自然场景图像中的文本检测综述

王润民[1,2] 桑 农[3] 丁 丁[4] 陈 杰[5] 叶齐祥[6] 高常鑫[3] 刘 丽[2,5]

摘 要 本文对自然场景文本检测问题及其方法的研究进展进行了综述. 首先, 论述了自然场景文本的特点、自然场景文本检测技术的研究背景、现状以及主要技术路线. 其次, 从传统文本检测以及深度学习文本检测的视角出发, 梳理、分析并比较了各类自然场景文本检测方法的优缺点, 并介绍了端对端文本识别技术. 再次, 论述了自然场景文本检测技术所面临的挑战, 探讨了相应的解决方案. 最后, 本文列举了测试基准数据集、评估方法, 将最具代表性的自然场景文本检测方法的性能进行了比较, 本文还展望了本领域的发展趋势.

关键词 文本检测, 场景文本, 深度学习, 手工设计的特征, 连通域分析

引用格式 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述. 自动化学报, 2018, **44**(12): 2113−2141

DOI 10.16383/j.aas.2018.c170572

Text Detection in Natural Scene Image: A Survey

WANG Run-Min[1,2] SANG Nong[3] DING Ding[4] CHEN Jie[5] YE Qi-Xiang[6] GAO Chang-Xin[3] LIU Li[2,5]

Abstract In this paper, the research progress of natural scene text detection problems and methods are reviewed. Firstly, the characteristics of natural scene text are introduced, and the text detection technology research background, status and the main technical route are illustrated respectively. Secondly, from the perspective of traditional text detection and deep learning text detection, the merits and demerits of various methods are analyzed, and the technology of end-to-end text recognition is introduced. Then, the challenges of the natural scene text detection technology and the corresponding solutions are discussed. Finally, the benchmark datasets are enumerated, the evaluation methods and the performances of the most representative approaches are fundamentally compared. Furthermore, potential application and development trend in this field are summarized.

Key words Text detection, scene text, deep learning, handcraft feature, connected component analysis

Citation Wang Run-Min, Sang Nong, Ding Ding, Chen Jie, Ye Qi-Xiang, Gao Chang-Xin, Liu Li. Text detection in natural scene image: a survey. *Acta Automatica Sinica*, 2018, **44**(12): 2113−2141

收稿日期 2017-10-10 录用日期 2018-03-12
Manuscript received October 10, 2017; accepted March 12, 2018
国家自然科学基金 (61502164), 湖南省自然科学基金 (2016JJ3090), 湖南省教育厅优秀青年项目 (16B155), 中国博士后科学基金 (2015T81130, 2014M562569) 资助
Supported by National Natural Science Foundation of China (61502164), Natural Science Foundation of Hunan Province (2016JJ3090), Foundation of Hunan Provincial Education Department (16B155), and China Postdoctoral Science Foundation (2015T81130, 2014M562569)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin

1. 湖南师范大学物理与信息科学学院 长沙 410081 中国 2. 国防科技大学信息系统与管理学院 长沙 410000 中国 3. 华中科技大学自动化学院 武汉 430074 中国 4. 国防科技大学教研保障中心 长沙 410000 中国 5. 奥卢大学电气与信息工程系 奥卢 FI-90014 芬兰 6. 中国科学院大学电子电气与通信工程学院 北京 101408 中国
1. School of Physics and Information Science, Hunan Normal University, Changsha 410081, China 2. Department of Information System and Management, National University of Defense Technology, Changsha 410000, China 3. School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China 4. Teaching and Research Support Center, National University of Defense Technology, Changsha 410000, China 5. Department of Electrical and Information Engineering, University of Oulu, Oulu, FI-90014, Finland 6. School of Electronics, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China

文本作为人类文明的标志、信息交流的载体, 广泛地存在于自然场景图像中 (如: 路牌、商品名称、车辆牌照等), 相较图像中的其他自然场景内容 (如: 树木、行人、建筑物等), 自然场景文本具有更强的逻辑性与更概括的表达性, 能更加有效地提供高层语义信息, 准确地识别图像中的文本将有助于场景内容的分析与理解.

1 研究背景与意义

基于文本具有高度的概括性和抽象的描述能力, 自然场景文本检测技术在智能交通系统[1−3]、视障人导盲[4−5]、基于内容的图像/视频检索[6] 以及可穿戴/便携式视觉系统[7−10] 等方面具有重要的应用. 随着互联网技术以及便携式移动设备的高速发展, 越来越多的应用场景需要利用图像中的文本信息. 目前自然场景文本检测已成为计算机视觉与模式识别、文档分析与识别领域的一个研究热点, 一些国际顶级会议, 如: CVPR、ICCV、ECCV,

已将自然场景文本检测列为其重要主题之一. 特别是自 2003 年以来, 作为文档分析与识别领域最重要的国际学术会议–文档分析与识别国际会议 (International Conference on Document Analysis and Recognition, ICDAR) 定期组织自然场景文本检测竞赛, 通过竞赛对该领域研究现状、发展趋势进行分析, 及时地跟踪并推动该技术的研究发展.

目前, 自然场景文本检测问题已受到国内外研究人员的广泛关注, 一些国外研究团队, 比如牛津大学视觉几何组 (Visual geometry group, VGG)、捷克理工大学机器感知中心 Jiri Matas 组、日本九州大学 Seiichi Uchida 组以及微软亚洲研究院等, 在该领域取得了一些里程碑式的研究成果. 国内研究机构与学者在文档分析与识别领域也发挥着举足轻重的作用[11−21]. 一些国内研究者, 比如, 中科院自动化所刘成林研究员、华中科技大学白翔教授、北京科技大学殷绪成教授、华南理工大学金连文教授、中国科学院大学叶齐祥教授以及中科院深圳先进技术研究院乔宇研究员、黄伟林博士等在历届 ICDAR 自然场景文本检测竞赛中获得了令人瞩目的成绩. 特别是, 华中科技大学白翔教授受邀作为 ICDAR 自举办 26 年来以来第一位来自中国的主讲嘉宾在日本京都举办的 ICDAR 2017 上作大会特邀报告, 展现了中国学者在此领域的影响力. 一些国内研究机构, 比如中科院自动化所、北京大学、清华大学、华中科技大学、北京科技大学、三星中国研究院、腾讯、百度、旷视科技等, 在 ICDAR 组织的一系列活动中表现活跃. 2011 年, 由清华大学与中科院自动化所合办的第十一届文档分析与识别国际会议 (ICDAR 2011) 在北京举办 (ICDAR 首次在国内举办), 清华大学丁晓青教授担任了大会主席. 2017 年, 由华中科技大学白翔教授等组织了 ICDAR 2017 自然场景中的中文文本识别竞赛 (RCTW-17), 共有来自高校、企业的 17 支队伍参赛了 RCTW-17 竞赛[1]. 2017 年, 中科院自动化所刘成林研究员领导的模式分析与学习团队 (PAL 团队) 与法国拉罗切尔大学、三星中国研究院等单位合作举办了多语言场景文本检测与语种判别的竞赛, 发布了包括 9 种语言, 18 000 幅图像的多语言场景文本数据库.

国内研究团队在包括 TPAMI、TIP、PR、CVPR 等各类主流国际期刊、会议的投稿数量逐年增加[11−19] , 在本次 ICDAR 2017 会议中来自国内学者的投稿论文数高居第一. 此外, 国内研究团队在该领域各项竞赛中也取得了瞩目的成绩, 在第 14 届国际文档分析与识别会议 (ICDAR 2017) 所组织的各项技术竞赛中, 中科院自动化所刘成林研究员领导的 PAL 团队在页面目标检测、中世纪文档版面分析、视频阿拉伯文本检测与识别、中文场景文本阅读等竞赛中获得了 8 项任务的第一名、2 项任务的第二名的突出成绩. 华南理工大学金连文教授带领的团队, 通过构建高性能的基于深度学习的文本检测与识别系统, 在场景文本检测, 端到端场景文本检测及识别两项任务中取得第一名的好成绩 (后者较其他参赛团队具有明显的优势), 在语种分类任务中以 0.4 % 的微弱差距位居第二名. 北京科技大学殷绪成教授团队再次 (连续三届) 荣获鲁棒阅读竞赛冠军.

尽管国内学者在自然场景文本检测领域取得了一些令人瞩目的成果, 在本领域重要的外文期刊上也发表了英文综述性论文[22−25], 然而我们以自然场景文本检测为关键词在国内中文期刊数据库中进行检索时却遗憾地发现, 除了出现个别手写文本识别的综述 [26] 外, 关于自然场景文本检测的中文综述几近空白. 据我们所知, 最近的英文综述 [22 − 25] 发表至今已逾两年, 然而在这两年以来, 一些新的测试数据库与一些新的检测结果的推出, 以及一些新型深度学习方法在自然场景文本检测领域的应用都极大地推动了相关技术的发展. 此外, 计算机视觉领域中的一些新的研究成果, 比如目标显著性检测、视觉上下文等, 也被引入到自然场景文本检测领域, 进一步提升了文本检测性能. 基于上述情况, 有必要对自然场景文本检测领域的相关研究工作进行全面综述和讨论. 本文系统综述了自然场景文本检测技术的研究进展和目前面临的挑战与困难, 以期为研究人员进一步深入研究自然场景文本检测以及拓展其应用领域提供帮助, 并期待能够启发更多的创新性工作.

本文首先论述了自然场景文本检测的研究背景、现状、自然场景文本特点以及主要技术路线. 接下来, 梳理、分析并比较了各类自然场景文本检测方法的动机、原理、优势与不足, 揭示了各类方法之间的区别与联系. 本文还介绍了端对端文本识别技术, 阐述并讨论了文本显著性、视觉上下文等其他领域知识在自然场景文本检测中的应用. 此外, 本文还论述了自然场景文本检测技术所面临的挑战, 并探讨了相应的解决方案. 列举了测试基准数据集、评估方法, 将最具代表性的自然场景文本检测方法的性能进行了比较. 最后, 给出了我们对该领域发展的一些思考.

2 研究现状

相对人脸检测、印刷体文档中的光学字符检测等经典问题, 自然场景文本检测研究还相对滞后, 直到 20 世纪 90 年代中期才开始出现该领域的研究

[1] 竞赛结果链接: http://mclab.eic.hust.edu.cn/icdar2017chinese/result.html

报道[27−29]. 目前, 自然场景文本检测已成为计算机视觉领域的热门研究课题, 吸引了国内外众多的研究机构与学者开展该课题的研究. 特别是国际文档分析与识别会议 (ICDAR) 定期举办的各项技术竞赛极大地推动了该领域的发展, 从而使得自然场景文本检测技术的瓶颈与难题不断地被突破. 比如在 2011 年, ICDAR 2011 自然场景文本检测竞赛冠军[30] 所获得的结果为召回率 (Recall) 0.63, 准确率 (Precision) 0.83, 综合指标 (F-measure) 0.71. 而在 2017 年, 文献 [31] 公布其在 ICDAR 2011 自然场景文本检测数据库上所获得的指标为召回率 (Recall) 0.82, 准确率 (Precision) 0.89, 综合指标 (F-measure) 0.86. 再如在 2015 年, ICDAR 2015 非受限环境下的自然场景文本 (Incidental scene text) 检测 (Task 4.1) 竞赛冠军[32] 的指标为召回率 (Recall) 0.37, 准确率 (Precision) 0.77, 综合指标 (F-measure) 0.50. 在 2017 年, 文献 [33] 公布对 ICDAR 2015 非受限环境下的自然场景文本检测 (Task 4.1) 所获得的结果为召回率 (Recall) 0.77, 准确率 (Precision) 0.73, 综合指标 (F-measure) 0.75. 由此可见, 自然场景文本检测技术在近几年取得了长足的发展.

目前针对自然场景文本处理的研究工作主要包括三个方面: 自然场景文本检测、自然场景文本识别、以及端对端 (End-to-end) 自然场景文本检测与识别. 分析 2017 年发表在 CVPR、ICCV、NIPS、IJCAI、AAAI、ICDAR 等各类顶级会议上的相关论文, 超过 80 % 的自然场景文本检测论文主要关注多方向排列的文本检测问题, 大部分文献主要处理英文文本, 较少的文献涉及自然场景文本识别以及端对端自然场景文本检测与识别问题. 从自然场景文本检测技术的处理对象来看, 主要经历了水平方向排列的文本检测[34−37] 到多方向排列的文本检测[15, 33, 38−42], 从单一的英文、阿拉伯数字的文本检测[34−36] 到多语种的文本检测[37, 41, 43−44]. 从自然场景文本检测所采用的描述特征来看, 主要经历了两个阶段: 首先是基于传统手工设计的特征 (Handcraft features), 然后在 2014 年前后出现了基于深度学习的自然场景文本检测方法[19, 31, 45−49]. 一些深度学习技术, 比如: 卷积神经网络 (Convolutional neural networks, CNN) 以及递归神经网络 (Recurrent neural networks, RNN) 等在自然场景文本检测领域得到了很好的应用, 目前采用深度学习方法检测自然场景文本已成为了该领域研究的主要技术手段.

不同于印刷体文档中的文本, 自然场景文本的字体大小、颜色、排列方向、稀疏性、对比度等有着很大的差异. 与此同时, 还受到光照变化、复杂背景、噪声干扰、拍摄视角等方面的影响. 尽管对仅包含英文与数字的 ICDAR 2011 数据集已取得综合指标 (F-measure) 0.86[31] 以及 ICDAR 2015 数据集已取得综合指标 (F-measure) 0.81 的性能[50], 但从最近刚落幕的 ICDAR 2017 自然场景中的中文文本识别竞赛 (RCTW-17)[51] 所公布的结果来看, 竞赛冠军所取得的准确率为 0.74, 召回率为 0.59, 综合指标为 0.66. 从上述文本检测结果中不难发现, 现有的自然场景文本检测技术受限于被检测文本的语种、排列方向、数据集公布的时间、数据集的规模等, 由此表明该技术的成熟度、鲁棒性还远非达到实用水平. 综上所述, 基于自然场景文本检测技术的研究现状以及该技术广阔的应用前景, 对该领域的深入研究具有重要的理论意义与应用价值.

3 自然场景文本的特点

图像中的文本根据其形成方式可以划分为场景文本和叠加文本[52]. 叠加文本是人们为了某种目的而后期人为添加的注释性内容 (比如新闻内容摘要、影视台词、体育赛况介绍等, 如图 1 所示); 场景文本作为固有的自然场景内容随机地存在于图像当中 (比如道路指示牌、商品名称、广告牌等, 如图 2 所示[2]). 与叠加文本相比, 自然场景文本在字体大小、类型、颜色、排列方向等方面具有更大的复杂性. 自然场景文本区域的视觉特性主要体现在以下几个方面:

1) 对比度属性: 文本作为人类社会人际沟通的信息载体, 可读性是自然场景文本具有存在意义的基本要求. 自然场景文本相对其背景区域而言, 其灰度、颜色信息往往具有较明显的对比度.

2) 梯度、边缘属性: 文本通常具有复杂的空间结构, 因此文本区域往往存在相对密集的边缘以及较明显的梯度信息.

3) 灰度、颜色属性: 为了视觉上的舒适性, 往往选择与背景有明显反差的颜色进行书写, 且文本的颜色与灰度分布均匀.

4) 笔画宽度属性: 尽管文本由不同的笔画组成, 但同一个文本中的笔画宽度近似相等.

5) 几何大小属性: 自然场景文本具有任意的尺寸大小, 但为了满足人眼视觉要求, 自然场景文本的尺寸通常满足一定的变化范围.

6) 视觉上下文属性: 在同一个文本行区域内, 相邻文本之间具有笔画宽度、颜色、高度以及像素灰度值相近等特点.

7) 空间分布属性: 文本在图像中通常以文本行的形式存在, 相邻文本之间的排列方式是任意的, 其

[2] 示例图像源自 ICDAR 系列数据库以及 MSRA-TD500 数据库.

图 1 叠加文本示例

Fig.1 Examples of overlay text

(a) (b) (c) (d) (e)

图 2 自然场景文本示例

Fig.2 Examples of natural scene text

间隔距离通常满足一定的规律.

4 自然场景文本检测所面临的挑战

除了受到复杂背景、光照变化、拍摄视角等外界因素影响外, 相比一般性物体检测问题, 自然场景文本检测技术还面临着自身特征的诸多挑战.

1) 从文本的种类来看: 自然场景文本包含了不同的语种, 每一种语种包含了数量规模不等的文本类型, 比如英文包括了 52 个大、小写字母, 而中文汉字所包含的文本类别则更多, 仅 1980 年制定的国标 GB2312-80 编码就定义了多达 6 763 类常用汉字, 不同语种或者相同语种的不同文本类型之间的视觉特征具有很大的差异.

2) 从文本排列方向来看: 相邻自然场景文本之间的排列是沿任意方向的, 从而导致很难找到合适的描述特征与边界框来表示文本区域.

3) 从文本行的组成来看: 自然场景文本行通常由不同类别的文本所构成, 尽管每一种文本具有固定的空间结构, 但是将不同类别的文本组合成文本行时就会呈现出杂乱的视觉特征. 正因为不同文本行之间的结构共性很少, 所以我们很难找到一个区分性好的描述特征来表达文本行区域.

4) 从文本行的大小及长宽比来看: 文本行大小不一, 且其长度与宽度的比值也不确定. 对文本行进行检测时, 我们不仅需要考虑尺度大小问题, 而且还需要考虑文本行长宽比问题, 从而大大增加了文本行检测的难度.

5) 从文本行检测结果来看: 文本行检测通常需要检测一个文本序列. 根据算法性能评测要求 (比如 ICDAR 竞赛测评标准), 所有的自然场景文本检测算法需要得到以单词为单元的检测结果, 而不同单词包括的字符数有所不同, 单词之间的间隔距离也会经常受到单个文本检测结果的影响, 因此自然场景文本检测较传统的独立目标检测更具挑战性.

5 自然场景文本检测方法

20 世纪 90 年代中期, 文献 [27−29] 等率先开展了自然场景文本检测研究. 经过 20 余年的发展, 该领域涌现出大量行之有效的解决方法. 特别是近年来目标检测技术与语义分割技术的快速发展使得文本检测领域取得了重大的突破. 依据文本检测技术的发展历程及文本区域描述特征分类标准, 自然场景文本检测方法大体上可以归纳为: 1) 传统的自然场景文本检测方法; 2) 基于深度学习的自然场景文本检测方法.

5.1 传统的自然场景文本检测方法

传统的自然场景文本检测方法主要沿用两条技术路线; 基于连通域分析的方法和基于滑动检测窗的方法. 该方法首先获得文本候选区域, 然后采用传统手工设计的特征 (Handcraft features) 对所获得的候选区域进行验证, 并最终获得文本位置信息.

5.1.1 基于连通域分析的方法

基于连通域分析的方法采用自底向上策略检测文本. 从获得连通域的途径来看可以分为边缘检测方法和文本级检测方法, 该类方法通常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.

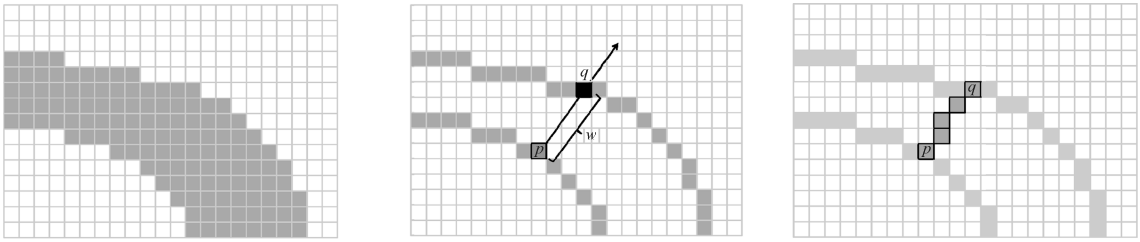
1) 边缘检测方法

考虑到自然场景文本具有丰富的边缘以及角点信息, 该类方法主要通过检测边缘或者角点等方式来获得文本候选区域, 然后对所获得的文本候选区域利用规则或者分类器进行分类. 文献 [38, 53 − 57] 等采用了一些边缘检测算子 (如: Sobel, Canny 等) 检测出图像的边缘信息, 然后对边缘图像进行形态学处理以剔除伪文本区域. 文献 [55] 首先提取水平、垂直、左上、右上方向边缘图像, 然后基于上述边缘图像采用 K 均值 (K-means) 聚类方法获得初始的文本区域检测结果, 最后对初始的检测结果采用经验规则以及投影分析来进行验证. 文献 [38] 采用 K 均值聚类方法对傅里叶−拉普拉斯滤波处理后的图像像素进行分类从而获得文本连通区域, 通过对各个连通区域的骨架进行分析, 将连通区域分为 “简单” 和 “复杂” 两类, 保留简单的连通区域并对复杂的连通区域进行进一步分析, 最后根据文本行平直度以及边缘密度等特征对文本候选区域进行判断以去除背景区域. 文献 [57] 首先提取边缘, 然后通过候选边缘重组以及识别的方法获得文本区域. Busta 等在文献 [53] 中通过定制 Fast 角点使其更有利于场景文本检测, 根据文献 [53] 报道的结果, 该方法所获得的场景文本检测召回率 (Recall) 较传统 MSER 方法高 25 %, 且速度是传统 MSER 方法的 4 倍以上. 除此以外, Jiri Matas 课题组还参与推出了 COCO-Text 自然场景文本数据集[58].

2) 文本级检测方法

该类方法利用自然场景文本通常具有像素灰度值近似相等、颜色近似相同以及笔画宽度相近等特点, 对自然场景图像进行特定处理后, 文本中的相邻像素在其空间结构上表现出连通性, 该类方法通过检测图像中的连通区域来获得文本候选区域. 为了获得文本连通区域, 该类方法采取了许多行之有效的手段, 比如极值区域 (Extremal regions, ERs)[47, 49, 59]、最大稳定极值区域 (Maximally stable extremal regions, MSER)[18, 60−64]、颜色对比度增强极值区域 (Color-enhanced contrasting extremal region, CER)[47, 65]、颜色聚类方法 (Color clustering)[66−68]、笔画宽度变换 (Stroke width transform, SWT)[34, 69−71]、笔画特征变换 (Stroke feature transform, SFT)[72]、级联空间变换[43]、图割二值化[35, 73−74]、手工阈值分割[75] 等. 在文本级检测方法中, 首先将图像分割成若干个连通区域, 然后对每一个连通区域的几何特征进行分析, 利用文本候选区域的边缘密度[38]、前景像素密度[18]、长宽比[34]、文本候选区域紧致度[13]、轮廓梯度[13]、笔画宽度变化率[18, 34]、平均方向偏差[76] 等特征通过设定判断规则, 或者提取文本区域的描述特征并结合已训练好的分类器对其进行判断, 从而将之分类为文本区域与背景区域.

在文本级检测方法中, 最为代表性的方法主要包括: 笔画宽度变换 (SWT)[34]、最大稳定极值区域 (MSER)[60] 等. 笔画宽度变换算法由 Epshtein 等[34] 于 2010 年首次提出 (如图 3 所示), 该方法主要利用了位于同一个文本中的笔画具有宽度近似相等的性质来获取文本候选区域. 在实施笔画宽度变换的过程中, 首先利用 Canny 算子对输入图像进行边缘检测, 并计算边缘像素点的梯度方向, 沿着梯度方向的路线寻找与之匹配的像素. 匹配像素 p 与 q 之间搜索路线上的每一个像素值被指定为上述两个像素之间的笔画宽度 (即像素点 p 与像素点 q 之间的欧氏距离). 对于某个像素而言, 若其属于多个搜索线路, 则其像素值为上述搜索线路对应的最小笔画宽度值. 文献 [39] 采用图 4 所示的检测框架, 较早地实现了任意方向排列的自然场景文本检测任务. 该文献通过笔画宽度变换 (SWT) 处理获得文本候选区域, 用文本级分类器 (简单特征 + 随机森林) 过滤非文本区域; 利用文本间的相似性连接成文本行, 再用文本行级的分类器 (简单特征 + 随机森林) 进一步过滤背景区域. 采用笔画宽度变换 (SWT) 处理可以提取出不同尺度和方向的文本候选区域, 然而该方法在图像边缘检测不准确以及背景复杂的情况下鲁棒性较差, 此外, 笔画宽度变换的运算效率



(a) (b) (c)

图 3 基于笔画宽度变换的自然场景文本检测[34]

Fig. 3 Natural scenes text detection based on stroke width transformation[34]

(a) (b)

图 4 任意方向文本检测方法[39]

Fig. 4 Detecting texts of arbitrary orientations in natural images[39]

也受到图像边缘像素数目的影响. 针对笔画宽度变换方法的一些不足, 在后续研究[70, 72] 中也出现了一些笔画宽度变换的变体, 比如文献 [72] 考虑到传统的笔画宽度变换方法在应对图像中包含一些具有不规则梯度方向的边缘时往往不能准确地计算出笔画宽度, 该文献利用了颜色信息来改进笔画宽度算子并提出了笔画特征变换 (Stroke feature transform) 算子. 最大稳定极值区域 (MSER) 基于分水岭的概念, 该方法取 $[0, 255]$ 范围的阈值对图像进行二值化处理, 所获得的二值化图像经历了一个从全黑到全白的过程 (犹如水位不断上升的俯瞰图). 在此过程中, 有些连通区域面积随阈值上升的变化很小, 定义该类区域为最大稳定极值区域 (MSER). 根据 MSER 的工作原理, 检测得到的 MSER 内部灰度值是小于边界的, 因此通过 MSER 方法检测不到位于黑色背景中的白色区域. 在实际处理的过程中, 通常需要对原图进行一次 MSER 检测后将其反转, 然后再做一次 MSER 检测, 上述两次操作分别称 MSER+ 和 MSER−. Neumann 等[60] 提出将 MSER 方法应用于自然场景文本检测 (如图 5 所示), 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. 最大稳定极值区域能够很好地描述文本内部颜色的一致性, 且可以克服噪声和仿射变换等因素的影响. 一些文献 [18, 62−64] 采用 MSER 方法在复杂的自然场景图像上取得了优异的文本检测性能. 此后, 在文献 [60] 的基础上, 最大稳定极值区域的一些变体[20, 47, 77−80] 相继被提出, 比如文献 [20, 80] 利用梯度信息来增强 MSER, 并提出了 Edge-preserving MSER 算子. 文献 [81] 采用局部自适应阈值方法来增强 MSER. 考虑到 MSER 在处理 “低对比度” 图像时不够鲁棒, Neumann 等在文献 [59] 中提出直接用极值区域 (ER) 来作为文本候选区域, 并设计了一套能够快速去除明显非文本区域的方法. Sun 等考虑到文献 [59] 所获得的极值区域的数量过大, 进而对后续的文本分类精度产生影响, 因此在文献 [82] 中提出了对比极值区域 (Contrasting extremal region, CER) 方法. 文献 [82] 所得到的 CER 是跟周围的背景有一定对比度的极值区域, 其数量远小于极值区域 (ER), 略多于最大稳定极值区域 (MSER), CER 应对 “低对比度” 图像更为鲁棒. 此后, Sun 等在文献 [83] 中又提出了颜色增强的对比极值区域 (Color-enhanced CER).

值得注意的是, 区别于前述边缘检测方法以及文本级检测方法. 一些文献 [84−87] 利用文本行上下边缘近似平行或者文本行的上下部分近似对称的性质, 实现对文本行候选区域的检测处理, 该类文

本行级检测方法主要应用于印刷体/手写文档中的文本处理[84−86]. 文献 [87] 利用了自然场景文本行上下结构相似的特点, 创新性地实现对场景文本的有效检测. 文献 [87] 设计了一个具有对称性的模板 (如图 6 (g) 所示), 通过该模板获得文本区域的自相似度与区分度, 即: 上半部和下半部的对称性、文本区域的上半部与背景的差异、文本区域的下半部与背景的差异等特征. 该模板在不同尺度下扫描图像, 通过其响应得到对称的中心点, 在得到对称中心点后通过文本的高度和连通性得到候选区域. 与传统的文本检测方法所采用的手工设计的特征所不同的是, 文献 [87] 使用了卷积神经网络 (CNN) 进行后续处理. 文本行级检测方法能有效地减少单个文本检测失误所带来的负面影响, 但该方法对文本行的边缘检测结果以及边缘对称性较为敏感.

基于连通域分析的自然场景文本检测方法主要通过提取图像中的连通区域来获得文本候选区域, 从而能有效地减少自然场景文本的搜索范围. 该类方法依赖于文本连通区域的检测结果, 连通区域的检测结果不仅影响文本检测召回率, 而且还会影响文本轮廓的准确性. 文本欠分割、过分割的处理结果将势必影响该文本候选区域的准确性, 进而对整个自然场景文本检测结果产生负面影响, 因此在保证文本连通区域检测高召回率的情况下, 获得准确的文本轮廓是提高文本检测性能的重要途径. 事实上, 在复杂的自然场景图像中准确地检测出文本连通区域是一件非常困难的事情, 光照变化、颜色褪色、噪声干扰等因素都将可能导致相邻文本出现粘连现象, 从而极大地影响文本检测系统的性能. 与此同时, 对每一个作为文本候选区域的连通区域进行验证时, 设计一个合理的连通区域分析器也是一件非常困难的事情. 受上述因素的影响, 基于连通区域分析的自然场景文本检测方法在背景复杂、噪声干扰、低对比度以及颜色多变等情况下难以鲁棒地检测自然场景文本.

5.1.2　基于滑动检测窗的方法

基于滑动检测窗的方法采用了自顶向下策略检测文本, 该类方法[88−92] 通过采用滑动检测

图 5　基于最大稳定极值区域的自然场景文本检测[18]

Fig. 5　Natural scenes text detection based on maximally stable extremal regions[18]

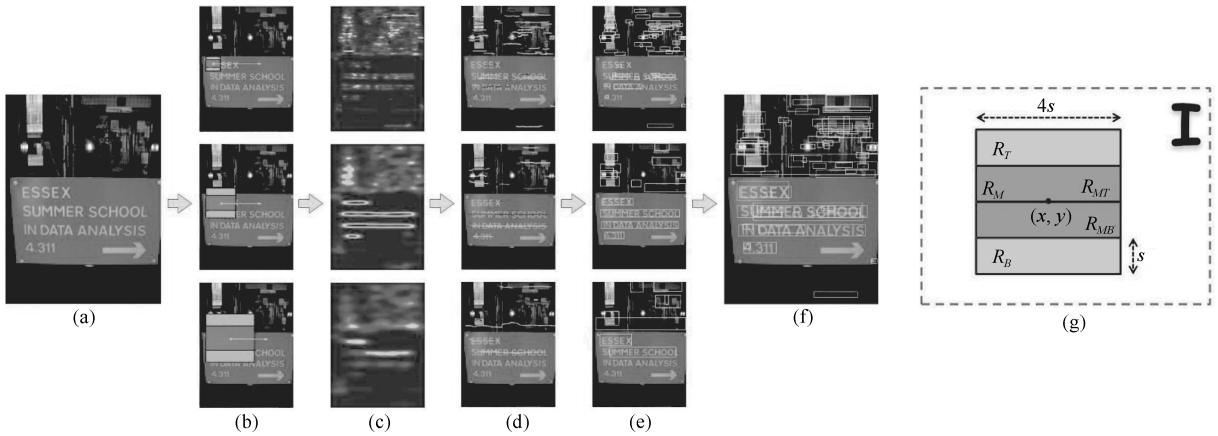


图 6　基于对称性的自然场景文本行检测[87]

Fig. 6　Symmetry-based text line detection in natural scenes[87]

窗口的方式对整幅自然场景图像进行扫描, 将每一个检测窗口所覆盖的图像区域视为文本候选区域, 然后提取文本候选区域的手工设计的特征, 结合已训练好的分类器获得该文本候选区域的置信度值, 通过将所获得的文本候选区域的置信度值与所设定的置信度阈值进行比较, 将文本候选区域分类为文本区域或者背景区域. 为了有效地应对文本大小、文本行长度多变的情况, 该类方法采用了多尺度滑动窗口的方式来得到文本候选区域. 基于滑动检测窗的自然场景文本检测方法主要采用了二种技术途径: 1) 一般性滑窗方法; 2) 基于特定单词的方法.

自然场景文本检测属于一种特定目标检测, 级联自适应增强算法 (Cascaded adaboost) 结合类哈尔 (Haar-like) 特征在人脸检测[93−95] 领域获得了巨大成功, 该技术方案为自然场景文本检测提供一种解决思路, 比如文献 [89, 96−98] 采用了 Adaboost 方法检测自然场景文本. 文献 [97] 采用级联 Adaboost 方式从一组特征池中选择 79 个特征并训练得到 4 个强分类器. 此后, 在文献 [96] 中, 他们进一步将文献 [97] 中的特征提取方法扩充至 6 种, 即 $X-Y$ 方向梯度信息、Gabor 滤波器的局部能量、图像纹理的统计直方图、图像小波变换系数的方差、边缘间距以及连通区域分析等, 进而使得检测性能得到了较大幅度的改善. 然而值得注意的是, 文献 [93 − 95] 之所以能够快速地实现人脸检测, 主要依赖于其采用的两个关键技术: 1) 构建了一种有效的级联检测框架; 2) 使用了一种计算高效且对人脸分类性能好的类哈尔 (Haar-like) 特征. 然而自然场景文本与人脸的视觉特性存在着较大的差异, 文献 [93−95] 中所采用的类哈尔 (Haar-like) 特征虽然能很好地反映人脸特征, 然而上述特征在描述文本区域时却表现差强人意. 文献 [88] 提出一种基于特定单词的自然场景文本检测方法. 该方法首先通过滑动检测窗口的方式获得单个的文本, 然后根据相邻文本之间的结构关系对可能的组合进行评分, 最后从给定的列表中选出最相近的组合作为输出结果. 区别于前述一般性的基于滑动检测窗的方法, 该方法只能检测事先给定列表中的单词, 对于列表之外的单词则无能为力. 然而, 在现实中不可能为每一幅图像指定一个包含所有可能出现的单词列表, 从而使得该方法的适用范围受到一定程度的限制.

基于滑动检测窗的自然场景文本检测方法的一个关键问题就是如何找到区分度好的描述特征来区分文本区域与背景区域. 传统的自然场景文本检测技术主要选择了手工设计的特征, 比如: 梯度边缘特征[99]、局部二值模式 (Local binary patterns, LBP)[100−101]、边缘局部二值模式 (Edge local binary patern, eLBP)[102]、方向梯度直方图 (Histograms of oriented gradients, HOG)[90, 103−104]、共生方向梯度直方图 (Co-occurrence histogram of oriented gradients, CoHOG)[105−106]、基于方向梯度直方图的纹理特征 (HOG-based texture descriptor, T-HOG)[107]、边缘方向梯度直方图 (Histogram of gradients at edges, eHOG)[20]、小波变换特征[17, 108−109]、离散小波变换特征[101] 等. 相应地, 为了分类文本区域与背景区域, 一些监督学习方法广泛地应用于自然场景文本检测领域, 比如支持向量机 (Support vector machine, SVM)[17, 101, 110−114] 、自适应增强算法 (Adaptive boosting, Adaboost)[98, 115−116]、随机森林 (Random forest, RF)[104, 117−119]、以及人工神经网络 (Artificial neural network, ANN)[120] 等. 大部分基于滑动检测窗的文本检测方法利用了文本候选区域的全局特征, 而文献 [121] 则从文本的局部特征出发, 提出了一种基于文本部件的树形结构模型 (Part-based tree-structured models), 该算法[121] 能较好地适应文本的字体变化, 对噪声、模糊等干扰因素也相对不敏感. 然而该模型依赖于详细的标注信息, 对不同语种文本的适应性也非常有限, 不能直接推广到新的语种文本. 若要处理新的语种文本, 则需要重新设计字符模板以及标注文本部件.

基于滑动检测窗的自然场景文本检测方法无需通过提取文本边缘、角点、连通区域或者文本行边缘等方式来获得文本/文本行候选区域, 该类方法在处理文本尺度较小或者对比度欠佳等情况具有较大的优势, 能有效地避免相邻文本间的粘连现象对文本候选区域提取的影响. 与此同时, 该类方法通常采用了区分性能好的手工设计的特征来区分文本区域与背景区域, 因此能较好地应对复杂自然场景中的文本检测问题. 考虑到自然场景图像中的文本区域通常由单个文本或者由多个文本构成, 除了文本位置随机分布以及相邻文本间隔距离多样化外, 文本大小尺寸以及文本区域的长宽比也存在着多个自由度, 此外, 文本行的排列方向通常比较随意, 有横行、竖行、斜行、甚至是弯曲的, 这对检测窗口的选取带来了很大的难度. 与此同时, 检测窗口的滑动步长的选取也是一个棘手的问题, 上述参数若设置不恰当将导致相当部分的文本漏检、欠分割、过分割以及出现虚警 (如图 7 所示), 从而影响文本检测性能. 基于滑动检测窗的自然场景文本检测方法采用多尺度滑动检测窗口的方式遍历整幅图像来获得文本候选区域. 为了有效地区分文本区域与背景区域, 一些复杂的手工设计的特征被大量使用, 从而增加了描述特征的计算复杂度, 进而导致了该类方法的检测效率通常不尽人意. 除此以外, 基于滑动检测窗的自然场景文本检测方法除了需要获得一个分

类性能好的描述特征外, 还对正、负训练样本的规模以及训练集的完备性提出了严格的要求. 不少算法的训练不仅需要知道每张训练样本中是否包含文本, 而且还需要知道每个文本所处的位置. 为了应对单文本与多文本情形, 还需要建立单文本训练样本、多文本训练样本. 与此同时, 为了获得良好的学习效果, 需要大量贴近真实场景的样本进行训练, 增加了标注工作量与训练时间. 尽管基于滑动检测窗口的方法在其他类型物体检测问题上取得了很好的结果, 但从 ICDAR 2011[122] 以及 ICDAR 2013[123] 的 "Robust Reading Competition Challenge 2" 的竞赛结果来看, 该类方法尚逊于基于连通区域分析的方法, 近几年基于滑动窗口的方法并没有成为文本检测算法的主流.

5.2　基于深度学习的自然场景文本检测方法

局限于手工设计的特征分类能力的不足, 文本检测性能在较长的一段时间内难以取得较大突破, 直至有了深度学习技术之后. 深度学习作为神经网络模型的新发展[124], 它模拟了人脑认识事物机理. 与传统的浅层人工神经网络相比, 深度学习含有多隐藏层的神经网络结构. 区别于传统的 "手工设计的特征提取 + 分类器" 的目标识别框架, 深度学习网络通过组合低层特征形成更加抽象的高层来表示属性类别, 使计算机自动学习数据的有效特征表示, 应用深度学习有一个很大的优势是可以避免繁琐低效的人工特征工程. 深度学习通过对训练样本进行学习以自动地获取描述特征[125] 的方式, 特别适合于物体识别与语音识别等模式识别问题. 典型的深度学习结构包括: 深度置信网络 (Deep belief network, DBN)[125−126]、卷积神经网络 (Convolutional neural network, CNN)[127] 以及递归神经网络 (Recurrent neural network, RNN)[128] 等.

深度学习 (如 CNN/LSTM 等模型) 在文本识别领域的应用有着较长的历史. 上世纪 90 年代, 深度学习的先驱者 Lecun 很早就用神经网络来解决文本识别, 1998 年, Lecun 等合作设计了 LeNet5 模型[127], 在 MNIST 数据集上的识别率高达 99.1 %; 在加上变形样本训练后, 其识别率进一步提升到 99.2 %. 2003 年微软研究院 Simard 等[129] 引入弹性变形 (Elastic distortion) 及仿射变形 (Affine distortion) 两种数据增广 (Data argumentation) 技术, 采用类似 CNN 的网络结构, 在 MNIST 数据集上将识别率提升至 99.6 %, 从而有效地解决了手写数字识别问题. 牛津大学 VGG 组的 Jaderberg 等[48] 较早地提出将深度学习方法运用于自然场景文本检测与识别领域, 他们在 2014 年利用深度卷积神经网络构建了如图 8 所示的自然场景文本识别框架. 在第一阶段, 学习一个不区分大小写的 CNN 文本分类器; 在第二阶段, 根据需要将结果特征映射应用于其他分类问题, 比如文本/背景分类器, 区分大小写的文本分类器以及二元分类器等.

目前已涌现出大量的基于深度学习的自然场景文本检测方法[12, 19, 31, 33, 37, 41, 45−49], 这些方法通过深度学习获得文本特征, 并依据上述特征对自然场景文本进行检测. 相比以前所使用的传统手工

图 7　基于自顶向下策略文本区域的错误提取结果[90]

Fig. 7　Error extraction result of text region based on top-down strategy[90]

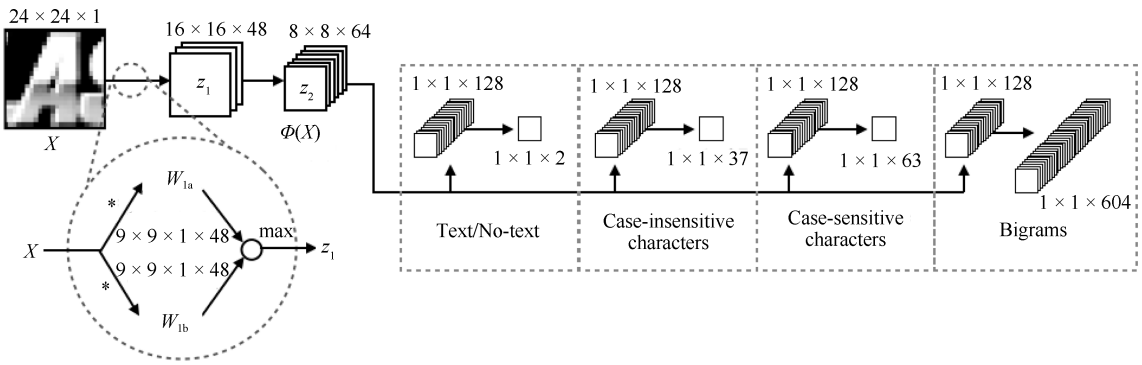


图 8　基于卷积神经网络的特征学习[48]

Fig. 8　Feature learning using a convolutional neural network[48]

设计的特征, 该类方法取得了更加令人鼓舞的检测结果. 从检测文本对象的排列方向这一角度来看, 基于深度学习的文本检测方法先后经历了水平方向排列的文本检测[31, 130]、任意方向排列的文本检测[15, 33, 50,131−134] 以及目前少数文献 [135−136] 所涉及的弧形排列方向的文本检测. 在基于深度学习的自然场景文本检测方法中, 基于文本区域建议 (Text region proposal) 的方法使用最为广泛, 其次是基于图像分割的方法. 主要的深度学习文本检测路线与一些代表性方法如图 9 所示.

5.2.1 基于区域建议的文本检测方法

基于区域建议的文本检测方法遵循一般目标检测的框架, 通常采用回归文本框的方式来获得文本区域信息. 文献 [119] 提出了分层文本检测策略, 该方法首先采用 CNN 提取特征, 从所获得的最大稳定极值区域中获得种子文本并依据种子文本来定位其他退化的文本区域, 然后采用随机森林结合文本行的上下文信息精细地分类文本候选区域. 文献 [130] 对 Faster RCNN 进行改进, 提出采用 Inception-RPN 方式获得文本候选区域, 然后利用一个文本检测网络去除背景区域, 最后对重叠的检测结果进行投票来获得最优的检测结果. 文献 [37] 首次将 RNN 引入到场景文本检测当中, 使用 CNN 得到深度特征, 然后用固定宽度的 Anchor 来检测文本建议区域 (Text proposal), 将同一行 Anchor 对应的特征输入到 RNN 中进行分类, 最后将正确的文本建议区域进行合并, 该方法得益于使用子块 (Block、Anchor) 对文本进行表示, 因此在一定程度上也能解决文本方向变化的问题. 文献 [12] 针对单词的分类问题, 将 CNN 与 RNN 进行联合训练, 首先, 采用标准 CNN 提取图像特征, 并利用 Map-to-sequence 表示成特征向量; 然后, 使用双向 LSTM (BLSTM) 学习场景文本的空间上下文信息; 最后, 对特征进行编码并得到最终的预测结果, 该方法将检测和识别模型结合

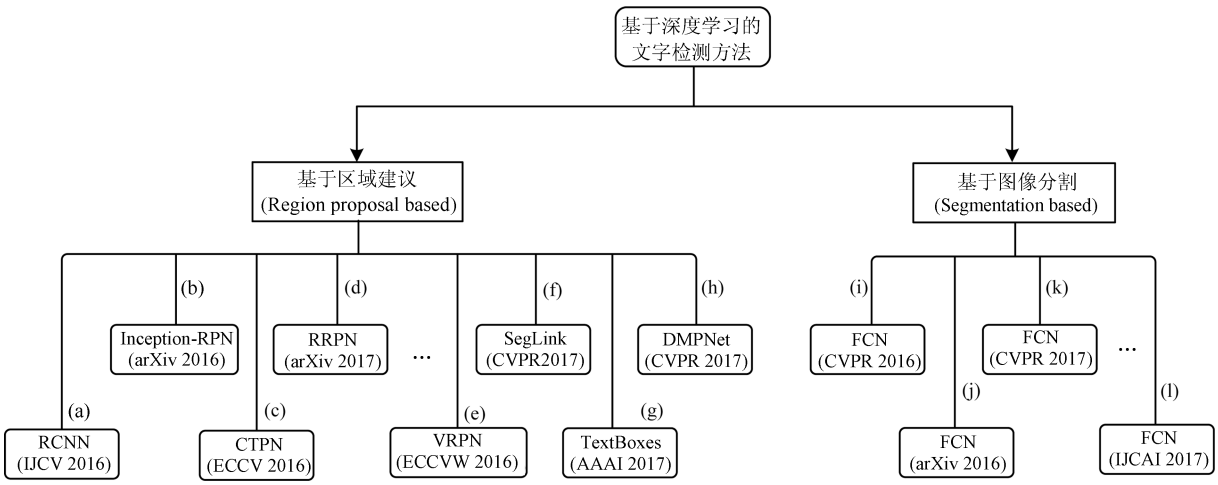


图 9 主要的深度学习文本检测路线与一些代表性方法 ((a) 文献 [137] 方法, 该方法采用 CNN 与 ACF 提取文本候选区域; (b) 文献 [130] 方法, 该方法对 faster RCNN 进行改进, 并提出 Inception-RPN 方式提取文本候选区域; (c) 文献 [37] 方法, 该方法提出了 Connectionist text proposal network 检测文本候选区域; (d) 文献 [138] 方法, 该方法提出旋转区域候选网络 (RRPN); (e) 文献 [139] 方法, 该方法提出了垂直回归建议网络 (VRPN); (f) 文献 [33] 方法, 该方法采用 Segment linking 方式解决多方向排列的文本检测问题; (g) 文献 [31] 方法, 该方法以 SSD 作为基础框架, 提出了一个端对端训练文本检测器 (TextBoxes); (h) 文献 [15] 方法, 该方法创新性提出采用四边形窗口 (非矩形) 的方式检测任意方向排列的文本; (i) 文献 [41] 方法, 该方法提出采用 Text-block 全卷积网络获得文本候选区域; (j) 文献 [140] 方法, 该方法采用 FCN 综合多信息属性来获得文本候选区域; (k) 文献 [50] 方法, 该方法参考了 DenseBox 的架构, 采用 FCN 网络检测任意方向排列的文本; (l) 文献 [141] 方法, 该方法采用深度卷积神经网络 (DCNN) 来学习文本的高级视觉表示 + 循环神经网络 (RNN) 处理文本序列.)

Fig. 9 The main deep learning text detection framework and some representative methods ((a) method[137], the CNN and the ACF are integrated to obtain the text region proposal. (b) method[130], the inception-RPN has been proposed in this work. (c) method[37], the connectionist text proposal network has been proposed in this work. (d) method[138], the RRPN has been proposed in this work. (e) method[139], the VRPN has been proposed in this work. (f) method[33], the segment and linking has been proposed in this work. (g) method[31], the TextBoxes method has been proposed in this work. (h) method[15], the deep matching prior network (DMPNet) with tighter quadrangle has been proposed in this work. (i) method[41], the text-block FCN has been proposed in this work. (j) method[140], the FCN and multi-channel prediction method has been proposed in this work. (k) method[50], the DenseBox framework has been followed and the FCN has been proposed in this work. (l) method[141], the DCNN and the RNN has been adopted in this work.)

之后得到了目前端到端模型中最好的文本检测结果. 文献 [33] 通过加入方向信息使得 SSD (Single shot detector) 检测器可以应对任意方向排列的文本检测问题. 该方法多尺度预测文本片段及其连接关系, 将文本信息转换为两个局部可检测信息, 即: 文本级或者单词级的 Segments 以及 Segments 之间的 Links. 其创新之处在于把这些 Links 加入到网络中去学习, 从而使得网络自动学习出哪些 Segments 属于同一个文本行 (或者单词). 文献 [137] 较早地开展了端到端 (End-to-end) 场景文本识别研究, 该方法针对文本检测问题对 R-CNN 进行了改造, 其工作内容主要分为二个部分: 基于目标候选区域 (Region proposal) 的文本检测部分以及基于卷积神经网络 (CNN) 的文本识别部分. 该方法获得了很好的场景文本识别效果, 并且在其后两年内一直保持领先地位. 文献 [79] 除了提出对比度增强的最大稳定极值方法 (Contrast-enhancement maximally stable extremal regions, CE-MSERs) 来提高文本检测召回率外, 还提出了基于多任务学习的文本注意卷积神经网络 (Text-attentional convolutional neural network, text-CNN) 模型, 该方法将底层像素级分割、高层的文本识别以及文本与背景分类融合到一个 text-CNN 模型中, 从而获得了较强的文本检测器. 传统的文本检测系统通常包含了多个处理流程, 各处理环节的性能均将直接影响到最终的检测结果, 文献 [15] 提出了一种深度匹配先验网络 (Deep matching prior network, DMPNet), 该方法考虑到原来的方法都专注于用矩形框来对文本进行定位, 然而实际上自然场景图像中的文本图像可能存在透视变换等问题, 从而导致图像中的文本区域并不是严格地呈现为矩形, 若继续采用矩形框来定位将出现错误的结果, 文献 [15] 创新性地提出采用四边形窗口 (非矩形) 来表示文本区域. 文献 [31] 提出了 Textboxes 文本检测方法, 该方法对 SSD 框架进行了改进, 其目的能实现快速地计算文本在每个区域存在的可能性, 文献 [31] 发现长条形的卷积核比常用的 1×1 或 3×3 卷积核更适合自然场景文本检测, 该方法在设计默认框 (Default box) 时考虑了包含较长的形状, 提出了一个实用的 "检测 + 识别" 框架对文本候选区域进行判断. 其后, Liao 等在文献 [131] 中对他们的前期工作[31] 进行了改进, 提出了 Textboxes++ 文本检测方法, 与前期工作 Textboxes 方法相比, 文献 [131] 除进一步修改网络结构以外, 其主要贡献是将 Textboxes 水平排列文本检测器扩展为任意方向排列文本检测器. 文献 [139] 提出了垂直回归建议网络 (VRPN). 为了生成具有文本方向角信息的倾斜候选框用于检测任意方向文本区域, 文献 [138] 提出了旋转区域候选网络 (Rotation region proposal networks, RRPN). 考虑到传统的 RoI 池化层只能处理轴对齐的候选框, 该文献还提出了旋转 RoI (RRoI) 池化层来调整 RRPN 生成的面向任意的候选框. 文献 [134] 为了检测任意方向的文本, 在 R-CNN[142] 构架的基础上提出了一种新的旋转区域 CNN (R^2CNN) 方法. 该文献使用 RPN 来生成轴对齐的包围不同方向的文本边界框, 通过合并 RPN 生成的不同大小的每个轴对齐文本框的特征来分类文本与非文本区域; 文献 [143] 针对端对端文本识别问题提出了一个统一的网络结构模型, 该模型主要包含了一个文本建议网络 (Text proposal network, TPN) 以及递归神经网络 (Recurrent neural network, RNN), 该模型可以直接通过一次前向计算就可以同时实现文本检测和文本识别任务. 对该网络模型进行训练时, 只需要输入图像、图像中文本的 Bbox 以及文本对应的标签信息. 与此同时, 文献 [143] 无需实施诸如文本行形成、单词分割等中间处理步骤, 从而可以减少错误. 文献 [133] 从实例感知语义分割 (Instance-aware semantic segmentation) 的角度提出了一种端对端训练框架 (Fused text segmentation networks, FTSN) 以应对多方向场景文本检测问题, 该方法采用 Resnet-101 backbone 提取特征后利用区域建议网络 (Region proposal network, RPN) 同时检测与分割文本实例, 通过非最大抑制方法 (Non-maximum suppression, NMS) 解决文本实例重叠的问题, 最后生成适合每个文本实例区域的最小四边形边界框作为整终的检测结果. 文献 [144] 为了应对任意方向的文本检测问题, 创新性地设计 RoIRotate 算法将任意方向特征转换为轴对齐特征. 近年来少数研究者基于深度学习方法对弧形排列方向的文本检测问题进行了研究. 文献 [136] 提出了滑动线点回归 (Sliding line point regression, SLPR) 方法检测任意方向排列的文本 (包括弧形排列方向文本), 该方法首先采用区域建议网络 (Region proposal network, RPN) 生成包含文本的最小矩形框, 然后分别沿着垂直方向和水平方向等距滑动线并回归文本的边缘点, 最后基于这些点获得文本的轮廓. 文献 [135] 提出了一种基于多边形的弧形文本检测算法 (Curve text detector, CTD), 此外该方法还提出了两个简单有效的后处理方法, 即: 非多边形抑制 (NPS) 和多边形非最大抑制 (PNMS), 以进一步提高文本检测精度. 除此以外, 文献 [135] 还推出了主要包含弧度方向排列文本的数据集 (SCUT-CTW1500), 该数据集共包含了 1 500 张图片, 其中 1 000 张图像作为训练集, 500 张图片作为测试集, 累积标注了约 10 000 个文本区域.

5.2.2 基于图像分割的文本检测方法

基于图像分割的文本检测方法[41, 50, 132, 140−141]将文本检测视为一种广义的“分割问题”. 该类方法通常利用语义分割中常用的全卷积网络 (FCN) 等方式来进行像素级别的文本/背景标注. 文献 [41] 首次采用了全卷积网络 (Fully convolutional network, FCN) 从像素层面对图像进行处理, 该方法首先利用 Text-block FCN 进行像素级的标定, 从而获得每个像素属于文本的概率, 进而得到文本区域显著图 (Salient map), 最后基于显著图得到文本候选区域 (如图 10 所示). 文献 [145] 提出了一种级联卷积文本网络 (Cascaded convolutional text network, CCTN), 该方法采用级联的方式检测文本, 具体处理步骤主要包括: 首先, 采用一个 Coarse-CNN 进行检测得到粗略的文本区域, 然后, 对所获得的文本区域检测结果进行判断是否需要进一步处理 (Refine), 若需要, 则采用 Fine-CNN 进行处理以得到更细致的文本线进行输出. 文献 [50] 提出了一种基于全卷积神经网络 (FCN) 与非最大抑制算法 (Non-maximum suppression, NMS) 的简单高效的文本检测框架, 该方法首先通过全卷积神经网络输出文本区域像素级检测结果, 然后将上述结果通过非最大抑制算法获得文本区域. 文献 [132] 提出基于深度直接回归的多方向场景文本检测方法, 该文献在其所提出的检测框架中对全卷积神经网络进行了端对端的优化并双任务输出, 其中一个任务是对文本与非文本进行像素级分类, 另一个任务则是采用该文献所提出的新贡献 直接回归的方式以确定四边形文本边界的顶点坐标. 文献 [141] 中先采用深度卷积神经网络 (DCNN) 来学习文本的高级视觉表示, 然后用循环神经网络 (RNN) 处理不规则文本 (Irregular text) 序列. 为了获得文本候选区域, 文献 [141] 采用了 FCN 网络来完成密集的文本检测任务. 文献 [140] 方法基于全卷积网络, 把“预测文本区域概率”、“预测字符概率”、“预测相邻字符连接概率”三个问题整合到一个网络中去进行整体学习以获得文本候选区域.

绝大部分基于深度学习的文本检测方法主要包含了两个部分内容, 即文本/非文本分类处理以及文本边界框回归处理, 尽管文本边界框回归处理并不是必须的处理步骤, 然而它对最终的检测结果产生重要影响. 区别于绝大部分基于深度学习的文本检测方法, 文献 [146] 直接通过实例分割处理来获得文本位置信息而无需进行文本边界框回归处理. 受到 SegLink[33] 方法的启发, 文献 [146] 所提出的 PixelLink 方法采用了一个深度神经网络 (Deep neural network, DNN) 来实现二种像素级预测, 即文本/非文本预测以及连接预测, 将所有文本实例中的像素进行标注并形成连通区域, 最后从分割结果中直接提取文本边界框. 文献 [147] 为了降低文本排列方向以及文本区域长宽比变化的影响, 该方法首先检测文本角点, 然后通过对角点进行采样和分组得到文

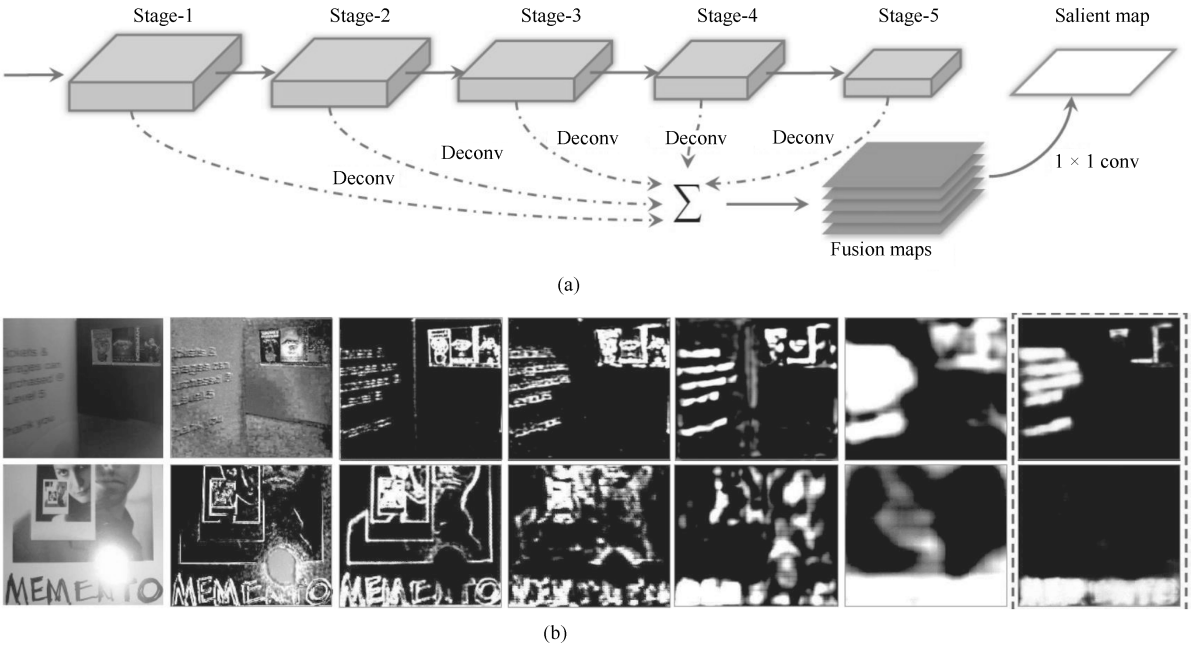


图 10 基于全卷积神经网络的自然场景文本检测[41] ((a) Text-Block 全卷积神经网络结构; (b) Text-Block 全卷积神经网络获得的结果)

Fig. 10 Natural scenes text detection based on fully convolutional networks[41] ((a) The network architecture of the Text-Block FCN, (b) The illustration of feature maps generated by the Text-Block FCN)

本候选区域的边框, 最后基于分割信息对所获得的文本候选区域的边框进行评价, 并采用非极大抑制处理 (NMS) 来获得最终的检测结果. 基于文本区域建议的文本检测方法[15, 31, 131, 134] 通常遵循一般目标检测的框架, 采用回归文本框的方式来获得文本区域的位置信息, 该类方法受到文本排列方向的任意性以及文本区域长宽比多样性的困扰. 基于图像分割的文本检测方法[41, 50, 132, 140−141] 从另外的视角出发, 视文本检测为一种广义的 "分割问题", 可以较好地避免文本排列方向以及文本区域长宽比变化的影响, 然而该类方法的后续处理通常比较的复杂. 此外, 由于目前绝大部分文本检测数据集的标注都是文本框类型, 仅仅将文本标记在某一矩形区域内而没有详细地标注出哪些像素点是文本哪些是背景, 因此基于图像分割的文本检测方法还面临着像素级别图像标注的困难. 考虑到人工标注像素 (Pixel) 级别的标记 (Label) 代价很高, 采用人工合成数据是一个值得尝试的替代手段.

为了获得较现有方法更优的检测结果, 选择或者设计更有效的深度学习文本检测框架显得格外重要. 仍然需要指出的是, 尽管深度学习方法在基于大量训练样本的情况下获得了较传统的手工设计的特征更优的区分性能, 但是自然场景文本检测系统通常包含了多个处理环节, 任何环节的处理结果都将会影响整个系统的检测性能. 深度学习方法虽然能很好地解决文本分类这一局部问题, 然而较难有效地利用文本的上下文信息以及其他知识. 虽然简单直接地应用深度学习技术可以达到还不错的检测结果, 但依然有必要将深度学习方法与其他的领域知识或者技巧相结合来设计文本检测系统. 此外, 采用深度学习方法进行训练时, 训练集的规模将对训练结果产生重要影响. 训练样本规模小将容易导致训练过拟合, 训练样本规模大则使得构建训练集及手工标注的工作量过大. 为了构建大规模训练集, 文献 [137, 141, 148−149] 等提出通过合成的方法生成含有文本信息的样本, 进而为扩充训练集的规模提供了一种有效的解决途径. 牛津大学 VGG 组的 Jaderberg 等除了在文献 [149] 中提出采用合成图 (Synthetic image) 的方式训练卷积神经网络 (CNN) 外, 他们还在文献 [148] 中详细地介绍了如何通过合成的方法生成自然场景文本样本, 文献 [148] 通过人工生成自然场景文本样本在 ICDAR 2011 数据集上获得了 F-measure 为 82.3 % 的成绩. 此外, 文献 [150] 考虑到现有的真实文本数据集大多是在单词或文本行级别进行标注的, 因此该文献提出了一个弱监督的框架, 基于单词级训练数据库来训练文本检测器以解决文本训练数据集不足的难题.

6 端对端文本识别方法

区别于单独的文本检测与单独的文本识别任务, 端到端文本识别包含了从自然场景图像中检测和识别文本的完整过程 (如图 11 所示). 在端到端文本识别任务中, 输入的是自然场景图像, 输出结果为图像中的文本内容. 从本质上来说, 文本检测和文本识别同属于模式分类问题. 文本检测的核心任务是区分图像中的文本和非文本成分, 因此文本检测是一个粗略的二分类问题; 而文本识别则需要在文本检测结果中进一步区分文本的所属类别, 因此文本识别则需要完成更精细的分类任务. 从针对自然场景文本检测与识别的研究内容来看, 目前大部分工作将文本检测与文本识别作为两个独立的内容来展开研究, 只有少数工作将文本检测与文本识别融合到一个框架中执行粗糙检测与精细化分类的两个任务, 从而达到同时进行文本检测和文本识别的目的. 相比单纯的文本检测与文本识别问题, 端对端文本识别更加具有挑战性. 从 ICDAR 2015 自然场景文本检测及识别竞赛[32] 的结果来看: 非受限环境下的自然场景文本 (Incidental scene text) 在无语料信息的真实环境下的端到端识别任务 (Task 4.4) 的最好识别率仅为 34.96 %, 可见端对端文本识别技术具有很大的提升空间.

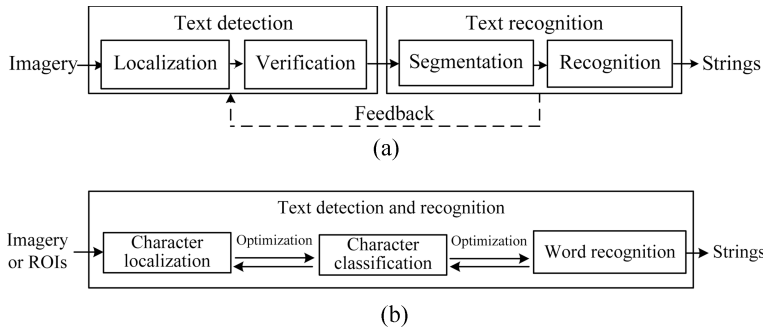


图 11 端到端场景文本识别框架[22]

Fig. 11 Scene text end to end recognition framework[22]

文献 [34, 73, 151−152] 较早地提出了端对端文本识别系统, 但这些系统主要关注文本检测部分, 文本的识别则依赖于已有的光学字符识别引擎. 文献 [153] 所提出的端对端文本识别系统则主要关注文本识别部分, 其文本检测部分利用了文献 [93, 154] 所提出的目标检测方法. 与文献 [34, 73, 151] 方法不同, Wang 等[91] 和 Neumann 等[60] 在他们所提出的端到端文本识别系统中并没有采用已有的光学字符识别软件, 而是自主设计了自然场景文本识别方法. Wang 等[91] 将单词作为一种特殊的待检测目标, 视字符为单词的组成部件, 通过各个字符的置信度以及字符之间的空间约束关系搜索最可能的检测和识别结果. 捷克理工大学 Jiri Matas 以及 Neumann 等[60] 通过提取图像中的最大稳定极值区域作为文本候选区域, 然后通过训练好的分类器剔除非文本区域, 将余下的候选区域输入到字符识别模型中进行识别. Neumann 等在前期工作[60] 的基础上, 在文献 [59] 中提出一个实时的端到端场景文本检测和识别方法, 其文本检测部分基于一种高效的序贯选择机制, 从一个极值区域集合中挑选可能的文本区域, 文本识别模型则由合成训练样本得到. 需要指出的是, 文献 [59] 是第一个在 ICDAR 2011 数据集上发布端对端文本识别结果报告的, 该文献所述方法现已被 OpenCV 3.0 所采用. 文献 [34, 59−60, 91, 151] 只能处理水平方向或者接近水平方向排列的自然场景文本. 考虑到上述方法的不足, 华中科技大学研究团队 Yao 等在文献 [76] 中率先提出了一种可以处理自然场景中任意方向文本的端到端识别方法. 该方法将文本检测和文本识别作为一个整体进行考虑, 在统一的框架中利用相同的特征和分类结构同时完成检测和识别任务, 此外该方法设计了一种基于字典搜索的纠错策略来提高文本识别准确性.

由于传统手工设计的特征不能有效地区分文本区域, 从而导致端对端文本识别性能在较长的一段时间里难以取得突破, 直至 2014 年前后深度学习方法为端对端文本识别问题提供了全新的解决方案. 在文献 [48, 137, 143, 148, 155−156] 等中设计了各种基于深度学习的端对端文本识别框架. 牛津大学 VGG 组在 2016 年 IJCV 期刊的首卷首期发表了基于区域建议 (Region proposal) 的方法[137], 该方法在端到端文本识别领域保持了近两年的领先地位. 文献 [137] 从两个方面展开对端到端 (End-to-end) 场景文本识别的研究 (如图 12 所示), 即: 基于目标区域建议 (Region proposal) 的文本检测部分以及基于卷积神经网络的文本识别部分. Shi 等在文献 [12] 中针对图像中的序列物体的识别问题提出了 Convolutional recurrent neural network (CRNN) 端对端检测框架. 针对单词的分类问题, 该方法首先采用标准 CNN 提取图像特征并利用 Map-to-sequence 表示成特征向量, 然后使用双向 LSTM (BLSTM) 学习场景文本的空间上下文信息, 最后对特征进行编码并得到最终的预测结果, 该方法得到了目前端到端模型中最好的文本检测结果. Alsharif 等[157] 采用了一种包含分割、矫正以及文本识别的 CNN 网络, 结合使用固定词典的隐马尔科夫模型 (HMM) 来获得最终的识别结果. Liao 等在文献 [31] 中对 SSD 框架进行了改进, 针对水平方向排列的文本提出了一种 “Textboxes”+“CRNN” 的端对端识别框架, 其中 Textboxes 用来实现文本检测, CRNN 则用来进行文本识别; 最近, Liao 等在文献 [131] 中对其前期工作[31] 进行了改进, 提出了一种 “Textboxes++”+“CRNN” 的端对端的文本识别框架, 文献 [131] 的主要贡献是将其前期工作

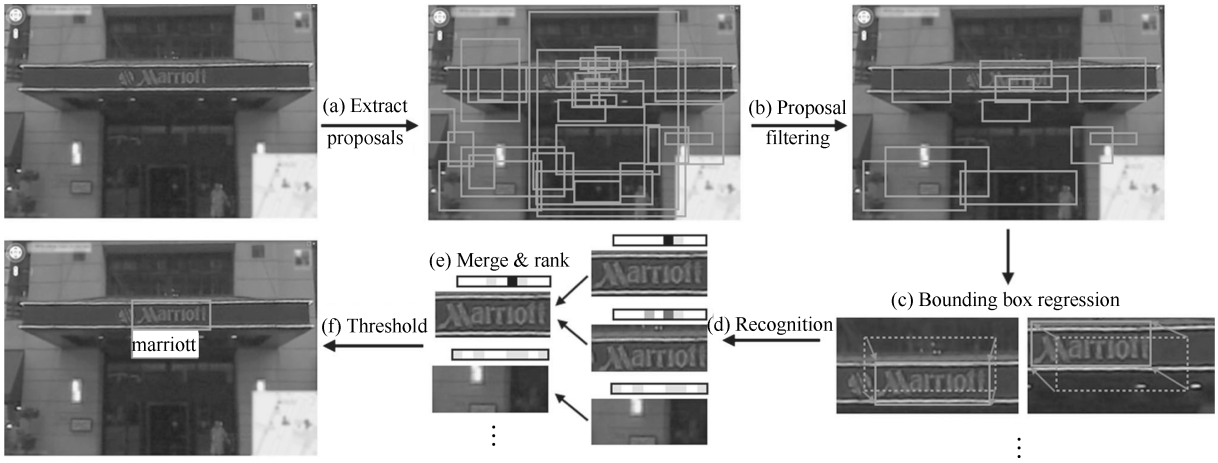


图 12 基于卷积神经网络的端对端自然场景文本识别方法[137]

Fig. 12 Feature learning using a convolutional neural network[137]

Textboxes[31] 进行了扩展, 在其所提方法中设计了 Textboxes++ 文本检测方法以应对任意方向排列的文本的检测问题. 文献 [158] 借鉴人类阅读文本的认知机制, 提出了一种基于卷积特征映射的端对端场景文本识别方法. 该方法首先采用滑动检测窗口的方式对输入图像进行扫描, 并将输入图像按照检测窗口大小切割出来, 然后获得所有切割后图片的特征向量, 由时序分类算法 (Connectionist temporal classification, CTC) 预测得到最终的识别结果. 该方法表现出一些优势, 比如避免了复杂的字符分割过程以及可以识别基于单词方法所不能识别的文本. 文献 [144] 提出了一种任意方向排列文本识别方法 (Fast oriented text spotting system, FOTS), 该方法考虑到提取特征是一个较耗时的过程, 为了加快系统的处理速度, 该方法采用共享特征的方式同时实现文本检测与文本识别处理. 为了应对文本排列方向的任意性, 文献 [144] 设计了 RoIRotate 算法将任意方向特征转换为轴对齐特征. 在文献 [148] 所提出的端对端文本识别方法中, 首先训练了一个全卷积回归网络 (Fully convolutional regression network, FCRN) 以及文本位置回归的方式检测文本, 然后通过文献 [149] 所采用的单词分类器 (Word classifier) 进行文本识别. Li 等在文献 [143] 中提出了一种端对端文本识别网络结构模型, 该模型主要包含了一个文本建议网络 (Text proposal network, TPN) 以及递归神经网络 (Recurrent neural network, RNN), 该模型可以直接通过一次前向计算就可以同时实现文本检测和文本识别任务. Patel 等在文献 [159] 中提出了 E2E-MLT 多语言场景文本识别算法, 该方法集成了多个卷积神经网络, 有效地实现了自然场景文本检测、识别以及文种分类等任务. 特别需要指出的是, 文献 [159] 所提的 E2E-MLT 模型是迄今为止第一个面向多种语言的文本识别方法. 与大部分传统的监督训练方式不同, Bartz 等在文献 [155, 160] 中对其提出的端对端文本检测与识别系统采用了半监督学习方法进行了训练. 为了应对任意方向排列的自然场景文本, 在文献 [156] 所提出的端对端文本识别系统中, 首先采用了旋转文本建议网络 (Region proposal network, RPN) 来获得文本区域, 然后采用基于合成文本样本训练后得到的文本分类器进行识别. 对于单个文本的识别问题, 基于深度学习文本识别方法的做法通常与传统方法类似, 采用 CNN 获取文本的描述特征并进行分类[46]; 对于由多个文本构成的单词, 主要采用了 CNN+LSTM 结构[12, 16], 首先利用 CNN 学习图像相邻像素之间的关系, 然后利用长短期记忆神经网络 (Long short-term memory, LSTM) 学习较长跨度的上下文关系.

7 性能评估

7.1 测试数据集

随着自然场景文本检测这一研究领域的不断发展, 越来越多的文本数据集被推出以供研究人员来检验其方法的性能. 最为常见且使用最为广泛的数据集有 ICDAR 自然场景文本检测竞赛的系列数据集. 除此以外, 自然场景文本检测数据集还包括了 MSRA-TD500、SVT、COCO-Text 等. 上述各种数据集的特点如表 1 所示.

表 1 常用自然场景文本检测数据集

Table 1 Widely used natural scene text detection datasets and their download link

数据集	年份	数据集大小	图像数目 (训练/测试)	文本数目 (训练/测试)	文本种类	文本排列方向
ICDAR′03[161]	2003	120.2 MB	509 (258/251)	2 276 (1 110/1 156)	英文	水平方向
ICDAR′11[30]	2011	266 MB	484 (229/255)	2 037 (848/1 189)	英文	水平方向
ICDAR′13[123]	2013	250 MB	462 (229/233)	1 943 (848/1 095)	英文	水平方向
ICDAR′15[32]	2015	131.8 MB	1 500 (1 000/500)	17 548	英文	水平方向
SVT[88]	2010	112 MB	350 (100/250)	904 (257/647)	英文	水平方向
MSRA-TD500[39]	2012	96 MB	500 (300/200)	1 719 (1 068/651)	中文/英文	任意方向
KIST[162]	2010	347.4 MB	3 000	> 5 000	英文/韩文	水平方向
OSTD[21]	2011	17.34 MB	89	218	英文	任意方向
NEOCR[163]	2011	1.3 GB	659	5 238	英文	任意方向
USTB-SV1K[164]	2015	36.1 MB	1 000 (500/500)	2 955	英文	任意方向
COCO-Text[58, 165]	2016	–	63 686	173 589	多语种	任意方向
RCTW-17[51]	2017	5.4 GB	> 12 000 (8 034/4 000)	–	中文	任意方向
SCUT-CTW1500[135]	2017	842 MB	1 500 (1 000/500)	10 000	英文	含弧形排列的任意方向

上述数据集的下载地址分别为: ICDAR′03[3], ICDAR′11[4], ICDAR′13[5], ICDAR′15[6], SVT[7], MSRA-TD500[8], KIST[9], OSTD[10], NEOCR[11], USTB-SV1K[12], COCO-Text[13], RCTW-17[14], SCUT-CTW1500[15].

7.2 评估方法

为了客观地评测各种方法的检测性能, 目前已推出了几种测评方法 (后续小节中将对各种评测方法进行详细介绍). 现有检测性能评测方法主要考虑三个性能参数, 即: 准确率 (Precision, P)、召回率 (Recall, R)、综合评价指标 (F-measure, F). 准确率 (P) 表示检测得到的真实文本与所有检测结果之间的比率, 召回率 (R) 表示检测得到的真实文本和所有手工标注的真实文本之间的比值, 综合评价指标 (F) 是准确率与召回率的调和平均值, 该值是评价文本检测方法性能的综合指标.

7.2.1 ICDAR 2003/2005 评估方法

具体办法是通过将检测结果的最小外接矩形与手工标注的文本区域矩形进行比较以获得其公共部分面积, 并通过公共部分面积计算出文本检测召回率、精确率以及综合评价指标, 最后根据上述三个性能指标对检测结果的优劣性进行评价. 匹配度 m_p 定义为上述两个矩形之间的公共部分面积与包含上述两个矩形的最小外接矩形的面积之比. 当两个矩形完全重合时, 匹配度 $m_p = 1$; 当两个矩形之间无公共部分时, 则匹配度 $m_p = 0$.

一个矩形 r 与一组矩形 R_e 之间的最佳匹配度采用式 1 进行定义.

$$m(r; R_e) = \max\{m_p(r; r_0) | r_0 \in R_e\} \tag{1}$$

召回率 (Recall, R) 和准确度 (Precision, P) 分别采用式 (2) 和式 (3) 进行定义.

$$R = \frac{\sum\limits_{r_t \in T} m(r_t, E)}{|T|} \tag{2}$$

$$P = \frac{\sum\limits_{r_e \in E} m(r_e, T)}{|E|} \tag{3}$$

其中, T 表示手工标注文本区域的矩形集合, E 表示检测结果的矩形集合. 综合评价指标 (F-measure) 为召回率 (Recall) 和精确率 (Precision) 的调和平均值, 其定义如式 (4) 所示.

$$F = \frac{1}{\dfrac{a}{P} + \dfrac{1-a}{R}} \tag{4}$$

其中, 参数 α 为检测召回率与准确率之间的权重, 通常取值为: $\alpha = 0.5$.

7.2.2 ICDAR 2011/2013 评估方法

ICDAR 2003 以及 ICDAR 2005 评估方法没有考虑检测结果与手工标注结果 (Ground-truth) 之间一对多 (One-to-many) 与多对一 (Many-to-one) 的匹配情形. 事实上, 检测结果与 Ground-truth 之间一对多 (One-to-many) 与多对一 (Many-to-one) 的匹配情形在实际检测结果中并不少见, 因此在采用 ICDAR 2003 以及 ICDAR 2005 评估方法时容易低估自然场景文本检测方法的实际性能. 考虑到上述情况, ICDAR 2011 以及 ICDAR 2013 自然场景文本检测竞赛采用了文献 [166] 所提出的评估方法. 需要指出的是, 文献 [166] 认为多对多 (Many-to-many) 的匹配情形并不常见, 因此在文献 [166] 中暂未考虑多对多匹配情形.

文献 [166] 所提出的评估方法主要考虑了检测结果与 Ground-truth 之间的三种匹配情形, 即: 一对一 (One-to-one)、一对多 (One-to-many) 以及多对一 (Many-to-one)(如图 13 所示). 准确率 (P) 与召回率 (R) 分别定义为

$$P(G, D, t_r, t_p) = \frac{\sum\limits_{j} Match_D(D_j, G, t_r, t_p)}{|D|} \tag{5}$$

$$R(G, D, t_r, t_p) = \frac{\sum\limits_{i} Match_G(G_i, D, t_r, t_p)}{|G|} \tag{6}$$

其中, G, D 分别表示 Ground-truth 与检测结果, $t_r \in [0, 1]$ 是召回率约束项, $t_p \in [0, 1]$ 是精确率约束项, 其取值分别为 $t_r = 0.8, t_p = 0.4$. 函数 $Match_D$ 与 $Match_P$ 用来区分匹配类型, 具体来说可以表示为

[3] Available at: http://algoval.essex.ac.uk/icdar/Datasets.html
[4] Available at: http://robustreading.opendfki.de/
[5] Available at: http://dag.cvc.uab.es/icdar2013competition
[6] Available at: http: //www.iapr.org/archives/icdar2015/index.html%3Fp=254.html
[7] Available at: http://vision.ucsd.edu/~kai/grocr/
[8] Available at: http://pages.ucsd.edu/~ztu/Download_front.htm
[9] Available at: http://www.iapr-tc11.org/mediawiki/index.php/KAIST_Scene_Text_Database
[10] Available at: http://media-lab.ccny.cuny.edu/ wordpress/cyi/www/ project_scenetextdetection.html
[11] Available at: http://www.iapr-tc11.org/mediawiki/index.php?title=NEOCR:Natural_Environment_OCR_Dataset
[12] Available at: http://prir.ustb.edu.cn/TexStar/MOMV-text-detection/
[13] Available at: https://vision.cornell.edu/se3/coco-text-2/
[14] Available at: http://mclab.eic.hust.edu.cn/icdar2017chinese/dataset.html
[15] Available at: https://github.com/Yuliang-Liu/Curve-Text-Detector

$$Match_D(D_j, G, t_r, t_p) = \begin{cases} 1, & \text{若 } D_j \text{ 匹配一个真值标注区域} \\ 0, & \text{若 } D_j \text{ 不匹配任何真值标注区域} \\ f_{sc}(k), & \text{若 } D_j \text{ 匹配多个真值标注区域} \end{cases} \tag{7}$$

$$Match_G(G_i, D, t_r, t_p) = \begin{cases} 1, & \text{若 } G_i \text{ 匹配一个检测结果} \\ 0, & \text{若 } G_i \text{ 不匹配任何检测结果} \\ f_{sc}(k), & \text{若 } G_i \text{ 匹配多个检测结果} \end{cases} \tag{8}$$

其中, $f_{sc}(k)$ 为针对欠分割与过分割情况的惩罚函数, 文献 [166] 取 $f_{sc}(k) = 0.8$.

7.2.3 ICDAR 2015 评估方法

ICDAR 2015 自然场景文本检测竞赛采用了文献 [167] 所提出的目标检测评价方法, 通过比较检测结果矩形框与 Ground-truth 矩形框之间的公共区域与并集区域之间的比值来进行衡量. 具体做法是, 定义覆盖面积比值为

$$a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{9}$$

其中, a_0 为覆盖率, B_p 和 B_{gt} 分别表示检测结果与 Ground-truth, $area(B_p \cap B_{gt})$ 与 $area(B_p \cup B_{gt})$ 分别表示 B_{gt}, B_p 之间的交集区域与并集区域. 若检测结果与 Ground-truth 之间的实际面积覆盖率 $a_0 > 0.5$, 则认为该检测结果为正确的; 反之, 则认为是虚警. 当同一文本行出现多个检测结果时, 根据降序顺序将除了第一个检测结果以外的其余检测结果视为虚警. 准确率 (P) 与召回率 (R) 分别定义为

$$P = \frac{|T_P|}{|E|} \tag{10}$$

$$R = \frac{|T_P|}{|T|} \tag{11}$$

其中 T_P, E, T 分别表示正确的检测结果集合, 检测结果集合以及 Ground-truth 集合, 综合评价指标 (F) 则定义为

$$F = \frac{2PR}{P + R} \tag{12}$$

7.2.4 MSRA-TD500 评估方法

文献 [39] 针对任意方向自然场景文本检测提出了一种有效的评估方法. 该方法采用了文献 [168] 所提出的最小面积矩形框对文本区域进行了标记, 图 14 (a) 为手工标记的结果. 对于任意方向排列的文本而言, 采用文献 [168] 所提出的最小面积矩形框相较轴对称矩形框更加紧致 (如图 14 (b) 所示). 采用图 14 (c) 所示方式计算检测结果与 Ground-truth 之间的覆盖率, 其中 G, D 分别表示为 Ground-truth 与检测结果. 考虑到在计算 G, D 之间的覆盖率时不够方便, 文献 [39] 将 G, D 按照其中心点 C_G, C_D 进行旋转至 G', D' 所示位置 (如图 14 (c) 所示). G, D 之间的覆盖率定义为

$$m(G, D) = \frac{A(G' \cap D')}{A(G' \cup D')} \tag{13}$$

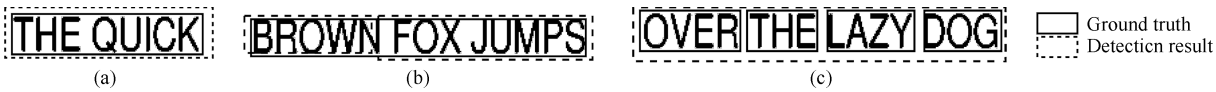


图 13　检测结果与 Ground-truth 匹配模式[166]

Fig. 13　Matching model of the detection results and ground-truth[166]

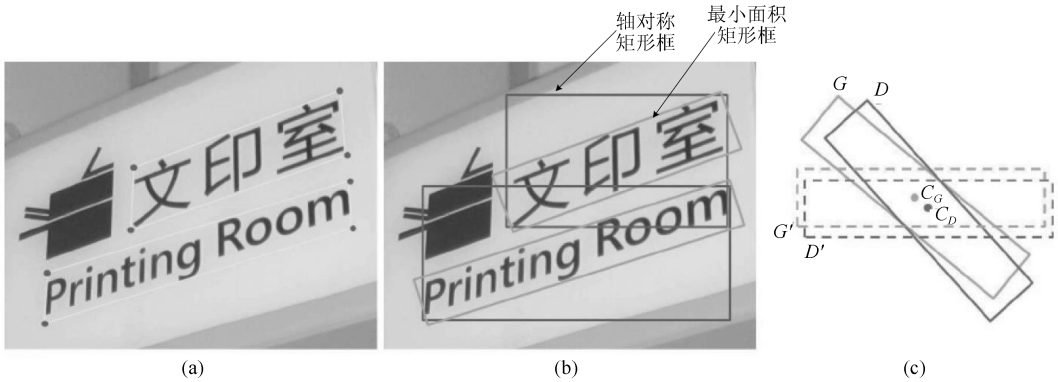


图 14　MSRA-TD500 数据集评估方法[39]

Fig. 14　Evaluation method of the MSRA-TD500 datasets[39]

$A(G' \cap D')$ 与 $A(G' \cup D')$ 分别表示 G', D' 之间的交集与并集区域, 文献 [39] 借鉴了 PASCAL 目标检测性能评估方法[169], 其具体做法是: 当 G, D 之间的倾斜角之差小于 $\pi/8$ 且覆盖率 m 大于 0.5 时, 则认为 D 为一个正确的检测结果. 对于同一文本行而言, 若出现多个检测结果, 则根据降序顺序将除了第一个检测结果以外的其余检测结果视为虚警. 准确率 (P) 与召回率 (R) 分别定义为

$$P = \frac{|T_P|}{|E|} \tag{14}$$

$$R = \frac{|T_P|}{|T|} \tag{15}$$

其中, T_P, E, T 分别表示正确的检测结果集合, 检测结果集合以及 Ground-truth 集合, 综合评价指标 (F) 定义为

$$F = \frac{2PR}{P+R} \tag{16}$$

7.3 测试结果

各种自然场景文本检测方法在各类公开数据集上进行测试, 一些代表性的文本检测方法的测评结果如表 2 所示[170−173]. 由表 2 可知, 自然场景文本检测技术在近几年得到了长足发展, 比如在 ICDAR 2011 数据库上的综合标价指标从 0.71 上升到 0.85 (2011 ∼ 2017 年), 在 ICDAR 2015 数据库上的综合标价指标从 0.50 上升到 0.81 (2015 ∼ 2017 年).

8 自然场景文本检测方法存在的问题

近年来, 尽管不少行之有效的文本检测方法被提出, 文本检测的性能也获得了大幅度提升, 但自然场景文本检测技术依然存在着一些不足.

8.1 端到端 (End-to-end) 场景文本识别问题

端到端场景文本识别需要同时应对文本检测与文本识别双重任务 (如图 11 所示), 目前针对端对端自然场景文本识别的研究还相对的单薄; 从已发表文献所采用的技术手段来看, 大部分工作[34, 49, 79, 88]将端到端场景文本识别的二个任务独立地进行处理, 只有少数文献 [59, 137, 143, 174] 将该问题作为一个整体来进行研究. 从 ICDAR 2015 自然场景文本检测及识别竞赛[32] 的结果来看, 在无语料信息的真实环境下的端对端 (End-to-end) 场景文本识别 (Task 4.4) 的最好识别率仅为 34.96 %, 由此可见端对端场景文本识别性能尚有很大的提升空间.

8.2 多方向与形变文本检测问题

从目前所提出的自然场景文本检测方法来看, 大多数方法主要针对直线方向排列 (即: 水平排列、垂直排列以及斜线排列) 的文本进行检测. 然而对于自然场景文本而言, 相邻文本之间的排列方向具有多样性, 除直线方向排列外, 还可能沿弧形等不规则方向排列. 对于多方向排列的场景文本, 其边界框可能是旋转的矩形或者四边形, 因此很难设计有效的方法来统计相邻文本之间排列方向的规律性. 此外, 形变场景文本的视觉特征的不规则性也阻碍了该技术的进一步发展.

8.3 少语种与混合语种文本检测问题

根据 2009 年联合国教科文组织《濒危语言图谱》统计结果表明, 全世界有 7 000 种语言, 其中有 83 种主要语言被 80 % 的人所使用. 目前大部分自然场景文本检测方法只能检测汉字、英文或者阿拉伯数字等单一语种文本或者极少数混合语种文本, 而其他的少数语种以及混合语种的文本检测问题却关注很少. 混合语种文本检测所遇到的挑战主要包括: 文本种类繁多且不同种类文本的空间结构存在多样性, 从而导致了文本区域的视觉特征具有很大的差异, 以致很难找到区分性好的描述特征来分类文本区域与背景区域, 此外, 构建多分类的识别框架也是一件困难的工作.

8.4 文本检测结果评价方法问题

目前的评价方法主要采用交并比 (Intersection-over-union, IoU) 指标来评价文本检测结果, 然而该方法并不能很好地反映文本检测方法的性能. 对于一般性的目标检测问题, 如果能检测出超过 50 % 的 IoU, 则表明获得了很好的检测结果, 然而对于文本检测问题而言, 即使 IoU 获得了大于 50 % 的结果, 也不能保证在后续的文本识别与语义理解中能得到很好的结果, 检测框内的内容和细节同样对后续处理产生很大的影响. 目前绝大部分自然场景文本检测方法采用了 ICDAR 性能评价标准, 在特定的、小规模的公开测试数据集上进行评估, 因此方法性能的鲁棒性还有待于进一步验证.

8.5 文本检测研究内容与创新性问题

对 2017 年在 CVPR、ICCV、NIPS、ICDAR 等顶级会议上发表的自然场景文本相关论文进行分析后发现, 超过 80 % 的自然场景文本检测论文主要关注多方向排列的场景文本检测问题, 较少文献涉及自然场景文本识别与端对端自然场景文本检测与识别问题, 从而导致了目前该领域的研究工作重检测轻识别; 另外, 目前所提出的大部分文本检测方法主要在一些公共数据集上测试性能, 相当一部分方法为了获得更高的测试性能, 往往简单地堆积一些领域知识与反复调整参数 (比如采用 Faster R-CNN, SSD, FCN, RNN 等模式识别领域知识), 从而导致缺乏创新和深度思考, 没有形成文档分析领域特色.

表 2 近期主流自然场景文本检测方法性能总结 (数据都是原文报道的结果, 带 (*) 标记的数据是引自相关论文)

Table 2 Performance summary of recent dominant natural scene text detection methods (All results are quoted directly from original papers, except for those marked with (*), which are from a recent related paper.)

方法	年份	数据集	精度 (P)	召回率 (R)	综合评价指标 (f)	检测耗时 (s)	方法亮点
Lucas[161]	2003	ICDAR′03	0.55	0.46	0.50	8.7	ICDAR′03 竞赛冠军
Hinnerk Becker[170]	2005	ICDAR′03	0.62	0.67	0.62	14.4	ICDAR′05 竞赛冠军
Yao[39]	2012	ICDAR′03	0.69	0.66	0.67	–	提出 MSRA-TD500 数据集, 检测任意方向文本
Epshtein[34]	2010	ICDAR′03	0.73	0.60	0.66	–	首次提出笔画宽度变换文本检测方法
SFT-TCD[72]	2013	ICDAR′03	0.81	0.74	0.72	–	提出笔画特征变换
Neumann[60]	2010	ICDAR′03	0.59	0.55	0.57	–	首次提出 MSER 文本检测方法
Kim[30]	2011	ICDAR′11	0.83	0.63	0.71	–	ICDAR′11 竞赛冠军
SFT-TCD[72]	2013	ICDAR′11	0.82	0.75	0.73	–	提出笔画特征变换
Yin[164]	2015	ICDAR′11	0.84	0.66	0.74	–	提出自适应聚类文本检测
Zhang[87]	2015	ICDAR′11	0.84	0.76	0.80	–	提出文本行上下结构相似的文本检测
Yin[14]	2014	ICDAR′11	0.86	0.68	0.76	–	提出基于 MSER 文本检测
Gupta[148]	2016	ICDAR′11	0.92	0.75	0.82	–	首次提出大规模合成场景文本数据集
Liao[31]	2017	ICDAR′11	0.89	0.82	0.86	0.73	提出端对段卷积神经网络
USTB TexStar[123]	2013	ICDAR′13	0.89	0.67	0.76	–	ICDAR′13 竞赛冠军
Yin[164]	2015	ICDAR′13	0.84	0.65	0.73	–	提出自适应聚类文本检测
Zhang[87]	2015	ICDAR′13	0.88	0.74	0.80	–	提出文本行上下结构相似的文本检测
Zhu[64]	2016	ICDAR′13	0.86	0.74	0.80	–	提出场景上下文检测文本
Zhang[41]	2016	ICDAR′13	0.88	0.78	0.83	–	首次提出基于 FCN 检测任意方向文本
Gupta[148]	2016	ICDAR′13	0.92	0.76	0.83	–	首次提出大规模合成场景文本数据集
Huang[42]	2016	ICDAR′13	0.88	0.72	0.79	–	提出基于视觉注意的文本检测方法
Liao[31]	2017	ICDAR′13	0.88	0.83	0.85	0.73	提出端对段卷积神经网络
Shi[33]	2017	ICDAR′13	0.88	0.83	0.85	20.6	提出改进版的 SSD 文本检测器
Stradvision-2[32]	2015	ICDAR′15	0.78	0.37	0.50	–	ICDAR′15 竞赛冠军
Zhang[41]	2016	ICDAR′15	0.71	0.43	0.54	2.1	首次提出基于 FCN 检测任意方向文本
Zheng[49]	2017	ICDAR′15	0.62	0.40	0.48	–	提出文本行熵方法
Liu[15]	2017	ICDAR′15	0.73	0.68	0.71	–	提出 DMPNet 文本检测网络
Shi[33]	2017	ICDAR′15	0.73	0.77	0.75	–	提出改进版的 SSD 文本检测器
Zhou[50]	2017	ICDAR′15	0.83	0.78	0.81		提出基于 FCN 与 NMS 简单高效的文本框架
Yao[39]	2012	MSRA-TD500	0.63	0.63	0.60	7.2	提出 MSRA-TD500 数据集, 检测任意方向文本
Zhang[41]	2016	MSRA-TD500	0.83	0.67	0.74	–	首次提出基于 FCN 检测任意方向文本
Huang[42]	2016	MSRA-TD500	0.74	0.68	0.71	–	提出基于视觉注意的文本检测方法
Shivakumara[171]	2017	MSRA-TD500	0.68	0.54	0.60	–	提出基于分形 (Fractals) 文本检测
Kang[63]	2014	MSRA-TD500	0.71	0.62	0.66	–	提出高阶关联聚类文本检测
Yin[14]	2014	MSRA-TD500	0.71	0.61	0.66	0.8	提出基于 MSER 文本检测
Yin[164]	2015	MSRA-TD500	0.81	0.63	0.71	1.4	提出自适应聚类文本检测
Zhou[50]	2017	MSRA-TD500	0.87	0.67	0.76		提出基于 FCN 与 NMS 简单高效的文本框架
Shi[33]	2017	MSRA-TD500	0.86	0.70	0.77	8.9	提出改进版的 SSD 文本检测器
Yi[21]	2011	OSTD	0.71	0.62	0.62	17.8	提出组件分析文本检测
Yao[39]	2012	OSTD	0.77	0.73	0.74	–	提出 MSRA-TD500 数据集, 检测任意方向文本
Yin[164]	2015	USTB-SV1K	0.45*	0.45*	0.45*	–	提出基于 MSER 文本检测
Yao[164]	2015	USTB-SV1K	0.46*	0.44*	0.45*	–	提出统一的文本检测与识别框架
Yin[164]	2015	USTB-SV1K	0.50	0.45	0.48	–	提出自适应聚类文本检测
Neumann[59]	2012	SVT	0.19	0.33	0.26	–	提出端对端的文本检测与识别方法
Zhu[64]	2016	SVT	0.41	0.34	0.37	–	提出场景上下文检测文本
Gupta[148]	2016	SVT	0.26	0.27	0.27	–	首次提出大规模合成场景文本数据集
SnooperText[172]	2014	SVT	0.36	0.54	0.43	–	提出自顶向下与自底向上的检测策略
Yin[164]	2015	NEOCR	0.41	0.25	0.31	–	提出自适应聚类文本检测
Yao[104]	2014	COCO-Text	0.3	0.27	0.33	–	提出 Strokelet 文本区域描述方法
Zhou[50]	2017	COCO-Text	0.50	0.32	0.40	–	提出 FCN 与 NMS 简单高效的文本框架
Jin[173]	2011	KAIST	0.85	0.90	–	–	提出 Touchline 文本检测方法
Foo and Bar[51]	2017	RCTW-17	0.74	0.59	0.66	–	RCTW-17 竞赛冠军
NLPR PAL[51]	2017	RCTW-17	0.77	0.57	0.66	–	RCTW-17 竞赛亚军

9 发展趋势及应用

9.1 任务实施步骤层面的几点思考

从自然场景文本检测任务的实施步骤来看, 主要需要解决三个问题: 如何获得文本候选区域、如何验证文本候选区域以及如何得到以单词为分割单元的检测结果.

对于第一个问题, 可以考虑将自顶向下检测方法与自底向上检测方法进行综合运用. 因为我们欣喜地发现, 文本在自然场景中通常表现出聚集性, 相邻文本往往具有高度、宽度与颜色的相似性, 即便是任意方向排列的文本区域, 其相邻文本间的排列方向也具有一定的规律, 因此自然场景文本相较其他的物体往往具有显著的视觉上下文信息. 基于自然场景文本的上述特点, 我们可以考虑将前期自底向上方式处理后所获得的文本区域作为种子区域, 然后将种子区域的大小、排列方向等信息作为先验知识, 为后续将要开展的自顶向下处理方法提供线索, 指导其检测窗口的大小以及搜索方向的设定, 进而可以兼顾检测效率与检测召回率.

对于第二个问题, 近年来基于深度学习的目标检测方法如 Faster R-CNN、YOLO、SSD、R-FCN 等为解决文本检测问题提供了全新的思路. 尽管文本检测属于目标检测中的一个特例, 但是简单地把深度学习中的目标检测框架应用于文本检测问题可能会达不到满意的效果. 然而, 自然场景文本所具有的独特性以及视觉上下文信息使得其具有了其他场景目标所不具备的优势. 如果能设法将文本上下文信息融入深度学习框架, 有望提升文本检测系统的整体性能; 另外, 从训练文本分类器的方式来看, 目前主要采用了监督学习方法, 半监督, 弱监督甚至无监督方法鲜有人关注, 而事实上, 上述学习方式可望在一定程度上减轻方法对大规模训练数据集的依赖.

对于第三个问题, 由于受到文本漏检、误检等因素的影响, 如果只是单向地通过经验或者统计学习方法来设计单词分割规则, 往往难以获取理想的分割结果. 然而我们注意到, 自然场景文本中的单词绝大部分都是常用单词, 尽管单词的总数有几十万个, 但是根据 Test your vocab 网站上两百万份测试的结果, 大部分母语为英语人的单词量为 20 000 ~ 35 000 之间, 国内英语专业研究生毕业应掌握单词量也才 8 000 以上, 因此我们可以考虑基于常用单词建立字典进而对所分割得到的结果进行对比, (尽管单词误分割后有可能刚好成为一个新的单词, 但是这样的几率相对很少), 在此基础上可尝试通过引入反馈机制来指导单词分割. 比如: 如果发生比对错误, 则将出错的信息反馈到单词分割处理的前端, 在单词分割的前端通过对该出错的单词调整其阈值以获得新的分割结果. 此外, 在设计自然场景文本检测方法时, 我们还应该借鉴计算机视觉与模式识别领域的一些新的研究成果, 并挖掘一些技巧性处理办法.

9.2 任务整体层面的思考

尽管自 2012 年以来, 任意方向排列的自然场景文本检测成为了该领域的研究热点, 然而我们却发现绝大部分任意方向排列的文本检测方法[15,33,50,131−134] 的检测对象仅仅是直线方向排列 (即: 水平排列、垂直排列以及斜线排列) 的文本, 只有极少的文献 [135, 136] 对包含弧形排列方向的任意方向排列文本开展了检测研究. 对任意方向排列的文本进行检测时将面临两个关键性问题: 1) 文本区域描述; 2) 文本行的形成.

1) 对于第一个问题, 为了适应文本的旋转变化, 需要设计文本级别以及文本行级别这两组旋转不变的描述特征. 所幸的是, 近年来所提出的深度学习方法[15, 33, 50, 131−134] + 合成文本数据技术[148−149] 已能较好地突破一些传统手工设计的特征[39, 101, 104] 的局限, 因此进一步提出更优的深度学习框架是一个有效的解决方法. 即便如此, 文本行相比其他独立的物体而言, 文本具有着特定的空间结构与语义属性, 因此文本满足一定的 “典型性” 与 “描述性” 特点. 根据文本行的组成特点, 除了设计更优的深度网络框架以外, 我们还可以借助视觉特性好的文本检测结果来提升视觉特性差的文本的检测效果. 文献 [74] 采用自底向上策略从局部特征提取角度来描述文本行区域, 为任意方向排列的文本行以及形变文本的检测问题提供了一种思路.

2) 对于第二个问题, 由于任意方向排列的文本区域其边界框可能是旋转的矩形或者是不规则的四边形, 从而导致传统的一般物体检测方法[93, 95, 175] 所采用的矩形检测框很难有效地应对任意方向排列的文本检测问题. 与此同时, 在文本行形成的过程中连接规则的设计也是一个非常重要的处理步骤, 一些基于连接的检测方法 (Linking methods)[33, 37, 150] 首先检测单个文本, 然后将单个文本通过一定的连接规则融合成文本行, 然而该方法有一定的缺点, 当出现大量堆叠的文本区域或者文本尺寸太小的情况时, 该类方法往往不能获得一个满意的效果. 区别于传统的基于连接的文本检测方法, 文献 [136] 针对任意方向排列的文本 (包含弧形排列方向) 的检测问题提供了一种新的解决思路, 该文献提出了滑动线点回归 (Sliding line point regression, SLPR) 方法. 文献 [135] 在提出基于多边形的弧形方向排列的文本检测方法的基础上, 首次推出了包含弧度方向排列文本的数据集 (SCUT-CTW1500), 从一定程度

上为更广泛的任意方向排列的文本检测研究提供了方便.

9.3 领域知识对文本检测性能影响的几点思考

自然场景文本检测属于一种典型的二分类模式识别问题, 计算机视觉与模式识别领域中的其他目标检测方法可以为自然场景文本检测提供思路.

9.3.1 视觉注意机制对文本检测的影响

文本在自然场景图像中表现出稀疏性特点, 大量的背景区域给真实文本检测带来了极大困难 (特别是一些类文本的背景区域). 事实上, 采用视觉注意机制对特定目标进行显著性检测时, 可以在突出特定目标的同时抑制其他背景信息, 进而减少背景干扰所带来的虚警. 一些研究者[20, 79, 176−182] 对自然场景文本的显著性检测问题展开了研究, 文献 [176−178] 的研究结果表明可以通过构建视觉注意模型来表示文本区域的显著性. 文献 [180] 认为图像中的文本区域并不是所谓 "最显著" 的区域, 因此只使用了文献 [183] 所提出的视觉注意模型中的强度突出图作为显著图. 文献 [79] 提出了一种文本–注意卷积神经网络 (Text-attentional convolutional neural network, Text-CNN), 并采用了多任务学习的方式训练 Text-CNN 模型. 在训练的过程中将低级的像素级信息 (分割问题)、高级的字符多类信息 (62 类字符识别问题) 以及字符与非字符信息 (2 类字符分类问题) 融合到 Text-CNN 模型中, 从而使得 Text-CNN 具有强大的识别歧义文本的能力, 同时也增强了算法在应对复杂背景时的鲁棒性, 最后通过采用训练后的 Text-CNN 对图像进行处理进而获得显著性图像, 在显著性图像中文本区域往往具有高的置信度值, 而背景区域所对应的置信度值较低. 最近, He 等在文献 [184] 中提出了一种视觉注意模型, 该方法通过自动地学习注意图来初步地获得文本区域. 区别于大部分文本显著性检测方法, 文献 [185] 关注于检测背景区域, 反向思维地将检测出的背景区域去除, 从而凸显待检测文本区域. 文献 [41] 则采用了全卷积网络 (Fully convolutional network, FCN) 直接得到文本区域的显著图 (Salient map), 然后基于该显著图进行后续处理. 通过对上述研究工作的分析我们可以发现, 结合自然场景文本的特点合理地设计一个视觉显著性模型将有助于自然场景文本检测问题的解决.

9.3.2 视觉上下文对文本检测的影响

诸如人脸检测、行人检测以及车辆检测等其他物体检测的对象往往是一些独立目标, 然而自然场景文本检测通常需要检测一个文本序列. 尽管自然场景文本种类多样, 由不同文本任意组合而成的文本行区域的视觉特征差异较大, 但是我们欣喜地发现: 对于某一特定语种其包含的文本种类是有限的, 而且文本序列中的相邻文本之间通常具有着独特的上下文信息, 比如: 相邻文本之间具有相近的高度、颜色、笔画宽度以及均匀的间隔距离等. 根据自然场景文本的上述特点, 如果我们能合理地利用相邻文本间的上下文信息, 无疑将有助于提高文本区域的分类正确性. 除此以外, 自然场景文本检测的目标是判断给定的图像区域中是否包含文本, 并不关心所包含文本的具体种类, 因此自然场景文本检测属于二分类模式识别问题, 从而为利用视觉上下文信息提供了便利. 近年来, 一些研究者开始关注自然场景文本视觉上下文信息对文本检测性能的影响, 文献 [35, 64, 74, 80, 184, 186−189] 通过利用相邻文本间的视觉上下文信息设计了不同的文本检测方法并取得了满意的检测结果. 通过对前期研究工作的分析我们可以发现, 在深度学习的框架内合理地融入文本视觉上下文信息可望有效地提升文本检测的性能.

9.4 应用层面的几点思考

基于文本的高度抽象描述能力, 自然场景文本检测技术具有广泛的应用价值. 在应用需求的驱动下, 目前自然场景文本检测技术在一些特定领域中获得了应用, 比如: 智能交通系统 (如: 美国 Hi-Tech 公司的 See/Car System 以及香港 Asia Vision Technology 公司的 VECON-VIS 等); 基于内容的视频检索系统 (如: 美国卡耐基梅隆大学的 Informedia Digital Video Library[190] 以及美国哥伦比亚大学的 WebSeek[191] 等); 可穿戴/便携式视觉系统 (美国麻省理工学院的 FingerReader[9] 以及 Goggles[10] 等). 除了上述应用以外, 一些研究者还将自然场景文本检测技术应用到图像理解[192], 文种识别[193] 等领域. 相比自然场景文本检测技术的潜在应用市场, 上述应用只是 "冰山一角".

文本具有高层的语义信息, 而语义信息往往能有助于解决计算机视觉中的一些传统问题以及拓展新的应用, 比如在特定目标 (如运动员、汽车) 的跟踪与重检测的问题上, 我们可以引入运动员的标牌或者汽车的车牌来帮助实现上述任务; 再如无人驾驶汽车的辅助导航问题上, 我们也可以通过引入自然场景文本检测技术来获得交通标识信息, 通过识别交通标识牌的语义信息来提高汽车的智能感知与行驶规划能力; 还有无纸化办公方面也可望使用文本检测技术, 对于会议后书写在白板上的工作安排, 我们只需用智能设备拍照留存与分析处理, 系统将根据白板上的文本识别结果来分类相关人员的后续工作. 另外, 自然场景文本检测技术还可以与音频信息结合起来共同解决诸如 "以词搜图" 的图片检索、

地图定点导航等实际问题. 作为一项面向具体应用场景的技术, 自然场景文本检测的应用领域将在各种应用需求的驱动下不断拓展、不断成熟.

9.5 其他问题的思考

1) 据报道, 人脸的识别在大脑中有专用机构[194−195], 那么是否在大脑里存在类似的专用机构处理文本的识别问题? 尽管目前的深度学习是最接近人脑思维过程, 相信神经生物学家未来的研究成果将有助于深入理解大脑的工作原理, 进而为构造更有效的文本识别机制提供依据.

2) 尽管深度学习在文本表示方面展现出显著的优势, 但是自然场景文本相对图片而言其尺寸较小, 网络的深度太深可能会对文本识别产生大的影响, 从而面临着网络的深度规模如何选取的问题.

3) 对于多语种文本检测是否会存在分类性能好且通用的描述特征? 通过观察我们发现: 如果一个中国小孩不学英文, 是不具备检测与识别英文文本的能力的.

10 结束语

自然场景文本检测是计算机视觉与模式识别领域中的一个新兴的研究课题, 具有重要的理论意义和实际应用价值. 国内外许多学者对该课题展开了大量研究, 然而复杂自然环境中所存在的诸多挑战使得该技术与实际实用仍然有一定距离. 为了全面分析文本检测中的问题, 本文对自然场景文本检测技术的研究背景与意义、发展现状等内容进行了阐述、对该技术的方法进行了详细的梳理和评述, 并揭示了它们之间内在联系、优势与不足. 与此同时, 本文介绍了端对端文本识别技术, 并对计算机视觉与模式识别领域中的一些新发展对自然场景文本检测技术的影响进行了介绍, 拓宽了研究思路; 本文还对一些主流数据库进行了总结和评述, 并列举了目前一些主流方法的性能参数; 在此基础上, 对自然场景文本检测技术的未来发展方向以及该技术的一些潜在的应用领域进行了分析与展望. 我们有理由相信, 计算机视觉与机器学习领域的进步, 将极大地促进自然场景文本检测问题的解决; 与此同时, 文本检测技术中的关键性问题的突破也将启发计算机视觉相关领域的发展.

References

1 González Á, Bergasa L M, Yebes J J. Text detection and recognition on traffic panels from street-level imagery using visual appearance. *IEEE Transactions on Intelligent Transportation Systems*, 2014, **15**(1): 228−238

2 Zhou W G, Li H Q, Lu Y J, Tian Q. Principal visual word discovery for automatic license plate detection. *IEEE Transactions on Image Processing*, 2012, **21**(9): 4269−4279

3 Greenhalgh J, Mirmehdi M. Recognizing text-based traffic signs. *IEEE Transactions on Intelligent Transportation Systems*, 2015, **16**(3): 1360−1369

4 Ezaki N, Kiyota K, Minh B T, Bulacu M, Schomaker L. Improved text-detection methods for a camera-based text reading system for blind persons. In: Proceedings of the 8th International Conference on Document Analysis and Recognition. Seoul, South Korea: IEEE, 2005. 257−261

5 Ezaki N, Bulacu M, Schomaker L. Text detection from natural scene images: towards a system for visually impaired persons. In: Proceedings of the 17th International Conference on Pattern Recognition. Cambridge, UK: IEEE, 2004. 683−686

6 Jung K, Kim K I, Jain A K. Text information extraction in images and video: a survey. *Pattern Recognition*, 2004, **37**(5): 977−997

7 Hedgpeth T, Black J A Jr, Panchanathan S. A demonstration of the icare portable reader. In: Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility. Portland, Oregon, USA: ACM, 2006. 279−280

8 Goto H, Tanaka M. Text-tracking wearable camera system for the blind. In: Proceedings of the 10th International Conference on Document Analysis and Recognition. Barcelona, Spain: IEEE, 2009. 141−145

9 Shilkrot R, Huber J, Liu C, Maes P, Nanayakkara S C. FingerReader: a wearable device to support text reading on the go. In: Proceedings of the CHI′14 Extended Abstracts on Human Factors in Computing Systems. Toronto, Ontario, Canada: ACM, 2014. 2359−2364

10 Google goggles [Online], available: http://www.google.com/mobile/goggles/#text, January 10, 2015

11 Bai X, Shi B G, Zhang C Q, Cai X, Qi L. Text/non-text image classification in the wild with convolutional neural networks. *Pattern Recognition*, 2017, **66**: 437−446

12 Shi B G, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(11): 2298−2304

13 Pan Y F, Hou X W, Liu C L. A hybrid approach to detect and localize texts in natural scene images. *IEEE Transactions on Image Processing*, 2011, **20**(3): 800−813

14 Yin X C, Yin X W, Huang K Z, Hao H W. Robust text detection in natural scene images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, **36**(5): 970−983

15 Liu Y L, Jin L W. Deep matching prior network: toward tighter multi-oriented text detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017. 3454−3461

16 He P, Huang W L, Qiao Y, Loy C C, Tang X O. Reading scene text in deep convolutional sequences. In: Proceedings of the 13th AAAI Conference on Artificial Intelligence. Phoenix, Arizona, USA: AAAI Press, 2016. 3501−3508

17 Ye Q X, Jiao J B, Huang J, Yu H. Text detection and restoration in natural scene images. *Journal of Visual Communication and Image Representation*, 2007, **18**(6): 504−513

18 Shi C Z, Wang C H, Xiao B H, Zhang Y, Gao S. Scene text detection using graph model built upon maximally stable extremal regions. *Pattern Recognition Letters*, 2013, **34**(2): 107−116

19 Huang W L, Qiao Y, Tang X O. Robust scene text detection with convolution neural network induced MSER trees. In: Proceedings of the 13th European Conference on Computer Vision. Cham, Switzerland: Springer, 2014. 497−511

20 Li Y, Jia W J, Shen C H, van den Hengel A. Characterness: an indicator of text in the wild. *IEEE Transactions on Image Processing*, 2014, **23**(4): 1666−1677

21 Yi C C, Tian Y L. Text string detection from natural scenes by structure-based partition and grouping. *IEEE Transactions on Image Processing*, 2011, **20**(9): 2594−2605

22 Ye Q X, Doermann D. Text detection and recognition in imagery: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**(7): 1480−1500

23 Zhang H G, Zhao K L, Song Y Z, Guo J. Text extraction from natural scene image: a survey. *Neurocomputing*, 2013, **122**: 310−323

24 Zhu Y Y, Yao C, Bai X. Scene text detection and recognition: recent advances and future trends. *Frontiers of Computer Science*, 2016, **10**(1): 19−36

25 Yin X C, Zuo Z Y, Tian S, Liu C L. Text detection, tracking and recognition in video: a comprehensive survey. *IEEE Transactions on Image Processing*, 2016, **25**(6): 2752−2773

26 Jin Lian-Wen, Zhong Zhuo-Yao, Yang Zhao, Yang Wei-Xin, Xie Ze-Cheng, Sun Jun. Applications of deep learning for handwritten Chinese character recognition: a review. *Acta Automatica Sinica*, 2016, **42**(8): 1125−1141
(金连文, 钟卓耀, 杨钊, 杨维信, 谢泽澄, 孙俊. 深度学习在手写汉字识别中的应用综述, 自动化学报, 2016, **42**(8): 1125−1141)

27 Ohya J, Shio A, Akamatsu S. Recognizing characters in scene images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994, **16**(2): 214−220

28 Zhong Y, Karu K, Jain A K. Locating text in complex color images. *Pattern Recognition*, 1995, **28**(10): 1523−1535

29 Lee C M, Kankanhalli A. Automatic extraction of characters in complex scene images. *International Journal of Pattern Recognition and Artificial Intelligence*, 1995, **9**(1): 67−82

30 Shahab A, Shafait F, Dengel A. ICDAR 2011 robust reading competition challenge 2: reading text in scene images. In: Proceedings of the 2011 International Conference on Document Analysis and Recognition. Beijing, China: IEEE, 2011. 1491−1496

31 Liao M H, Shi B G, Bai X, Wang X G, Liu W Y. Textboxes: a fast text detector with a single deep neural network. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, CA, USA: AAAI, 2017. 4161−4167

32 Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, et al. ICDAR 2015 competition on robust reading. In: Proceedings of the 13th International Conference on Document Analysis and Recognition. Tunis, Tunisia: IEEE, 2015. 1156−1160

33 Shi B G, Bai X, Belongie S. Detecting oriented text in natural images by linking segments. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017. 3482−3490

34 Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2010. 2963−2970

35 Wang R M, Sang N, Gao C X. Text detection approach based on confidence map and context information. *Neurocomputing*, 2015, **157**: 153−165

36 Yi C C, Tian Y L. Text extraction from scene images by character appearance and structure modeling. *Computer Vision and Image Understanding*, 2013, **117**(2): 182−194

37 Tian Z, Huang W L, He T, He P, Qiao Y. Detecting text in natural image with connectionist text proposal network. In: Proceedings of the 14th European Conference on Computer Vision. Cham, Switzerland: Springer, 2016. 56−72

38 Shivakumara P, Phan T Q, Tan C L. A Laplacian approach to multi-oriented text detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(2): 412−419

39 Yao C, Bai X, Liu W Y, Ma Y, Tu Z W. Detecting texts of arbitrary orientations in natural images. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 1083−1090

40 Yao C, Zhang X, Bai X, Liu W Y, Ma Y, Tu Z W. Rotation-invariant features for multi-oriented text detection in natural images. *PLoS One*, 2013, **8**(8): e70173

41 Zhang Z, Zhang C Q, Shen W, Yao C, Liu W Y, Bai X. Multi-oriented text detection with fully convolutional networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 4159−4167

42 Huang W Y, He D F, Yang X, Zhou Z H, Kifer D, Giles C L. Detecting arbitrary oriented text in the wild with a visual attention model. In: Proceedings of the 2016 ACM on Multimedia Conference. Amsterdam, The Netherlands: ACM, 2016. 551−555

43 Raza A, Siddiqi I, Djeddi C, Ennaji A. Multilingual artificial text detection using a cascade of transforms. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, D. C., USA: IEEE, 2013. 309−313

44 Ren X H, Zhou Y, He J H, Chen K, Yang X K, Sun J. A convolutional neural network based Chinese text detection algorithm via text structure modeling. *IEEE Transactions on Multimedia*, 2017, **19**(3): 506−518

45 Yan Jian-Qiang. Text Detection and Recognition in Complex Scene of Image and Video [Ph. D. dissertation], Xidian University, China, 2014
(颜建强. 图像视频复杂场景中文字检测识别方法研究 [博士学位论文], 西安电子科技大学, 中国, 2014)

46 Wang T, Wu D J, Coates A, Ng A Y. End-to-end text recognition with convolutional neural networks. In: Proceedings of the 21st International Conference on Pattern Recognition. Tsukuba, Japan: IEEE, 2012. 3304−3308

47 Sun L, Huo Q, Jia W, Chen K. Robust text detection in natural scene images by generalized color-enhanced contrasting extremal region and neural networks. In: Proceedings of the 22th International Conference on Pattern Recognition. Stockholm, Sweden: IEEE, 2014. 2715−2720

48 Jaderberg M, Vedaldi A, Zisserman A. Deep features for text spotting. In: Proceedings of the 13th European Conference on Computer Vision. Cham, Switzerland: Springer, 2014. 512−528

49 Zheng Y, Li Q, Liu J, Liu H P, Li G, Zhang S W. A cascaded method for text detection in natural scene images. *Neurocomputing*, 2017, **238**: 307−315

50 Zhou X Y, Yao C, Wen H, Wang Y Z, Zhou S C, He W R, et al. East: an efficient and accurate scene text detector. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017. 2642−2651

51 Shi B G, Yao C, Liao M H, Yang M K, Xu P, Cui L Y, et al. ICDAR2017 competition on reading Chinese text in the wild (RCTW-17). arXiv preprint arXiv: 1708.09585, 2017

52 Lienhart R W, Stuber F. Automatic text recognition in digital videos. In: Proceedings Volume 2666, Image and Video Processing IV. San Jose, CA, USA: SPIE, 1996. 180−188

53 Busta M, Neumann L, Matas J. FASText: efficient unconstrained scene text detector. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1206−1214

54 Liu X Q, Samarabandu J. Multiscale edge-based text extraction from complex images. In: Proceedings of the 2006 IEEE International Conference on Multimedia and Expo. Toronto, Ont. Canada: IEEE, 2006. 1721−1724

55 Liu C M, Wang C H, Dai R W. Text detection in images based on unsupervised classification of edge-based features. In: Proceedings of the 8th International Conference on Document Analysis and Recognition. Seoul, South Korea: IEEE, 2005. 610−614

56 Jamil A, Siddiqi I, Arif F, Raza A. Edge-based features for localization of artificial urdu text in video images. In: Proceedings of the 11th International Conference on Document Analysis and Recognition. Beijing, China: IEEE, 2011. 1120−1124

57 Yu C, Song Y H, Meng Q, Zhang Y L, Liu Y. Text detection and recognition in natural scene with edge analysis. *IET Computer Vision*, 2015, **9**(4): 603−613

58 Veit A, Matera T, Neumann L, Matas J, Belongie S. COCO-text: dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv: 1601.07140, 2016

59 Neumann L, Matas J. Real-time scene text localization and recognition. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 3538−3545

60 Neumann L, Matas J. A method for text localization and recognition in real-world images. In: Proceedings of the 10the Asian Conference on Computer Vision. Berlin, Heidelberg, Germany: Springer, 2010. 770−783

61 González A, Bergasa L M, Yebes J, Bronte S. Text location in complex images. In: Proceedings of the 21st International Conference on Pattern Recognition. Tsukuba, Japan: IEEE, 2012. 617−620

62 Gómez L, Karatzas D. MSER-based real-time text detection and tracking. In: Proceedings of the 22th International Conference on Pattern Recognition. Stockholm, Sweden: IEEE, 2014. 3110−3115

63 Kang L, Li Y, Doermann D. Orientation robust text line detection in natural images. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 4034−041

64 Zhu A N, Gao R W, Uchida S. Could scene context be beneficial for scene text detection. *Pattern Recognition*, 2016, **58**: 204−251

65 Sun L, Huo Q. An improved component tree based approach to user-intention guided text extraction from natural scene images. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, DC, USA: IEEE, 2013. 383−387

66 Mancas-Thillou C, Gosselin B. Color text extraction with selective metric-based clustering. *Computer Vision and Image Understanding*, 2007, **107**(1−2): 97−107

67 Lai A N, Park K, Kumar M, Lee G. Korean text extraction by local color quantization and k-means clustering in natural scene. In: Proceedings of the 1st Asian Conference on Intelligent Information and Database Systems. Dong Hoi, Vietnam: IEEE, 2009. 138−143

68 Garg R, Hassan E, Chaudhury S, Gopal M. A CRF based scheme for overlapping multi-colored text graphics separation. In: Proceedings of the 11th International Conference on Document Analysis and Recognition. Beijing, China: IEEE, 2011. 1215−1219

69 Zhou G, Liu Y H, Tian Z Q, Su Y Q. A new hybrid method to detect text in natural scene. In: Proceedings of the 18th IEEE International Conference on Image Processing. Brussels, Belgium: IEEE, 2011. 2605−2608

70 Mosleh A, Bouguila N, Hamza A B. Image text detection using a bandlet-based edge detector and stroke width transform. In: Proceedings of the British Machine Vision Conference. BMVA Press, 2012. 63.1−63.12

71 Karthikeyan S, Jagadeesh V, Manjunath B S. Learning bottom-up text attention maps for text detection using stroke width transform. In: Proceedings of the 20th IEEE International Conference on Image Processing. Melbourne, VIC, Australia: IEEE, 2013. 3312−3316

72 Huang W L, Lin Z, Yang J C, Wang J. Text localization in natural images using stroke feature transform and text covariance descriptors. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia: IEEE, 2013. 1241−1248

73 Milyaev S, Barinova O, Novikova T, Kohli P, Lempitsky V. Image binarization for end-to-end text understanding in natural images. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, DC, USA: IEEE, 2013. 128−132

74 Wang R M, Sang N, Gao C X. Scene text identification by leveraging mid-level patches and context information. *IEEE Signal Processing Letters*, 2015, **22**(7): 963−967

75 Wei Y W, Zhang Z J, Shen W, Zeng D, Fang M, Zhou S F. Text detection in scene images based on exhaustive segmentation. *Signal Processing: Image Communication*, 2017, **50**: 1−8

76 Yao Cong. Research on text detection and recognition in natural images [Ph. D. dissertation], Huazhong University of Science and Technology, China, 2014
(姚聪. 自然图像中文字检测与识别研究 [博士学位论文], 华中科技大学, 中国, 2014)

77 Chen H Z, Tsai S S, Schroth G, Chen D M, Grzeszczuk R, Girod B. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In: Proceedings of the 18th IEEE International Conference on Image Processing. Brussels, Belgium: IEEE, 2011. 2609−2612

78 Koo H I, Kim D H. Scene text detection via connected component clustering and nontext filtering. *IEEE Transactions on Image Processing*, 2013, **22**(6): 2296−2305

79 He T, Huang W L, Qiao Y, Yao J. Text-attentional convolutional neural network for scene text detection. *IEEE Transactions on Image Processing*, 2016, **25**(6): 2529−2541

80 Li Y, Shen C H, Jia W J, van den Hengel A. Leveraging surrounding context for scene text detection. In: Proceedings of the 20th IEEE International Conference on Image Processing. Melbourne, VIC, Australia: IEEE, 2013. 2264−2268

81 González Á, Bergasa L M. A text reading algorithm for natural images. *Image and Vision Computing*, 2013, **31**(3): 255−274

82 Sun L, Huo Q. A component-tree based method for user-intention guided text extraction. In: Proceedings of the 21st International Conference on Pattern Recognition. Tsukuba, Japan: IEEE, 2012. 633−636

83 Sun L, Huo Q, Jia W, Chen K. A robust approach for text detection from natural scene images. *Pattern Recognition*, 2015, **48**(9): 2906−2920

84 Louloudis G, Gatos B, Pratikakis I, Halatsis C. Text line detection in handwritten documents. *Pattern Recognition*, 2008, **41**(12): 3758−3772

85 Rabaev I, Biller O, El-Sana J, Kedem K, Dinstein I. Text line detection in corrupted and damaged historical manuscripts. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, D. C., USA: IEEE, 2013. 812−816

86 van Beusekom J, Shafait F, Breuel T M. Combined orientation and skew detection using geometric text-line modeling. *International Journal on Document Analysis and Recognition*, 2010, **13**(2): 79−92

87 Zhang Z, Shen W, Yao C, Bai X. Symmetry-based text line detection in natural scenes. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 2558−2567

88 Wang K, Belongie S. Word spotting in the wild. In: Proceedings of the 11th European Conference on Computer Vision. Berlin, Heidelberg, Germany: Springer, 2010. 591−604

89 Hanif S M, Prevost L, Negri P A. A cascade detector for text detection in natural scene images. In: Proceedings of the 19th International Conference on Pattern Recognition. Tampa, FL, USA: IEEE, 2008. 1−4

90 Mishra A, Alahari K, Jawahar C V. Top-down and bottom-up cues for scene text recognition. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 2687−2694

91 Wang K, Babenko B, Belongie S. End-to-end scene text recognition. In: Proceedings of the 2011 IEEE International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011. 1457−1464

92 Tian S X, Pan Y F, Huang C, Lu S J, Yu K, Tan C L. Text flow: a unified text detection system in natural scene images. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 4651−4659

93 Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA: IEEE, 2001. I-511−I-518

94 Lienhart R, Kuranov A, Pisarevsky V. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: Proceedings of DAGM 25th Pattern Recognition. Berlin, Heidelberg, Germany: Springer, 2003. 297−304

95 Viola P, Jones M J. Robust real-time face detection. *International Journal of Computer Vision*, 2004, **57**(2): 137−154

96 Lee J J, Lee P H, Lee S W, Yuille A, Koch C. Adaboost for text detection in natural scene. In: Proceedings of the 2011 International Conference on Document Analysis and Recognition. Beijing, China: IEEE, 2011. 429−434

97 Chen X R, Yuille A L. Detecting and reading text in natural scenes. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC, USA: IEEE, 2004. II-366−II-373

98 Hanif S M, Prevost L. Text detection and localization in complex scene images using constrained AdaBoost algorithm. In: Proceedings of the 10th International Conference on Document Analysis and Recognition. Barcelona, Spain: IEEE, 2009. 1−5

99 Shivakumara P, Huang W H, Phan T Q, Tan C L. Accurate video text detection through classification of low and high contrast images. *Pattern Recognition*, 2010, **43**(6): 2165−2185

100 Kim W, Kim C. A new approach for overlay text detection and extraction from complex video scene. *IEEE Transactions on Image Processing*, 2009, **18**(2): 401−411

101 Wei Y C, Lin C H. A robust video text detection approach using SVM. *Expert Systems with Applications*, 2012, **39**(12): 10832−10840

102 Anthimopoulos M, Gatos B, Pratikakis I. A two-stage scheme for text detection in video images. *Image and Vision Computing*, 2010, **28**(9): 1413−1426

103 Fabrizio J, Marcotegui B, Cord M. Text detection in street level images. *Pattern Analysis and Applications*, 2013, **16**(4): 519−533

104 Yao C, Bai X, Shi B G, Liu W Y. Strokelets: a learned multi-scale representation for scene text recognition. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 4042−4049

105 Su B L, Lu S J, Tian S X, Lim J H, Tan C L. Character recognition in natural scenes using convolutional co-occurrence HOG. In: Proceedings of the 22nd International Conference on Pattern Recognition. Stockholm, Sweden: IEEE, 2014. 2926−2931

106 Tian S X, Lu S J, Su B L, Tan C L. Scene text recognition using co-occurrence of histogram of oriented gradients. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, D. C., USA: IEEE, 2013. 912−916

107 Minetto R, Thome N, Cord M, Leite N J, Stolfi J. T-HOG: an effective gradient-based descriptor for single line text regions. *Pattern Recognition*, 2013, **46**(3): 1078−1090

108 Yan J Q, Li J, Gao X B. Chinese text location under complex background using Gabor filter and SVM. *Neurocomputing*, 2011, **74**(17): 2998−3008

109 Leon M, Vilaplana V, Gasull A, Marques F. Caption text extraction for indexing purposes using a hierarchical region-based image model. In: Proceedings of the 16th IEEE International Conference on Image Processing. Cairo, Egypt: IEEE, 2009. 1869−1872

110 Ye Q X, Huang Q M, Gao W, Zhao D B. Fast and robust text detection in images and video frames. *Image and Vision Computing*, 2005, **23**(6): 565−576

111 Kim K I, Jung K, Kim J H. Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, **25**(12): 1631−1639

112 Nguyen C D, Ardabilian M, Chen L M. Robust car license plate localization using a novel texture descriptor. In: Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance. Genova, Italy: IEEE, 2009. 523−528

113 Chen D T, Bourlard H, Thiran J P. Text identification in complex background using SVM. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA: IEEE, 2001. II-621−II-626

114 Lee C Y, Bhardwaj A, Di W, Jagadeesh V, Piramuthu R. Region-based discriminative feature pooling for scene text recognition. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 4050−4057

115 Chen X R, Yuille A L. Detecting and reading text in natural scenes. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC, USA: IEEE, 2004. II-366−II-373

116 Yin X W, Yin X C, Hao H W, Iqbal K. Effective text localization in natural scene images with MSER, geometry-based grouping and AdaBoost. In: Proceedings of the 21st IEEE International Conference on Pattern Recognition. Tsukuba, Japan: IEEE, 2012. 725−728

117 Shi C Z, Wang C H, Xiao B H, Gao S, Hu J L. End-to-end scene text recognition using tree-structured models. *Pattern Recognition*, 2014, **47**(9): 2853−2866

118 Zhang Y, Wang C H, Xiao B H, Shi C Z. A new method for text verification based on random forests. In: Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition. Washington, DC, USA: IEEE, 2012. 109−113

119 Xu H L, Su F. A robust hierarchical detection method for scene text based on convolutional neural networks. In: Proceedings of the 2015 IEEE International Conference on Multimedia and Expo. Turin, Italy: IEEE, 2015. 1−6

120 Li H P, Doermann D, Kia O. Automatic text detection and tracking in digital video. *IEEE Transactions on Image Processing*, 2000, **9**(1): 147−156

121 Shi C Z, Wang C H, Xiao B H, Zhang Y, Gao S, Zhang Z. Scene text recognition using part-based tree-structured character detections. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2013. 2961−2968

122 Karatzas, D, Mestre S R, Mas J, Nourbakhsh F, Roy P P. ICDAR 2011 Robust Reading Competition - Challenge 1: Reading Text in Born-Digital Images (Web and Email). In: Proceedings of the International Conference on Document Analysis and Recognition. Beijing, China: IEEE, 2011. 1485−1490

123 Karatzas D, Shafait F, Uchida S, Iwamura M, i Bigorda L G, Mestre S R, et al. ICDAR 2013 robust reading competition. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, D. C., USA: IEEE, 2013. 1484−1493

124 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, **521**(7553): 436−444

125 Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527−1554

126 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504−507

127 Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324

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

129 Simard P Y, Steinkraus D, Platt J C. Best practices for convolutional neural networks applied to visual document analysis. In: Proceedings of the 7th International Conference on Document Analysis and Recognition. Washington, DC, USA: IEEE, 2003. 958−963

130 Zhong Z Y, Jin L W, Zhang S Y, Feng Z Y. DeepText: a unified framework for text proposal generation and text detection in natural images. arXiv preprint arXiv: 1605.07314, 2016

131 Liao M H, Shi B G, Bai X. Textboxes++: a single-shot oriented scene text detector. arXiv preprint arXiv: 1801.02765, 2018

132 He W H, Zhang X Y, Yin F, Liu C L. Deep direct regression for multi-oriented scene text detection. arXiv preprint arXiv: 1703.08289, 2017

133 Dai Y C, Huang Z, Gao Y T, Xu Y X, Chen K, Guo J, et al. Fused text segmentation networks for multi-oriented scene text detection. arXiv preprint arXiv: 1709.03272, 2017

134 Jiang Y Y, Zhu X Y, Wang X B, Yang S L, Li W, Wang H, et al. R2CNN: rotational region CNN for orientation robust scene text detection. arXiv preprint arXiv: 1706.09579, 2017

135 Liu Y L, Jin L W, Zhang S T, Zhang S. Detecting curve text in the wild: new dataset and new solution. arXiv preprint arXiv: 1712.02170, 2017

136 Zhu Y X, Du J. Sliding line point regression for shape robust scene text detection. arXiv preprint arXiv: 1801.09969, 2018

137 Jaderberg M, Simonyan K, Vedaldi A, Zisserman A. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 2016, **116**(1): 1−20

138 Ma J Q, Shao W Y, Ye H, Wang L, Wang H, Zheng Y B, et al. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, DOI: 10.1109/TMM.2018.2818020

139 Xiang D L, Guo Q, Xia Y. Robust text detection with vertically-regressed proposal network. In: Proceedings of the European Conference on Computer Vision. Cham, Switzerland: Springer, 2016. 351−363

140 Yao C, Bai X, Sang N, Zhou X Y, Zhou S C, Cao Z M. Scene text detection via holistic, multi-channel prediction. arXiv preprint arXiv: 1606.09002, 2016

141 Yang X, He D F, Zhou Z H, Kifer D, Giles C L. Learning to read irregular text with attention mechanisms. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017. 3280−3286

142 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of the International Conference on Neural Information Processing Systems. Istanbul, Turkey, 2015. 91−99

143 Li H, Wang P, Shen C H. Towards end-to-end text spotting with convolutional recurrent neural networks. arXiv preprint arXiv: 1707.03985, 2017

144 Liu X B, Liang D, Yan S, Chen D G, Qiao Y, Yan J J. FOTS: fast oriented text spotting with a unified network. arXiv preprint arXiv: 1801.01671, 2018

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

146 Deng D, Liu H F, Li X L, Cai D. PixelLink: detecting scene text via instance segmentation. arXiv preprint arXiv: 1801.01315, 2018

147 Lyu P, Yao C, Wu W H, Yan S C, Bai X. Multi-oriented scene text detection via corner localization and region segmentation. arXiv preprint arXiv: 1802.08948, 2018

148 Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 2315−2324

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

150 Hu H, Zhang C Q, Luo Y X, Wang Y Z, Han J Y, Ding E R. WordSup: exploiting word annotations for character based text detection. arXiv preprint arXiv: 1708.06720, 2017

151 Chen D T, Odobez J M, Bourlard H. Text detection and recognition in images and video frames. *Pattern Recognition*, 2004, **37**(3): 595−608

152 Neumann L, Matas J. Scene text localization and recognition with oriented stroke detection. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia: IEEE, 2013. 97−104

153 Bissacco A, Cummins M, Netzer Y, Neven H. PhotoOCR: reading text in uncontrolled conditions. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia: IEEE, 2013. 785−792

154 Pan Y F, Hou X W, Liu C L. Text localization in natural scene images based on conditional random field. In: Proceedings of the 10th International Conference on Document Analysis and Recognition. Barcelona, Spain: IEEE, 2009. 6−10

155 Bartz C, Yang H J, Meinel C. STN-OCR: a single neural network for text detection and text recognition. arXiv preprint arXiv: 1707.08831, 2017

156 Bušta M, Neumann L, Matas J. Deep TextSpotter: an end-to-end trainable scene text localization and recognition framework. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2223−2231

157 Alsharif O, Pineau J. End-to-end text recognition with hybrid HMM maxout models. arXiv preprint arXiv: 1310.1811, 2013

158 Yin F, Wu Y C, Zhang X Y, Liu C L. Scene text recognition with sliding convolutional character models. arXiv preprint arXiv: 1709.01727, 2017

159 Patel Y, Bušta M, Matas J. E2E-MLT — an unconstrained end-to-end method for multi-language scene text. arXiv preprint arXiv: 1801.09919, 2018

160 Bartz C, Yang H J, Meinel C. SEE: towards semi-supervised end-to-end scene text recognition. arXiv preprint arXiv: 1712.05404, 2017

161 Lucas S M, Panaretos A, Sosa L, Tang A, Wong S, Young R. ICDAR 2003 robust reading competitions. In: Proceedings of the 7th International Conference on Document Analysis and Recognition. Edinburgh, UK: IEEE, 2003. 682−687

162 Lee S, Cho M S, Jung K, Kim J H. Scene text extraction with edge constraint and text collinearity. In: Proceedings of the 20th International Conference on Pattern Recognition. Istanbul, Turkey: IEEE, 2010. 3983−3986

163 Nagy R, Dicker A, Meyer-Wegener K. NEOCR: a configurable dataset for natural image text recognition. In: Proceedings of the International Workshop on Camera-Based Document Analysis and Recognition. Berlin, Heidelberg, Germany: Springer, 2011. 150−163

164 Yin X C, Pei W Y, Zhang J, Hao H W. Multi-orientation scene text detection with adaptive clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**(9): 1930−1937

165 Su P. COCO-text explorer. Cornell University CS Department MEng Report, 2016.

166 Wolf C, Jolion J M. Object count/area graphs for the evaluation of object detection and segmentation algorithms. *International Journal of Document Analysis and Recognition*, 2006, **8**(4): 280−296

167 Everingham M, Eslami S M, van Gool L, Williams C K I, Winn J, Zisserman A. The pascal visual object classes challenge: a retrospective. *International Journal of Computer Vision*, 2015, **111**(1): 98−136

168 Freeman H, Shapira R. Determining the minimum-area encasing rectangle for an arbitrary closed curve. *Magazine Communications of the ACM*, 1975, **18**(7): 409−413

169 Everingham M, van Gool L, Williams C K I, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010, **88**(2): 303−338

170 Lucas S M. ICDAR 2005 text locating competition results. In: Proceedings of the 8th International Conference on Document Analysis and Recognition. Seoul, South Korea: IEEE, 2005. 80−84

171 Shivakumara P, Wu L, Lu T, Tan C L, Blumenstein M, Anami B S. Fractals based multi-oriented text detection system for recognition in mobile video images. *Pattern Recognition*, 2017, **68**: 158−174

172 Minetto R, Thome N, Cord M, Leite N J, Stolfi J. SnooperText: a text detection system for automatic indexing of urban scenes. *Computer Vision and Image Understanding*, 2014, **122**: 92−104

173 Jung J, Lee S, Cho M S, Kim J H. Touch TT: scene text extractor using touchscreen interface. *ETRI Journal*, 2011, **33**(1): 78−88

174 Lyu P Y, Liao M H, Yao C, Wu W H, Bai X. Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes. In: Proceedings of the Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 71−88

175 Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. I-511−I-518

176 Judd T, Ehinger K, Durand F, Torralba A. Learning to predict where humans look. In: Proceedings of the 12th International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 2106−2113

177 Wang H C, Pomplun M. The attraction of visual attention to texts in real-world scenes. *Journal of Vision*, 2012, **12**(6): 26

178 Shahab A, Shafait F, Dengel A, Uchida S. How salient is scene text? In: Proceedings of the 10th IAPR International Workshop on Document Analysis Systems. Washington D. C., USA: IEEE, 2012. 317−321

179 Karaoglu S, van Gemert J C, Gevers T. Object reading: text recognition for object recognition. In: Proceedings of the 12th European Conference on Computer Vision. Berlin, Heidelberg, Germany: Springer, 2012. 456−465

180 Sun Q Y, Lu Y, Sun S L. A visual attention based approach to text extraction. In: Proceedings of the 20th International Conference on Pattern Recognition. Istanbul, Turkey: IEEE, 2010. 3991−3995

181 Mesquita R G, Mello C A B. Finding text in natural scenes by visual attention? In: Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics. Manchester, UK: IEEE, 2013. 4243−4247

182 Gao R W, Uchida S, Shahab A, Shafait F, Frinken V. Visual saliency models for text detection in real world. *PLoS One*, 2014, **9**(12): e114539

183 Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, **20**(11): 1254−1259

184 He P, Huang W L, He T, Zhu Q L, Qiao Y, Li X L. Single shot text detector with regional attention. arXiv preprint arXiv: 709.0138, 2017

185 Karaoglu S, Tao R, van Gemert J C, Gevers T. Con-text: text detection for fine-grained object classification. *IEEE Transactions on Image Processing*, 2017, **26**(8): 3965−3980

186 Du Y N, Duan G Q, Ai H Z. Context-based text detection in natural scenes. In: Proceedings of the 19th IEEE International Conference on Image Processing. Orlando, FL, USA: IEEE, 2012. 1857−1860

187 Pan J Y, Chen Y, Anderson B, Berkhin P, Kanade T. Effectively leveraging visual context to detect texts in natural scenes. In: Proceedings of the 11th Asian Conference on Computer Vision. Daejeon, South Korea, 2012. 1−14

188 Lin L, Qu Y Y, Liao W M. Structure context clues for Chinese text detection. In: Proceedings of the International Conference on Internet Multimedia Computing and Service. Xiamen, China: ACM, 2014. 327

189 He D F, Yang X, Huang W Y, Zhou Z H, Kifer D, Giles C L. Aggregating local context for accurate scene text detection. In: Proceedings of the Asian Conference on Computer Vision. Cham, Switzerland: Springer, 2016. 280−296

190 Hauptmann A G, Witbrock M J. Informedia: news-on-demand multimedia information acquisition and retrieval. *Intelligent Multimedia Information Retrieval*. Cambridge, MA: AAAI Press, 1997.

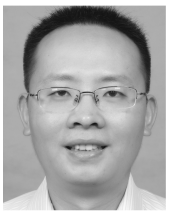
191 Smith J R, Chang S F. VisualSEEk: a fully automated content-based image query system. In: Proceedings of the 4th ACM International Conference on Multimedia. Boston, Massachusetts, USA: ACM, 1996. 87−98

192 Bai X, Yang M K, Lyu P, Xu Y C, Luo J B. Integrating scene text and visual appearance for fine-grained image classification. arXiv preprint arXiv: 1704.04613, 2017

193 Shi B G, Yao C, Zhang C Q, Guo X W, Huang F Y, Bai X. Automatic script identification in the wild. arXiv preprint arXiv: 1505.02982, 2015

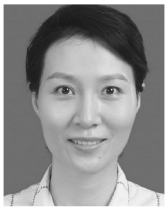
194 Bruce V, Young A. Understanding face recognition. *British Journal of Psychology*, 1986, **77**(3): 305−327

195 Kanwisher N, McDermott J, Chun M M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 1997, **17**(11): 4302−4311

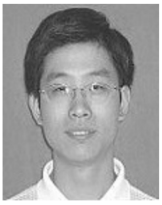


王润民 国防科技大学博士后. 湖南师范大学物理与信息科学学院讲师. 2015年获得华中科技大学博士学位. 主要研究方向为计算机视觉与模式识别.
E-mail: runminwang@hust.edu.cn
(**WANG Run-Min** Postdoctor at National University of Defense Technology. Lecturer at the School of Physics and Information Science, Hunan Normal University. He received his Ph. D. degree from the Huazhong University of Science and Technology in 2015. His research interest covers computer vision and pattern recognition.)

桑 农 华中科学技术大学自动化学院教授. 2000 年获得华中科技大学博士学位. 主要研究方向为计算机视觉与模式识别. E-mail: nsang@hust.edu.cn
(**SANG Nong** Professor at the School of Automation, Huazhong University of Science and Technology. He received his Ph. D. degree from Huazhong University of Science and Technology in 2000. His research interest covers computer vision and pattern recognition.)



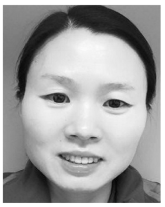
丁 丁 国防科技大学教研保障中心讲师. 2010 年获得国防科技大学博士学位. 主要研究方向为计算机视觉与模式识别.
E-mail: nudtdd@163.com
(**DING Ding** Lecturer at Teaching and Research Support Center, National University of Defense Technology. She received her Ph. D. degree from National University of Defense Technology in 2010. Her research interest covers computer vision and pattern recognition.)



陈 杰 芬兰奥卢大学电气与信息工程系资深教授. 2007 年获得哈尔滨工业大学博士学位. 主要研究方向为计算机视觉与模式识别.
E-mail: jie.chen@oulu.fi
(**CHEN Jie** Senior research scientist in the Department of Electrical and Information Engineering, University of Oulu, Finland. He received his Ph. D. degree from Harbin Institute of Technology in 2007. His research interest covers computer vision and pattern recognition.)

叶齐祥 中国科学院大学电子电气与通信工程学院教授, 2006 年获得中国科学院计算技术研究所博士学位. 主要研究方向为机器学习与视觉目标感知.
E-mail: qxye@ucas.ac.cn
(**YE Qi-Xiang** Professor at the University of the Chinese Academy of Sciences. He received his Ph. D. degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2006. His research interest covers visual object sensing and machine learning.)

高常鑫 华中科学技术大学自动化学院副教授, 2010 年获得华中科技大学博士学位. 主要研究方向为计算机视觉与模式识别.
E-mail: cgao@hust.edu.cn
(**GAO Chang-Xin** Associate professor at School of Automation, Huazhong University of Science and Technology. He received his Ph. D. degree from Huazhong University of Science and Technology in 2010. His research interest covers computer vision and pattern recognition.)



刘 丽 国防科技大学信息系统与管理学院副教授. 2012 年获得国防科技大学博士学位. 主要研究方向为图像理解, 计算机视觉, 模式识别. 本文通信作者.
E-mail: liuli_nudt@nudt.edu.cn
(**LIU Li** Associate professor at the College of Information System and Management, National University of Defense Technology. She received her Ph. D. degree from National University of Defense Technology in 2012. Her research interest covers image understanding, computer vision, and pattern recognition. Corresponding author of this paper.)