提取图像中的文本区域. 随后文本行级 CNN 对该区域进一步加工, 提取其中的文本行.

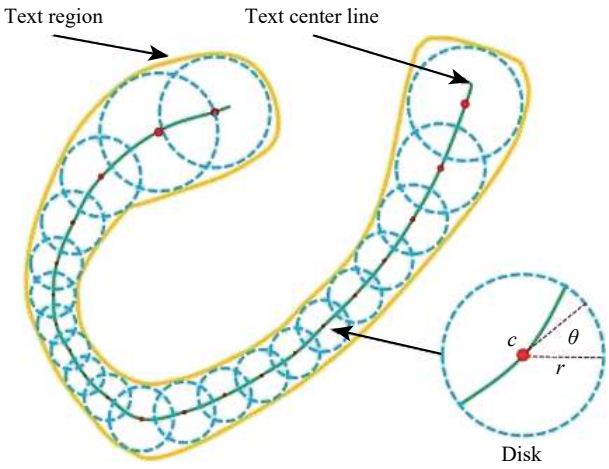

图 4　Text Snake 表征图示[54]

Fig.4　Illustration of the proposed Text Snake representation[54]

3.2.4　基于 FCIS 的方法

FCIS[49] 中采用了类似于 Fast R-CNN 的结构, 其主要区别在于 FCIS 去掉了原 R-CNN 框架中的边界框回归单元. FCIS 采用实例相关的位置敏感信息为指导, 进行特征提取与融合, 进而利用这些特征完成实例分割与分类任务.

文献 [56] 提出 PixelLink 模型, 通过深度学习网络预测与文字相关的像素与连接关系, 采用实例分割的方法, 分割出文本行区域, 然后直接找对应文本行的外接矩形框. 整个过程包括两部分: 根据“链接为正”的预测结果实现对“正像素”的预测和连通, 进而得到文本实例的分割图, 然后从分割图中直接提取文本行的边界框. 由于文字检测的定位与图像分割相比要更加精确, 而仅仅采用分割的方法不能精确的将距离近的文本很好的定位, 所以文献 [56] 采用 SegLink 中 link 的思想, 在预测中不仅预测出哪些像素是否为文本, 还要预测出这些像素能否连接进而组成一个好的文本框, 从而输出更为精确的检测区域, 其结构图如图 5 所示.

文献 [57] 将 Inception 结构集成于 FCIS 分割框架, 针对自然场景下文字的特点设计网络, 通过不同尺寸的卷积核检测不同大小和宽高比的文字; 该方法设计了柔性可变的卷积层和位置敏感的候选区域池化, 用以提升任意方向文字的检测效果. 文献 [58] 提出的 FTSN(Fused text segmentation networks)模型是 FCIS 和 FPN 的一个组合, 它是基于实例分割的端到端可训练多方向文本检测方法, 去除了中间冗余的步骤. 该文献提出了融合文本分割网络, 在特征提取过程中结合了多级特征, 并利用分割模型和基于区域建议的对象检测任务的优点同时检测和分割文本实例.

相较于一般的基于分割的方法, 实例分割方法不仅可以像素级别的分类, 而且可以通过聚类、度量学习等手段区分并定位不同的实例. 这种方法能够保持更好的底层特征(细节信息和位置信息), 但由于泛化能力较差, 因此无法应对实例类别多的复杂场景.

3.2.5　其他基于分割的方法

考虑到现有文本检测方法多基于四边形或旋转矩形, 很难对任意形状的文本进行包闭操作, 且大多数基于分割的方法不能很好地区分邻近的文本实例, 文献 [59] 提出了基于分割的单文本实例多预测的方法, 用于检测任意方向的文本. 该算法网络框架从特征金字塔网络中受到启发, 采用了 U 形的网络框架, 先将网络提取出的特征进行融合, 然后再利用分割的方式将提取出的特征进行像素分类, 最后利用像素的分类结果通过一些后处理得到文本检测结果. 该方法既能避免现有边界框回归方法所产生的对弯曲文字检测不准确的缺点, 也能改善现有基于分割的方法所产生的对“文字紧靠”现象不易分割的问题. 文献 [60] 提出基于像素聚合网络(Pixel aggregation network, PAN)的文本检测方法. 该方法的分割模块包含特征金

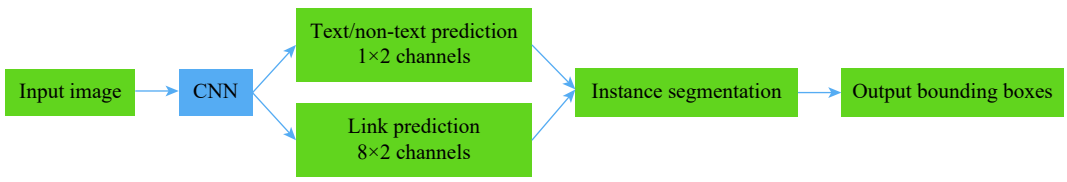

图 5　PixelLink 结构图[56]

Fig.5　Architecture of PixelLink[56]

字塔增强子模块和特征融合子模块两个部分. 其分割网络可预测文字区域、内核(Kernel, 区分文本实例的一组权重)以及相似向量, 利用该 Kernel 可重建完整的文字实例. 为了保证网络的高效率, PAN 选用了更轻量级的主干网络 ResNet18[61] 和更简单的后处理(Pixel aggregation)来降低上述两步的耗时, 从而在不损失精度的情况下, 极大加快了网络的速度. 文献 [62] 提出一种基于字符识别的文字检测方法(Character region awareness for text detection, CRAFT). 该方法的思路是利用基于单字符分割的方法, 先检测单个字符及字符间的连接关系, 然后根据字符间的连接关系确定最终的文本行. 文献 [63] 中设计了一种细分网络对文本对象进行互为独立的实例级分割和定位, 同时在特征空间中进行优化, 力求使得属于同一文本实例的像素彼此靠近, 不同文本实例的像素彼此疏远. 该方法引入形状感知损失, 对相邻的文本实例进行分割, 并能够对任意形状的文本实例, 特别是尺寸较大、较长的文本实例进行有效检测.

基于分割的文本检测方法的后处理过程往往非常耗时, 为此文献 [64] 提出 Differentiable Binarization module(DB module)来简化基于分割的方法中繁琐的后处理过程, 即如何将分割结果转换为文本框或文本区域. 有别于其他方法采用基于固定阈值的二值化手段生成边界框, 该方法采用了像素级的自适应二值化过程, 通过网络去预测图片每个位置上的阈值, 更为有效的区分出文本. 由于避免了繁琐的后处理过程, 该方法运行速度更快, 且在多个数据集上取得了目前最好的精度.

3.3 混合方法

基于分割的方法由于学习到了像素级的语义信息, 其准确率较高, 但由于小文本区域的特征响应信号较低, 容易被漏检, 降低了这类方法的召回率. 基于区域建议的方法能够捕捉小文本, 但往往对文本尺寸不够敏感, 易造成文本密集区域的锚点匹配困难情况. 因此综合两种方法优势的混合方法往往能够进一步提高文本检测精度.

文献 [65] 融合了检测和分割的思路, 首先通过检测过程中的回归方法获得角点, 然后对角点进行采样和重组, 获取候选框. 之后通过对旋转位置敏感分割图进行预测得到得分, 利用分割图的信息来辅助判断候选框的好坏, 进而可以通过 NMS 排除次优框, 得到最终的文本检测结果. 类似地, 文献 [66] 提出一种 Pixel-Anchor 方法, 该方法结合了基于锚点和基于像素分割的检测方法的特性. 文中设计了基于锚点的模块和基于像素分割的模块共享主干网络提取的特征, 在基于锚点的模块中只保留小候选框和长候选框; 在基于像素分割的模块中移除小的候选框, 保留中等大小的候选框; 然后再聚合两者的候选框, 并通过一个级联 NMS 来得到最终的检测结果. 文献 [67] 提出了一种基于 Faster-RCNN 的双任务检测模型 DSTD(Deep scene text detection). 第一个任务为文本像素分割预测, 即区分图片中的文本像素与非文本像素, 然后在此基础上利用组件连接生成候选框; 第二个任务为字符候选框检测, 输出一组候选字符, 结合之前生成的候选框, 通过保留有字符的候选框得到最终的检测结果.

3.4 端到端文本识别方法

从本质上来说, 文本检测和文本识别都属于分类问题. 文本检测主要是区分图像中的文本和非文本区域, 因此可以粗略地看作为一个二分类问题; 文本识别是要在文本检测结果中进一步区分字符, 因此可以看作为一个更精细的分类任务. 常见的 OCR 方法中往往都是把文本检测和文本识别拆分成两个部分独立进行研究. 近年来, 一些方法将文本检测和识别融合到同一个框架中完成, 同样也能达到很好的效果. 一方面文本检测和文本识别可以共享底层特征, 这降低了检测到识别过程的运算参数; 另一方面通过反向传播算法利用文本识别的损失能够优化底层特征的提取和文本检测. 目前已经出现许多优秀的端到端文本识别方法.

文献 [68] 提出一种端到端的文本检测、识别方法 Text Perceptron, 这种方法通过基于分割的文字检测方法得到文本的轮廓点, 进而通过形状转换模块对文本区域进行校正, 将校正后的结果输入文本识别模型, 其识别模型的误差可以回传给检测模型用于检测模型的进一步优化. 大多数经典文本检测方法和许多深度学习文本检测方法为多步骤方法, 其训练过程需要多个环节的调优. 这种多步结构一方面非常耗时, 另一方面, 每一步误差的累积往往会影响最终的结果. 因此文献 [69] 提出一种端到端的文本检测方法 EAST(Efficient and accurate scene text detector), 省略了候选区域聚合、文本切分、后处理等中间步骤, 直接对文本行进行预测. 该方法先利用 FCN 预测单词和文本行, 输出旋转的矩形的文本候选框或者四边形的文本候选框, 然后使用 NMS 算法过滤掉冗余的候选框, 得到最后的结果. 现有的端到端方法中, 检

测和识别两个子任务被串行连接, 文本检测任务和文本识别任务耦合度高, 且对识别噪声较为敏感. 文献 [70] 利用并行检测−识别的方法进行端到端的场景文本提取, 在检测模块和识别模块之间构建弱连接, 指导模型参数更新. 这种方法能够平衡检测和识别两个分支对系统性能的影响. 文献 [71] 提出的端到端识别方法中, 采用 PVAnet[72] 代替 EAST[73] 算法中的 ResNet50[70] 框架进行文本检测, 得到任意方向的文本候选区域. 该文中提出一种 Text-alignment 方法, 将文本候选区域固定为统一大小的特征图. 这些特征图经由一个递归神经网络进行分析处理, 得到最终的文本识别结果.

3.5　其他基于深度学习的方法

除前文所述方法外, 研究人员针对自然场景文本检测问题, 在深度学习领域还开展了很多有价值的研究. 如文献 [74] 提出一种新的实例转换网络(Instance transformation network, ITN), 它使用一种网内转换嵌入方法(In-network transformation embedding), 对自然场景中的文本行进行自适应表征, 同时提出了表征文本行的特定几何结构, 无需后处理步骤即可实现对多尺度、多方向、多语言文本的端到端检测. 文献 [75] 提出针对自然场景下文字检测的几何归一化网络(Geometry normalization networks, GNNets). GNNets 通过对待处理图像的特征图进行几何变换, 将几何分布差异较大的文本框归一化到一定的几何分布范围内, 提高了自然场景下文本测检的效果. 作者研究了几何分布对场景文本检测的影响, 发现基于 CNN 的检测器只能捕获有限的文本几何分布, 但充分利用所有训练的样本可以提高其泛化能力. 该文还提出了一种新颖的几何规范化模块(Geometry normalization module, GNM)用于归一化文本实例, 而被归一化的一组实例可被用于训练共享的文本检测器. 针对不规则形状、尺度的曲线文字检测问题, 文献 [76] 提出一种条件空间膨胀(Conditional spatial expansion, CSE)机制, 取代了传统的边框回归或分割策略. CSE 随机在文本区域初始化种子区域, 并依靠卷积网络提取的区域特征和已融合区域的上下文信息对临近的区域进行进一步融合. 文献 [77] 提出自适应贝塞尔曲线网络(Adaptive bezier-curve network, ABCNet)对场景文本实时检测. 该方法通过参数化的贝塞尔曲线以极低的计算开销自适应拟合文本形状. 文中设计了新颖的 Bezier Align 层, 可提取用于任意形状文本实例的卷积特征. 文献 [78] 提出用边界点表示任意形状文本的方法, 该模型包含多方向矩形包围框检测器、边界点检测器和识别网络三个部分. 多方向矩形包围框检测器在 RPN 提取的区域中通过对目标框的中心偏移量、宽度、高度和倾斜角度进行回归以产生多方向的矩形框, 同时利用该区域特征对文字边界点进行回归. 预测得到的边界信息进一步对文本区域特征进行矫正, 一方面能够描述精准的文本形状, 消除背景噪声所带来的负面影响; 另一方面由于边界点的表示可导, 该识别结果支持对检测结果的梯度反向传播优化. 文献 [79] 引入特征金字塔结构, 采用多尺度文本特征提取网络融合深层语义信息、浅层位置信息, 减弱了文本大小与多样性对检测结果的影响. 文献 [80] 提出了一种基于 YOLO 算法[81] 的 YOLO_BOX 定位模型. 该模型对聚类后的边界框进行灰度化处理, 然后通过计算像素灰度值的方差来得到文字的倾斜角度并进行角度矫正, 提升了倾斜文本区域定位的准确度. 文献 [82] 在 EAST[69] 算法的主干网络 PVANet[72] 中引入注意力机制模块, 使网络在特征提取时能够有效捕捉价值较高的信息, 这有效地改善了 EAST[69] 算法在预测长文本方向信息时视野不足的问题.

4　常用数据集、OCR 工具与开源项目

4.1　常用数据集

目前比较常用的自然场景文本检测基准数据集有 CTW[83]、ICDAR[84, 31−32]、MSRA-TD500[14]、COCO-Text[85]、RCTW[86]、Total-Text[87]、MLT[88] 等. CTW 数据集[83] 是由清华大学与腾讯共同推出的中文自然场景文本数据集, 包含 32285 张图像和 1018402 个中文字符. 该数据集具有高度多样性, 包含了平面文本、凸出文本、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、部分显示文本等不同的文本类型. 标注信息除真实字符、边界框外, 还包含是否被遮挡、有无复杂的背景、是否凸出、是手写体还是打印体等属性. 由于规模大, 多样性强, CTW 被广泛地应用于文本检测和文本识别模型的训练和验证.

ICDAR 系列数据集[84, 31−32] 由国际文档分析和识别会议推出. 自 2003 年开始, ICDAR 设立了和会议同名的竞赛, 并正式发布了 ICDAR2003 数据集[84]. 该数据集中的大部分文本是水平的, 且均为英文. ICDAR2011[31]、ICDAR2013[32] 是对 ICDAR2003 数据集[84] 的扩展, 增加了多方向场景文本和扭曲形式的场景文本图片, 以及视频文本、网页等. ICDAR2013 除英文外, 还补充了西班牙文、法文图片.

此外,RCTW-17[86]、Total-Text[87]、MLT[88]、COCO-Text[85]、ArT[89]数据集也发布于ICDAR会议. RCTW-17[86]数据集共12263张图像,其中8034张为训练集,4229张为测试集;这些图像多为手机摄像头于室外采集的自然场景,如建筑、标志牌、条幅等街道场景和商场墙壁等室内场景,还有一小部分是屏幕截图. Total-Text[87]是基于单词级别的英语曲线文本数据集,数据集涵盖各文本的图片,共1555张图像,其中1255作为训练集,300作为测试集. Total-Text[87]由马来亚大学发布于ICDAR,用于任意形状文本识别任务中的算法评价. ArT由Total-Text、SCUT-CTW1500和Baidu Curve Scene Text三个数据集组合而成[89],该数据集中共有10166张图像,其中5603张为训练集,4563张为测试集. 数据集中的文本形状多样,包括水平文本、多向文本和弯曲文本. MLT数据集[88]出自ICDAR MLT Challenge挑战赛,侧重于多语种场景文本的检测. COCO-Text数据集[85]由微软公司提供,源于大规模自然场景图像数据集MS COCO[90],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有效文本内容[91].

SCUT-CTW1500[92]由华南理工大学提出,包含1500张图像,主要为曲线文本数据集,其中1000张用于训练,500张用于测试. 该数据集的图像部分来自于互联网、部分通过手机摄像头收集,每张图像至少有一个曲线文本,还包括大量的水平和多向文本.

MSRA-TD500数据集[14]由华中科技大学于2012年发布于CVPR,该数据集包含多种类、多语种的500张图片,300张用于训练,200张用于测试. 这些图片由袖珍相机拍摄于室内、室外场景,室内场景图片以标志,门板和警示牌为主,室外场景图片则涉及复杂背景中的导板、广告牌等对象. 图像分辨率从1296像素×864像素到1920像素×1280像素不等.

表1中统计了上述数据集的具体信息.

文本检测结果主要采用交并比(Intersection-over-union, IoU)指标来评价,不同数据集有不同的评测方法,但现有检测性能评测方法都主要考虑三个性能参数:召回率(Recall),准确率(Precision),和综合指标(F-measure). 召回率是指在实际为正的样本中被预测为正样本的概率,准确率是指在被所有预测为正的样本中实际为正样本的概率,综合指标是召回率和准确率的加权调和平均,该值是评价文本检测方法性能的综合指标. 一般来

表1 文本检测常用数据集

Table 1 Common datasets for text detection

Dataset	Presenter	Type	Sample size(Training/Test)	Language	Direction
CTW	THU, Tencent	Scene	32285	Chinese	Horizontal
ICDAR2003	ICDAR	Scene	2276(1110/115)	English	Horizontal
ICDAR2011	ICDAR	Scene	484(229/255)	English	Horizontal
		Graph	522(420/102)	English	Curve
ICDAR2013	ICDAR	Scene	463(229/233)	English	Horizontal
		Graph	551(410/141)	English	Multiple
		Video	28(13/15)	English, French, Spanish	Multiple
MSRA-TD500	HUST	Scene	500(300/200)	English Chinese	Multiple
COCO-Text	Microsoft	Scene	63686	English	Multiple
RCTW-17	HUST	Scene	12263(8034/4229)	Chinese English	Horizontal
MLT2017	ICDAR	Scene	18000(7200/10800)	Multi-lingual	Horizontal
MLT2019	ICDAR	Scene	20000(10000/10000)	Multi-lingual	Horizontal
Total-Text	UM	Scene	1525(1225/300)	English	Multiple
SCUT-CTW1500	SCUT	Scene	1500(1000/500)	Multi-lingual	Multiple
ArT	UM, SCUT, Baidu	Scene	10166(5603/4563)	English Chinese	Multiple

说三个指标的值越高, 检测算法的性能越好.

4.2　OCR 工具介绍

面向生产、生活场景中文字提取与分析的需求, 腾讯、百度等信息技术企业已开发出若干功能强大、应用方便的 OCR 工具, 且支持以 API 的形式方便用户二次开发.

腾讯云 OCR[93] 在卡证文字、票据单据、资产证件等文档材料中优势明显, 识别过程中不需要切分单字, 直接对整行字符进行识别, 可缓解图像采集过程中的文字倾斜、模糊和畸变. 有道智云 OCR[94] 可有效识别手写体文字和扫描文件, 同样适用于身份证、购物小票和证件的识别. 特别地, 该 OCR 支持繁简中文、英语、日语、韩语、西班牙语、德语、俄语、法语、意大利语等 27 种语言文字的自动识别. 百度云 OCR[95] 常被应用于拍照文字、截图文字的识别, 可应用于搜索、书摘、笔记、翻译等移动应用中, 方便用户进行文本的提取和录入. 百度云 OCR 同样支持多语种识别, 并针对图片模糊、倾斜、翻转等情况进行了优化, 鲁棒性强, 识别速度快, 且支持超过 2 万体量的大字库, 总体识别准确率高. 创蓝万数 OCR[96] 可对图片进行自动拉伸、矫正、增强对比度处理, 并集成了智能纠错功能, 支持在干扰环境下的可靠工作.

4.3　典型开源项目

PixelLink 项目[56, 97] 采用 TensorFlow 框架实现了该算法, 项目中提供了详细的 conda 基础环境包, 方便研究人员配置环境. AdvancedEAST 开源项目[98] 基于 Keras 框架编写, 对 EAST 文本检测算法[69] 进行优化, 使算法对长文本的预测更加准确. 该项目在检测图片后会生成三个文件, 分别是检测过程的结果图、检测出最终文本边界框的圈定结果图和检测文本边界框的位置坐标文件. SegLink 开源项目[37, 99] 采用 TensorFlow 框架实现, 该项目针对 384 像素×384 像素和 512 像素×512 像素的图片提供了两个训练好的模型. CTPN 开源项目[100−101] 采用 TensorFlow 框架, 项目同样提供了训练好的模型, 支持对用户指定的测试图片进行测试; 检测结果由两部分组成, 一部分为带透明文本边界框的图片, 另一部分为包含检测文本边界框位置坐标和置信度分数的 txt 文件.

5　分析与展望

通过前文的回顾与分析, 可以看出, 自然场景文本检测方法已逐步从基于数字图像处理、连通域分析与统计学习的经典方法转为基于深度学习方法. 特别是近几年出现了大量的基于深度学习目标检测和语义分割技术的自然场景文本检测技术, 取得了优于经典方法的效果.

然而, 自然场景文本检测依然面临着一些问题待解决. 首先是形变文本的检测问题, 尽管目前已有较多研究工作针对任意方向、任意形状的自然场景文本展开, 其检测准确率仍明显低于对直线方向排列文本的检测结果. 自然场景中相邻文本之间的排列情况往往十分复杂, 多样性强, 仍需针对实际应用设计有效的方法来体现文本排列规律. 其次是混合语种的文本检测问题, 目前大多数的自然场景文本检测方法只能检测单一语种文本和极少数的混合语种文本, 对原始语料和字符一致性依赖度极高. 在开放语料环境中, 现有方法难以对混合语种文本进行检测和识别. 针对这些问题, 对自然场景文本检测的研究可从以下几个方面进一步展开:

1) 文本级特征是影响检测性能的关键因素, 许多方法在提取网络各层特征的过程中没有明确主要信息, 注意力机制可对全局信息中的显著信号进行增强, 从而减少非文本目标的误判; 针对非常规形状文本和跨语种文本设计具有针对性的注意力机制 (Attention, 如 Local attention、Self attention 等) 与感知策略有利于提高算法的区分性能.

2) 形变文本检测性能仍需进一步提高; 现有形变文本检测方法 (如基于 Mask R-CNN 的 SPC Net[48]、基于 FCN 的的 Text Snake[50] 等) 多针对曲型文本, 对于实际场景中多样化、特别不规则的文本, 现有方法识别效果较差; 可进一步分析人类对复杂形变文本的认知形式, 探索可行的方法改进策略.

3) 对现有算法和模型进一步改良、融合, 使其能够更有效地利用图片中的特征.

4) 构建大规模、多语种的文本检测数据集仍具有较大意义, 特别是针对具有形态差异性的语种如中文、英文、维文、蒙文等, 这有利于提高文本检测模型的普适性; 也可以构建先语种识别, 后文本检测的多步文本检测系统.

5) 设计端到端的文本识别模型, 将检测、识别任务集成到统一框架内进行处理, 一方面可以提高模型效率, 另一方面检测、识别的结果也可以互为作用, 提高算法性能; 如在端到端模型中对检测、识别任务作一定程度的并行处理, 减少两个子任务的依赖性; 可分别采用、改进、设计不同的检测模块 (如基于区域建议策略的 SegLink[33]、

TextBoxes[34]，基于分割策略的 pixelLink[52] 等）与识别模块（如基于 RNN 的模型、Encoder-Decoder 模型等），进行多维度的组合与替换，促进整个端到端模型的性能提升.

6）由于文本检测技术在自动驾驶领域应用极为广泛，检测实时性仍需进一步提高；一方面可改进现有文本检测结构以提高其对背景区域的识别能力，进而优化、简化后处理过程；另一方面则有赖于通用计算机领域的进步，如设计规模较小的轻量型主干网络、模型压缩技术的发展与计算硬件的改良.

6 结束语

由于自然场景中文本对象的特殊性，文本检测方法也具有不同于计算机视觉算法的一些特点. 本文首先对该问题进行了阐述和分析，接着主要从经典方法和深度学习两个方面，对自然场景文本检测技术进行了较为系统的综述与回顾. 最后经过分析，对该问题未来可能的研究方向进行展望.

致谢：感谢北京科技大学计算机与通信工程学院殷绪成教授、侯杰波博士对本文修改工作的帮助与支持.

参 考 文 献

[1] Dai J. Review of research on text detection technology in natural scenes. *Comput CD Software Appl*, 2013(18): 104
（戴津. 自然场景中文本检测技术研究综述. 计算机光盘软件与应用, 2013(18): 104）

[2] Zhuo L, Long H X, Peng Y F, et al. Image processing in encrypted domain: a comprehensive survey. *J Beijing Univ Technol*, 2016, 42(2): 174
（卓力, 龙海霞, 彭远帆, 等. 加密域图像处理综述. 北京工业大学学报, 2016, 42(2): 174）

[3] Fan Y L.*Natural Scene Text Detection Algorithm Research Based on Mobile Terminal* [Dissertation]. Xi 'an: Xidian University, 2015
（樊亚玲. 移动终端自然场景文本检测算法研究[学位论文]. 西安: 西安电子科技大学, 2015）

[4] Wang R M, Sang N, Ding D, et al. Text detection in natural scene image: a survey. *Acta Autom Sin*, 2018, 44(12): 2113
（王润民, 桑农, 丁丁, 等. 自然场景图像中的文本检测综述. 自动化学报, 2018, 44(12): 2113）

[5] Li J G, Li L J, Zhang Y, et al. A method which is suitable for the training of convolutional neural networks with multiple classifiers. *J Beijing Univ Technol*, 2018, 44(10): 1291
（李建更, 李立杰, 张岩, 等. 适用于具有多分类器的卷积神经网络训练方法. 北京工业大学学报, 2018, 44(10): 1291）

[6] Li X L, Zhang B, Wang K, et al. The development and application of artificial intelligence. *J Beijing Univ Technol*, 2020, 46(6): 583
（李晓理, 张博, 王康, 等. 人工智能的发展及应用. 北京工业大学学报, 2020, 46(6): 583）

[7] Canny J. A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell*, 1986, 8(6): 679

[8] Liu C M, Wang C H, Dai R W. Text detection in images based on unsupervised classification of edge-based features // *Eighth International Conference on Document Analysis and Recognition* (*ICDAR*'05). Seoul, 2005: 610

[9] Sobel I E. *Camera Models and Machine Perception* [Dissertation]. San Francisco: Stanford University, 1970

[10] Shivakumara P, Phan T Q, Tan C L. A laplacian approach to multi-oriented text detection in video. *IEEE Trans Pattern Anal Mach Intell*, 2011, 33(2): 412

[11] Yu C, Song Y, Meng Q, et al. Text detection and recognition in natural scene with edge analysis. *IET Computer Vision*, 2015, 9(4): 603

[12] Buta M, Neumann L, Matas J. FASText: Efficient unconstrained scene text detector // *Proceedings of the IEEE International Conference on Computer Vision*. Santiago, 2015: 1206

[13] Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform // 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. California, 2010: 2963

[14] Yao C, Bai X, Liu W Y, et al. Detecting texts of arbitrary orientations in natural images // 2012 *IEEE Conference on Computer Vision and Pattern Recognition*. Providence, 2012: 1083

[15] Huang W L, Lin Z, Yang J C, et al. Text localization in natural images using stroke feature transform and text covariance descriptors // *Proceedings of the* 2013 *IEEE International Conference on Computer Vision*. Sydney, 2013: 1241

[16] Matas J, Chum O, Urban M, et al. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vision Computing*, 2004, 22(10): 761

[17] Gomez L, Karatzas D. Object proposals for text extraction in the wild // 2015 13*th International Conference on Document Analysis and Recognition* (*ICDAR*). Tunis, 2015: 206

[18] Neumann L, Matas J. A method for text localization and recognition in real-world images // 10*th Asian Conference on Computer Vision*. Queenstown, 2010: 770

[19] Neumann L, Matas J. Real-time scene text localization and recognition // 2012 *IEEE Conference on Computer Vision and Pattern Recognition*. Providence, 2012: 3538

[20] Sun L, Huo Q. A component-tree based method for user-intention guided text extraction // *Proceedings of the* 21*st International Conference on Pattern Recognition* (*ICPR*2012). Tsukuba, 2012: 633

[21] Sun L, Huo Q, Jia W, et al. A robust approach for text detection from natural scene images. *Pattern Recognit*, 2015, 48(9): 2906

[22] Zhou P F.*Research on Text Detection and Recognition in Natural Scene Images* [Dissertation]. Xi 'an: Xi 'an University of Technology, 2019
(周鹏飞. 自然场景图像中的文本检测与识别技术研究[学位论文]. 西安: 西安理工大学, 2019)

[23] Chen X R, Yuille A L. Detecting and reading text in natural scenes // *Proceedings of the* 2004 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington, 2004: II

[24] Lee J J, Lee P H, Lee S W, et al. AdaBoost for text detection in natural scene // 2011 *International Conference on Document Analysis and Recognition*. Beijing, 2011: 429

[25] Liu Y L, Jin L W. Deep matching prior network: Toward tighter multi-oriented text detection // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, 2017: 1962

[26] Yin B C, Wang W T, W L C. Review of deep learning research. *J Beijing Univ Technol*, 2015, 41(1): 48
(尹宝才, 王文通, 王立春. 深度学习研究综述. 北京工业大学学报, 2015, 41(1): 48)

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

[28] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector // *European Conference on Computer Vision*. Amsterdam, 2016: 21

[29] Dai J F, Li Y, He K M, et al. R-FCN: object detection via region-based fully convolutional networks// *Proceedings of the* 30*th International Conference on Neural Information Processing Systems*. Barcelona, 2016: 379

[30] Yu Z, Wang Q Q, Lü Y. Scene text detection based on feature fusion network. *Comput Syst Appl*, 2018, 27(10): 1
(余峥, 王晴晴, 吕岳. 基于特征融合网络的自然场景文本检测. 计算机系统应用, 2018, 27(10): 1)

[31] Karatzas D, Mestre S R, Mas J, et al. ICDAR 2011 robust reading competition-challenge 1: reading text in born-digital images (web and email) // 2011 *International Conference on Document Analysis and Recognition*. Beijing, 2011: 1485

[32] Karatzas D, Shafait F, Uchida S, et al. ICDAR 2013 robust reading competition // 2013 12*th International Conference on Document Analysis and Recognition*. Washington, 2013: 1484

[33] Ma J Q, Shao W Y, Ye H, et al. Arbitrary-oriented scene text detection *via* rotation proposals. *IEEE Trans Multimedia*, 2018, 20(11): 3111

[34] Jiang Y Y, Zhu X Y, Wang X B, et al. R2CNN: rotational region CNN for orientation robust scene text detection [J/OL]. *arXiv preprint* (2017-06-30)[2020-03-01]. https://arxiv.org/abs/1706.09579

[35] Zhong Z Y, Sun L, Huo Q. An anchor-free region proposal network for Faster R-CNN-based text detection approaches. *Int J Doc Anal Recognit*, 2019, 22(3): 315

[36] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. *arXiv preprint* (2015-04-10)[2020-03-01]. https://arxiv.org/abs/1409.1556

[37] Shi B G, Bai X, Belongie S. Detecting oriented text in natural images by linking segments // *Proceedings of the* 2017 *IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, 2017: 2550

[38] Liao M H, Shi B G, Bai X, et al. TextBoxes: a fast text detector with a single deep neural network // *Thirty-First AAAI Conference on Artificial Intelligence*. San Francisco, 2017: 4161

[39] Liao M H, Shi B G, Bai X. TextBoxes++: a single-shot oriented scene text detector. *IEEE Trans Image Process*, 2018, 27(8): 3676

[40] He P, Huang W L, He T, et al. Single shot text detector with regional attention // *Proceedings of the* 2017 *IEEE International Conference on Computer Vision*. Venice, 2017: 3047

[41] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning // *Thirty-First AAAI Conference on Artificial Intelligence*. San Francisco, 2017: 4278

[42] Liao M H, Zhu Z, Shi B G, et al. Rotation-sensitive regression for oriented scene text detection // *Proceedings of the* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, 2018: 5909

[43] Liu X, Zhang R, Zhou Y S, et al. Scene text detection with feature pyramid network and linking segments // 2019 *International Conference on Document Analysis and Recognition* (*ICDAR*). Sydney, 2019: 508

[44] Zhang S, Liu Y L, Jin L W, et al. Feature enhancement network: a refined scene text detector // *Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, 2018: 2612

[45] Tian Z, Huang W L, He T, et al. Detecting text in natural image with connectionist text proposal network // *European Conference on Computer Vision*. Munich, 2016: 56

[46] Wang X B, Jiang Y Y, Luo Z B, et al. Arbitrary shape scene text detection with adaptive text region representation // *Proceedings of the* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 2019: 6449

[47] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN // *Proceedings of the* 2017 *IEEE International Conference on Computer Vision*. Venice, 2017: 2961

[48] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell*, 2017, 39(4): 640

[49] Li Y, Qi H Z, Dai J F, et al. Fully convolutional instance-aware semantic segmentation // *Proceedings of the* 2017 *IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, 2017: 4438

[50] Lyu P Y, Liao M H, Yao C, et al. Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes // *Proceedings of the European Conference on Computer Vision*. Munich, 2018: 67

[51] Liao M H, Lyu P Y, He M H, et al. Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes. *IEEE Trans Pattern Anal Machine Intelligence*, 2019: 1

[52] Xie E Z, Zang Y H, Shao S, et al. Scene text detection with supervised pyramid context network. *Proc AAAI Conf Artif Intell*, 2019, 33: 9038

[53] Zhang Z, Zhang C Q, Shen W, et al. Multi-oriented text detection with fully convolutional networks // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, 2016: 4159

[54] Long S B, Ruan J Q, Zhang W J, et al. TextSnake: a flexible representation for detecting text of arbitrary shapes // *Proceedings of the European Conference on Computer Vision*. Munich, 2018: 19

[55] He T, Huang W L, Qiao Y, et al. Accurate text localization in natural image with cascaded convolutional text network [J/OL]. *arXiv preprint* (2016-03-31)[2020-03-01]. https://arxiv.org/abs/1603.09423

[56] Deng D, Liu H F, Li X L, et al. PixelLink: Detecting scene text via instance segmentation [J/OL]. *arXiv preprint* (2018-01-04)[2020-03-01]. https://arxiv.org/abs/1801.01315

[57] Yang Q P, Cheng M L, Zhou W M, et al. IncepText: a new inception-text module with deformable PSROI pooling for multi-oriented scene text detection // *Proceedings of the* 27*th International Joint Conference on Artificial Intelligence*. Stockholm, 2018: 1071

[58] Dai Y C, Huang Z, Gao Y T, et al. Fused text segmentation networks for multi-oriented scene text detection //2018 24*th International Conference on Pattern Recognition* (*ICPR*). Beijing, 2018: 3604

[59] Wang W H, Xie E Z, Li X, et al. Shape robust text detection with progressive scale expansion network //*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 2019: 9328

[60] Wang W H, Xie E Z, Song X G, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network // *Proceedings of the* 2019 *IEEE/CVF International Conference on Computer Vision*. Seoul, 2019: 8439

[61] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition // 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Las Vegas, 2016: 770

[62] Baek Y, Lee B, Han D, et al. Character region awareness for text detection // *Proceedings of the* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 2019: 9357

[63] Tian Z T, Shu M, Lyu P Y, et al. Learning shape-aware embedding for scene text detection // *Proceedings of the* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 2019: 4229

[64] Liao M H, Wan Z Y, Yao C, et al. Real-time scene text detection with differentiable binarization [J/OL]. *arXiv preprint* (2019-12-03)[2020-03-01]. https://arxiv.org/abs/1911.08947

[65] Lyu P Y, Yao C, Wu W H, et al. Multi-oriented scene text detection *via* corner localization and region segmentation // *Proceedings of the* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, 2018: 7553

[66] Li Y, Yu Y J, Li Z F, et al. Pixel-Anchor: a fast oriented scene text detector with combined networks [J/OL]. *arXiv preprint* (2018-11-19)[2020-03-01].http://export.arxiv.org/abs/1811.07432

[67] Jiang F, Hao Z H, Liu X R. Deep scene text detection with connected component proposals [J/OL]. *arXiv preprint* (2017-08-17)[2020-03-01]. http://export.arxiv.org/abs/1708.05133

[68] Qiao L, Tang S L, Cheng Z Z, et al. Text perceptron: towards end-to-end arbitrary-shaped text spotting [J/OL]. *arXiv preprint* (2020-02-17)[2020-03-01]. https://arxiv.org/abs/2002.06820

[69] Zhou X Y, Yao C, Wen H, et al. EAST: an efficient and accurate scene text detector // *Proceedings of the* 2017 *IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, 2017: 2642

[70] Li J R, Zhou Z J, Su Z Z, et al. A new parallel detection-recognition approach for end-to-end scene text extraction // 2019 *International Conference on Document Analysis and Recognition* (*ICDAR*). Sydney, 2019: 1358

[71] He T, Tian Z, Huang W L, et al. An end-to-end TextSpotter with explicit alignment and attention // 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Salt Lake City, 2018: 5020

[72] Kim K H, Hong S, Roh B, et al. PVANET: Deep but lightweight neural networks for real-time object detection. *arXiv preprint* (2019-09-30)[2020-03-01]. https://arxiv.org/abs/1608.08021

[73] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition // *IEEE Conference on Computer Vision & Pattern Recognition*. Las Vegas, 2016: 770

[74] Wang F F, Zhao L M, Li X, et al. Geometry-aware scene text detection with instance transformation network // *Proceedings of the* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, 2018: 1381

[75] Duan J Q, Xu Y J, Kuang Z H, et al. Geometry normalization networks for accurate scene text detection // *Proceedings of the* 2019 *IEEE/CVF International Conference on Computer Vision*. Seoul, 2019: 9136

[76] Liu Z C, Lin G S, Yang S, et al. Towards robust curve text detection with conditional spatial expansion // *Proceedings of the* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 2019: 7261

[77] Liu Y L, Chen H, Shen C H, et al. ABCNet: real-time scene text spotting with adaptive Bezier-curve network [J/OL]. *arXiv preprint* (2020-02-25)[2020-03-01]. https://arxiv.org/abs/2002.

10200v2

[78] Wang H, Lu P, Zhang H, et al. All you need is boundary: toward arbitrary-shaped text spotting [J/OL]. *arXiv preprint* (2019-11-21)[2020-03-01]. https://arxiv.org/abs/1911.09550

[79] Zhang A X. *Research on Natural Scene Text Detection Algorithms Based on Deep Learning* [Dissertation]. Beijing: North China University of Technology, 2019
(张艾萱. 基于深度学习的自然场景文本检测算法研究[学位论文]. 北京: 北方工业大学, 2019)

[80] Zhou X Y, Gao Z H. Research on inclined text location method of natural scene based on YOLO. *Comput Eng Appl*, 2020, 56(9): 213
(周翔宇, 高仲合. 基于YOLO的自然场景倾斜文本定位方法研究. 计算机工程与应用, 2020, 56(9): 213)

[81] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection // 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Las Vegas, 2016: 779

[82] Niu Z D, Li H D. Natural scene text detection algorithm with attention mechanism. *Comput Appl Software*, 2019, 36(9): 198
(牛作东, 李捍东. 引入注意力机制的自然场景文本检测算法研究. 计算机应用与软件, 2019, 36(9): 198)

[83] Yuan T L, Zhu Z, Xu K, et al. Chinese text in the wild [J/OL]. *arXiv preprint* (2018-02-26)[2020-03-01]. https://arxiv.org/abs/1803.00085

[84] Lucas S M, Panaretos A, Sosa L, et al. ICDAR 2003 robust reading competitions // *Seventh International Conference on Document Analysis and Recognition*. Edinburgh, 2003: 682

[85] Veit A, Matera T, Neumann L, et al. COCO-Text: dataset and benchmark for text detection and recognition in natural images [J/OL]. *arXiv preprint* (2016-06-19)[2020-03-01]. https://arxiv.org/abs/1601.07140

[86] Shi B G, Yao C, Liao M H, et al. ICDAR2017 competition on reading Chinese text in the wild (RCTW-17) // 2017 14*th IAPR International Conference on Document Analysis and Recognition* (*ICDAR*). Kyoto, 2017: 1429

[87] Ch'ng C K, Chan C S. Total-Text: a comprehensive dataset for scene text detection and recognition // 2017 14*th IAPR International Conference on Document Analysis and Recognition* (*ICDAR*). Kyoto, 2017: 935

[88] Nayef N, Yin F, Bizid I, et al. ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification-RRC-MLT // 2017 14*th IAPR International Conference on Document Analysis and Recognition* (*ICDAR*). Kyoto, 2017: 1454

[89] Chng C K, Liu Y L, Sun Y P, et al. ICDAR2019 robust reading challenge on arbitrary-shaped text-RRC-ArT. // 2019 *International Conference on Document Analysis and Recognition* (*ICDAR*). Sydney, 2019: 1571

[90] Liu Y L, Jin L W, Zhang S T, et al. Detecting curve text in the wild: new dataset and new solution. *arXiv preprint* (2017-12-06)[2020-3-1]. https://arxiv.org/abs/1712.02170

[91] Wang J X, Wang Z Y, Tian X. Review of natural scene text detection and recognition based on deep learning. *J Software*, 2020, 31(5): 1465
(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)

[92] Liu Y L. Jin L W. Zhang S T, et al. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognit*, 2019, 90: 337

[93] Peng B F. Tencent cloud university big players share | decryption OCR text recognition technology [EB/OL]. *Tencent Cloud Community Column* (2019-08-13) [2020-03-01]. https://cloud.tencent.com/developer/article/1473262
(彭碧发. 腾讯云大学大咖分享 | 解密OCR文字识别技术 [EB/OL] 腾讯云社区专栏 (2019-08-13) [2020-03-01]. https://cloud.tencent.com/developer/article/1473262)

[94] Youdao Z Y. Text recognition OCR service [EB/OL]. *Youdao Intelligent Cloud·AI Open Platform* (2019-12-17) [2020-03-01]. https://ai.youdao.com/product-ocr.s
(有道智云. 文字识别OCR服务 [EB/OL]. 有道智云·AI开放平台 (2019-12-17) [2020-03-01]. https://ai.youdao.com/product-ocr.s)

[95] Baidu Clound Engine. Universal text recognition [EB/OL]. *Baidu Intelligent Cloud* (2020-02-05) [2020-03-01]. https://cloud.baidu.com/product/ocr/general
(百度云. 通用文字识别 [EB/OL]. 百度智能云 (2020-02-05) [2020-03-01]. https://cloud.baidu.com/product/ocr/general)

[96] Chuangyejun. Chuang Lan 253- the image recognition OCR technology of Chuanglan Myriads platform [EB/OL]. *Chuang Lan 253 Column* (2018-07-19) [2020-03-01]. https://blog.csdn.net/chuangyejun/article/details/81113833
(Chuangyejun. 创蓝253-创蓝万数平台图像识别OCR技术 [EB/OL] 创蓝253专栏 (2018-07-19) [2020-03-01]. https://blog.csdn.net/chuangyejun/article/details/81113833)

[97] ZJULearning. Pixel_link [EB/OL]. *GitHub* (2019-11-21) [2020-03-01]. https://github.com/ZJULearning/pixel_link

[98] Huoyijie. AdvancedEAST [EB/OL]. *GitHub* (2020-4-3) [2020-03-01]. https://github.com/huoyijie/AdvancedEAST

[99] Dengdan. Seglink [EB/OL]. *GitHub* (2018-5-3) [2020-03-01]. https://github.com/dengdan/seglink

[100] Tianzhi0549. CTPN [EB/OL]. *GitHub* (2020-4-3) [2020-03-01]. https://github.com/tianzhi0549/CTPN

[101] Tian Z, Huang W L, He T, et al. Detecting text in natural image with connectionist text proposal network // *ECCV* 2016: *European Conference on Computer Vision*. Amsterdam, 2016: 56