《文本检测方法综述》

目　　录

引言

研究背景与意义

文字是现代社会信息和思想交流的重要手段，包含着大量的高级语义信息。文本是文字的不同组合的，是交流语义信息的最重要媒介，生活中随处可见它的身影，例如书籍，街道名牌，商店标志，产品包装，餐厅菜单，视频字幕等。一直以来，场景文本的检测和识别都是机器视觉中的两个重要任务。自动检测和识别这些场景图片中的文本可以用于很多方面，例如实时文本翻译，盲人协助，广告推荐，自动驾驶，智能机器人和在线教育，情感分析，情报获取以及敏感信息过滤等。

文本检测和文本识别通常被串在一起组成一个 OCR 系统。文本检测通常作为文本识别任务的置步骤，比如字幕提取，身份证识别等 OCR 任务都是先对图片上文本进行检测，然后再对检测到的文本区域进行识别。在文本检测中，检测文本区域并标记其边界框。在文本识别中，从检测到的文本区域中解读文本信息。文本检测是端到端文本识别的重要步骤，没有文本检测，就很难从场景图像中精确识别文本字符。在 OCR 系统中，文本检测中文本定位的精度直接影响了 OCR 系统最终的识别结果好坏，因此文本检测是一项十分重要又具有挑战性的计算机视觉任务。2020 年 9 月 28 日，中国发布了国内首份智能文字识别（OCR）能力评测和应用白皮书，这从侧面反映了场景文本检测和识别任务的重要性。

图 1-1　CTW-1500 曲形文本数据集标注示例。

传统的文本检测方法一般采用手工特征提取的方式进行检测文本，比如SWT、MSER 等方法，然后采用模板匹配或模型训练的方法对检测到的文本进行识别。而现在的深度学习方法使用卷积神经网络代替手工提取特征方法进行文本检测，然后神经网络对检测到的文本进行识别。

近年来，人工智能和深度学习技术的发展，为计算机视觉领域的蓬勃发展注入了强有劲的动力。因此基于深度学习技术的文本检测方法也受到了广泛的关注，并且取得了良好的检测效果，也被广泛地应用于字幕提取，车牌号码识别、无人驾驶，广告推荐等领域。尽管这些方法在水平文本以及多方向场景文本上取得了很好的检测效果，但是由于这些方法大多从通用目标检测方法演化而来，受到文本表示的限制，从而只能检测四边形文本，无法适应任意形状文本检测的需求。水平场景文本文本检测方法和多向场景文本检测方法在检测这些扭曲的文本时，不仅会造成多个文本行之间边界框的重叠问题， 还会带来大量的背景噪，重叠和背景噪声都会直接影响到文本检测和后续识别等任务的性能。四边形文本只是一个文本的表示特例，而现实世界中，大多数文本为不规则形状（任意形状），如图 1-1，大多数柱状物体（瓶子、石桩等）、球形物体、折叠平面、硬币、标志、招牌上等物体上的文本行可能有横向、竖向、弯曲、旋转、扭曲等式样。

图 1-2　文本检测示意图，左图为词级标注，右图为行级标注。

任意形状文本检测的研究最早始于 2017 年中科院自动化所的刘成林老师团队提出的单词级别的曲形文本检测数据集 Total-Text 和华南理工大学金连文教授团队提出的文本行级别的曲形文本检测数据集 CTW-1500。从 2018 年开始，陆续就有相关的研究成果发表在国际顶级学术会议和顶级期刊上。曲形文本检测数据集的发布标志着文本检测领域由传统的四边形文本检测迈向了更加普遍、更加通用的文本检测。

研究现状与挑战

文本检测研究的现状

文本检测的研究大致分为三个阶段。第一阶段是在深度学习技术发展起来前，在这一时期文本检测主要是利用一些传统的方法，比如基于最大稳定极值区域（MSER）的方法和基于笔划宽度变换（SWT）。这些传统的文本检测方法在简单场景下取得了良好的检测结果，但是无法处理复杂的应用场景。第二阶段是深度学习早期，这一阶段的文本检测方法[1] [2] 3] 4] 5] [6] [7] [8]多是借鉴了基于深度学习的物体检测的一些方法[9] [10] [11] [12] [13]，因此这些文本检测方法只能检测水平文本和多方向文本。无论是水平文本还是多方向文本都是可以用四边形表示的文本。这些早期的基于深度学习的文本检测方法与传统方法相比，它们可以处理更加复杂场景下的文本检测问题。第三阶段是从 2017 年中科院自动化所的刘成林老师团队提出的单词级别的曲形文本检测数据 Total-Text 和南方科技大学金连文教授团队提出的文本行级别的曲形文本检测数据集 CTW-1500 开始的，这两个曲形文本检测数据集的公开发布，标志着文本检测正是迈入任意形状文本检测阶段。在这一阶段，文本检测方法不再依赖四边形表示，对任意形状都具有很好的适应性和鲁棒性。

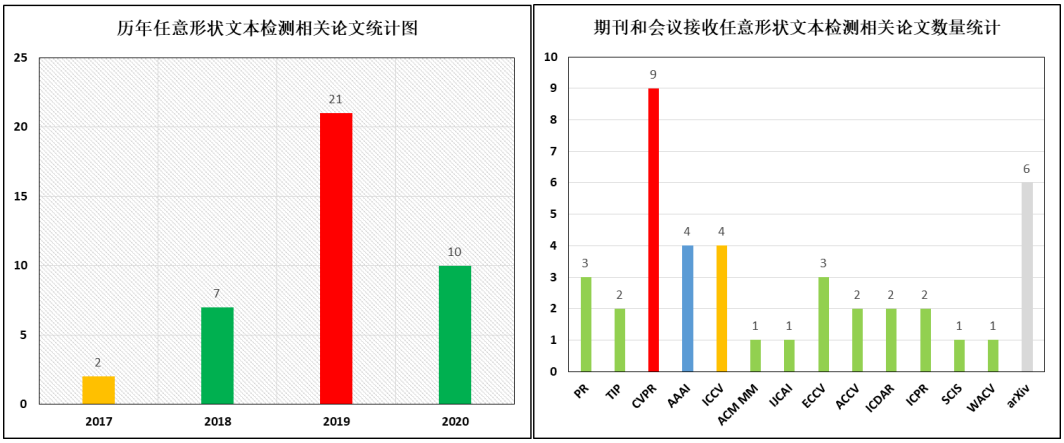

图 1-3　任意形状文本检测相关论文统计。左图为历年任意形状文本检测相关研究的论文数量的统计图，2017 年曲形文本检测数据集 Total-Text 和 CTW-1500 公开发布。右图各期刊会议自 2017 年以来接收和发表的有关任意形状文本检测的论文数量的统计。

从 2017 至今，已经有很多优秀的关于任意形状文本检测的研究成果[14] [15] [16] [17] [18] [19] [20]陆续发布在各大人工智能和计算机视觉的国际顶级会议上。如图 1-3 所示，本文统计了自 2017 年以来在国际学术会议和期刊上发表的有关任意形状文本检测研究的论文。从图 1-3 可以看出，任意形状文本自 2017 年以

来逐步成为文本检测研究的的主要方向，在 2019 年就有超过 20 篇的顶级学术会议和顶级期刊论文发表。在这些研究中有不少都是发表在国际计算机视觉以及人工智能顶会上，如 CVPR2019 就有 6 篇相关研究论文被接收和发表，AAA12019 和 ICCV2019 也各接收了 4 篇任意形状文本检测相关研究的论文。今年（2020 年），国际顶级学术会议接收的文本检测相关论文也基本都是和任意形状文本检测研究相关的论文。

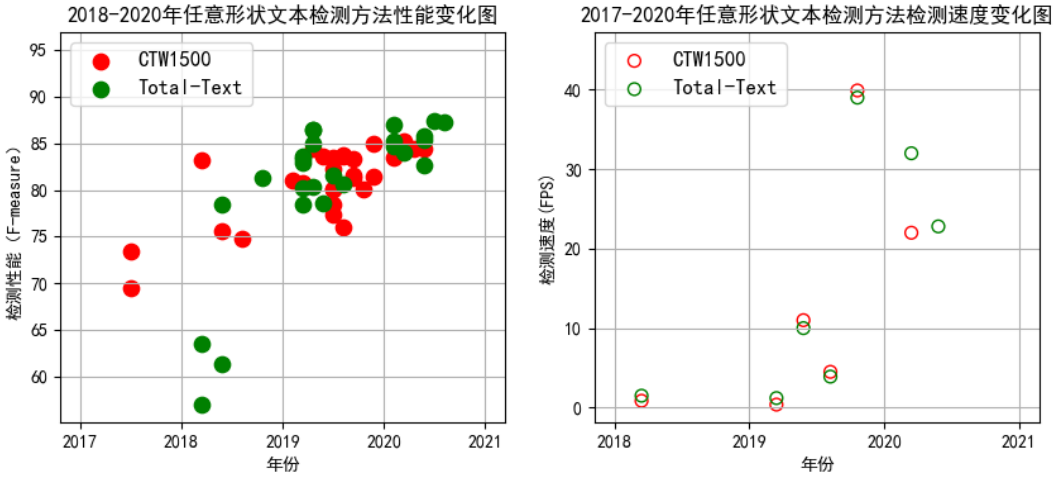

图 1-4　2017 年以来任意形状文本检测方法检测性能和检测速度的变化趋势。左图为 2018 年到 2020 年已有的一些任意文本检测方法的性能统计结果；右图为 2017 年到 2020 年已有的一些任意文本检测方法的检测速度的统计结果。注：不同的方法按照论文的发表的先后顺序或挂到 arXiv 网站上的时间顺序进行排序。

在短短 3 年的时间里，任意形状文本检测已经取代了了多方向的四边形文本检测成为文本检测研究领域的主要研究方向，并在检测精度和速度上取得了质的的飞跃。如图 1-4 所示，本文对现有的一些任意性形状文本检测方法的检测性能和检测速度做了一个统计。从图 1-4 中可以看到的是，自 2017 年以来，任意形状文本检测方法不论是检测性能还是检测速度都有很明显的提升；在检测性能上，任意形状文本检测模型 Boundary、DB 检测的精度基本达到了多方向的一些顶尖的四边形文本的检测模型的检测精度；在检测速度上，任意形状文本检测模型 DB、PAN 的检测速度基本到达了实时检测。

目前，基于深度学习的多方向文本检测技术（CTPN、EAST、Pixel-Link 等）已经比较成熟，并且在商业领域实现了广泛的应用，比如字幕提取，广告图片文字识别等领域。2020 年百度开源了一个基于 DB 和 CRNN 的端到端的 OCR 模

型，这意味着任意形状文本检测方法也在开始尝试实现商用。此外，最近越来越多的研究[21] [22] [23] [24] [25] [26] [27] [28] [29] [30]。开始关注任意形状文本检测和识别端到端一体化 OCR 模型的研究，这表明任意形状文本检测与识别在未来一段时间内仍然将是一个研究热点。

文本检测的挑战

文本检测也是一种目标检测，因此早期的一些文本检测算法[14] [15] [16] [17] [19] [36]多是借鉴了物体间算法的[9] [10] [24]一些思路，然后再根据文本自身的一些特点做了一些特别的设计。例如 RRPN 将 Faster-RCNN 中是的水平锚点（Anchor）改成旋转的锚点（Rotated Anchor）来解决文本实例多方向的问题。如图 1-5 所示，展示了一些具有挑战性的场景中的文本实例图像。

图 1-5　一些比较具有挑战性场景图片中的文本实例。

不同于一般的通用物体检测（Object Detection），文本实例通常有自己独有的一些特点。使用通用物体检测方法来直接解决文本检测问题，往往检测效果不甚理想，这是因为通用物体检测方法并不能很好地适应文本图像的一些特有的属性：

1) 文本实例大多数大多呈线性分布（带状分布），四边形文本和曲形文本有这个规律，因此文本实例的长宽比变化范围比较大（尺度的多样化）。而普通的目标检测中的物体通常长宽比接近于 1。

2) 文本实例通常不存在明显的闭合边缘轮廓，而普通物体通常都存在比较明显的闭合边缘轮廓。因此在实例分割数据集中会提供物体的像素级的标注，而文本检测数据集一般没有像素级的标注。

3) 一个文本实例中通常包含多个文字，并且文字之间是有间隔的。而文本检测通常需要检测方法能够准确地定位出整个单词或者句子的轮廓。如果检测方法设计的不好，就会很容出现将一个文本实例分成了多段情况，这与人们的预期效果是不一致的；同时这种情况也会影响到后续的一些过程，比如文本识别。

4) 文本实例形状和方向的多样化，如水平、垂直、倾斜、扭曲等。多变的形状也是曲形文本检测区别于四边形文本检测的一个重要特征，如何解决文本实例的形状表示问题也是任意形状文本检测方法研究的一个重点。

5) 文本实例的颜色、字体、语种的多样化；场景图片中的文字的颜色和字体变化是没有规律的，颜色和字体的取值空间都很大。此外，同一张图片可能存在不同语言的文本，这些不同语言的文本可能相似，也可能差异很大。

除了以上这些文本实例自身的特点是导致文本检测复杂度较高的原因外，外在的一些环境因素也会增大文本检测的复杂度。自然场景图像的背景是多样的、复杂的；例如文字可以出现在平面、曲面或折皱面上；文字区域附近有复杂的干扰纹理、或者非文字区域有近似文字的纹理，比如沙地、草丛、栅栏、砖墙。这些复杂的背景纹理会导致误检测的发生，另外文字区域还可能会产生变形(透视、仿射变换)、残缺、模糊等现象，也会导致误检测和漏检测的发生。

本篇文章是对近年来（2017-2021）一些重要的深度学习方向上的文本检测方法的综述。

文本检测传统方法

传统的文本检测方法通常包含：图像预处理，版面处理，图像切分，特征提取、匹配及模型训练，识别后处理。

1) 预处理：灰度化、二值化、倾斜检测与矫正，平滑、规范化
2) 版面处理：版面分析、版面理解、版面重构
3) 图像切分：行（列）切分和字切分
4) 特征提取与模型训练：特征提取及匹配、模型训练
5) 识别后处理：版面恢复和识别矫正

在传统的文本检测方法中比较有名的两种方法分别是：SWT(Stroke Width Transform) 和 MSER(Robust wide-baseline stereo from maximally stable extremal regions)。其中前者利用笔画宽度变化来检测文本，后者利用最大稳定极值区域来检测文本区域。

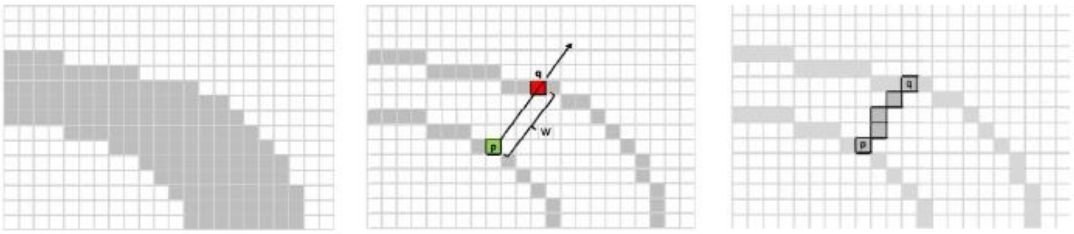

图 2-1 笔画宽度变换

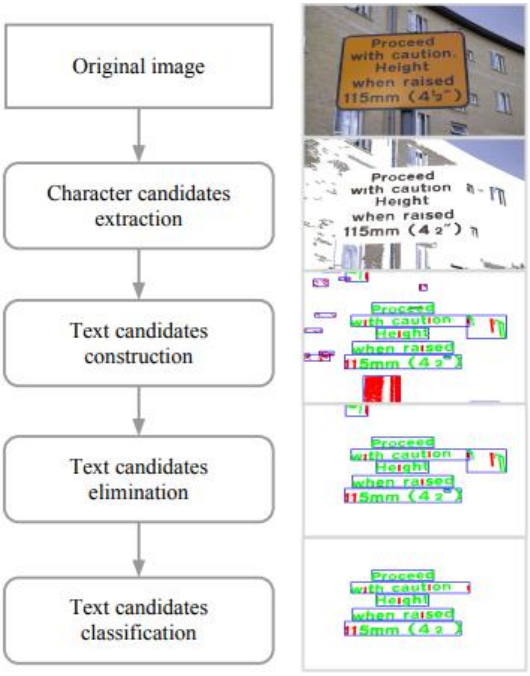

图 2-2 笔画宽度变换

笔画宽度变换（SWT）是一种文本区域检测算法，光学字符识别(OCR)一般在有噪声的图像上效果很差。SWT 就可以从有噪声的图像中提取文本，它通过提取出具有一致宽度的带状目标来实现。由此得到的图像消除了大部分噪

声，并保留了文本，从而得到更可靠的光学字符识别结果。由于文本具有一致宽度的线条，所以能得到很好的文本识别结果。SWT 用于检测图像中的笔画——这些是有限宽度的形状，由两条大致平行的边构成。笔画形状是在手写和打字场景中的主要元素，但在其他地方相对少见。大多数自然场景中的景物没有类似笔画的特征，即不具有一致的笔画宽度。SWT 从图像中的高对比度边缘上的一点开始，通过在垂直于边缘的方向上探索像素，我们可以找到另一条与之平行的边缘上的一点，由这两条边缘的点构成一个笔画横截面。通过连接很多宽度相似的笔画横截面，从而产生一个完整的笔画。如图 2-1 里面 p 和 q 两个像素相连得到笔画的一个横截面。字符可以被过滤、分组(成词)，从而得到文本区域，而不需要知道文本的语言或字体类型。无噪声的图像更容易被光学字符识别处理。

MSER 对灰度图像取阈值进行二值化处理，阈值从 0 到 255 依次进行递增，阈值的递增类似于分水岭算法中的水平面的上升，随着水平面的上升，有一些山谷和较矮的丘陵会被淹没，如果从天空往下看，则整个区域被分为陆地和水域两个部分，这类似于二值图像。图像中灰度值的不同就对应地势高低的不同，每个阈值都都会生成一个二值图。随着阈值的增加，首先会看到一个全白图像，然后出现小黑点，随着阈值的增加，黑色部分会逐渐增大， 这些黑色区域最终会融合，直到整个图像变成黑色。在得到的所有二值图像中，图像中的某些连通区域变化很小，甚至没有变化，则该区域就被称为最大稳定极值区域。MSER 是最大稳定极值区域：是对一幅灰度图像（灰度值为 0～255）取阈值进行二值化处理，阈值从 0 到 255 依次递增。阈值的递增类似于分水岭算法中的水面的上升，随着水面的上升，有一些较矮的丘陵会被淹没，如果从天空往下看，则大地分为陆地和水域两个部分，这类似于二值图像。在一幅含有文字的图像上，由于文字区域的灰度值是一致的，而且和文字周边像素的灰度值差别较大，因此在水平面（阈值）持续增长的一段时间内它们都不会被覆盖，直到阈值涨到文字本身的灰度值时才会被淹没，所以文字区域可以作为最大稳定极值区域。所以如果一个区域在给定的阈值范围内保持其形状和大小基本稳定不变，而不会与其他区域合并，该区域被认为是稳定的。

但是传统的文本检测方法存在很明显的缺陷，对文字形状变化（文字模糊、笔画粘连、断笔、黑白不均、油墨反透）的适应性和抗干扰性比较差。

文本检测深度方法

再过去的几年时间里，随着深度学习的繁荣，基于深度学习的场景文本检测方法也取得广泛的应用，并取得了很好地检测性能。通常，这些基于深度学习的文本检测方法方法可以分为五大类类：基于回归的文本检测方法，基于分割的文本检测方法，基于组件连接的文本检测方法，基于关键点预测的文本检测方法，以及基于检测和识别一体的文本检测方法。

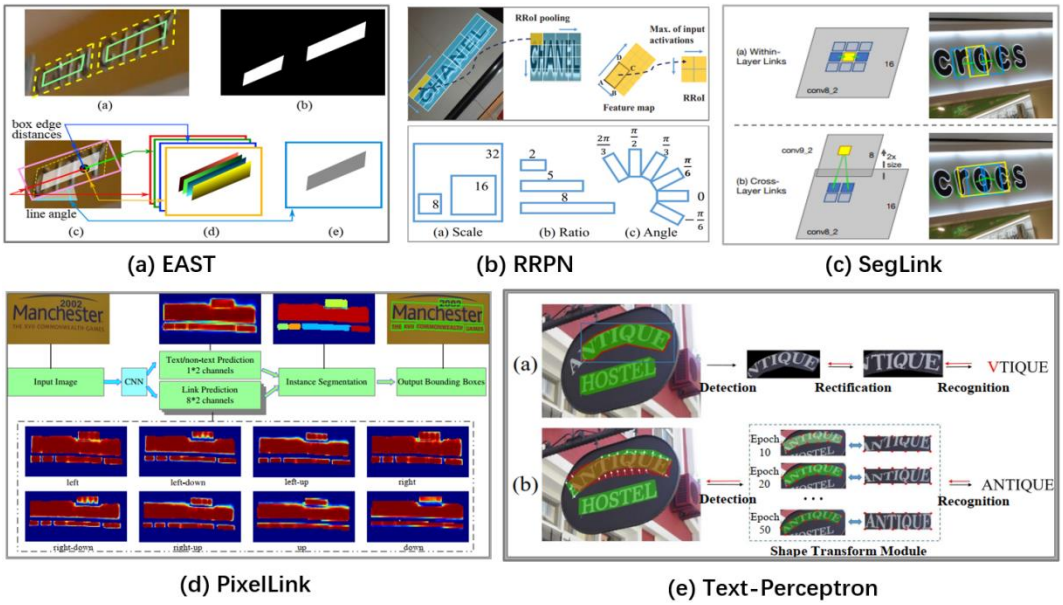

图 3-1 (a) EAST 直接回归；（b）RRPN 锚点回归；（c）SegLink 组件连接；（d）PixelLink 文本实例分割；（e）Text-Perceptorn 端到端模型。

目前，任意形状文本检测多以基于分割和基于组件连接的文本检测方法为主，也有少量方法通过关键点来解决任意形状文本的检测问题。基于分割和组件这两类方法通常需要对像素或者文本组件进行聚类才能得到完整的文本实例。因此在这两类方法中常常还需要对像素之间的关系或者文本组件之间的关系进行有效地学习和推理。

基于回归的文本检测方法

基于回归文本检测方法 CTPN，EAST，TextBoxes，TextBoxes++，RRPN，DDR，RRD 多是受基于锚点（Anchor）的物体检测方法 SSD，Faster R-CNN 和基于非锚点（Anchor Free）的物体检测方法 DenseBox，UnitBox 等物体检测方法

的启发，将文本实例作为一个对象，通过估计检测框到锚点的偏移或者直接估计文本实例像素到文本边界的距离来生成文本实例的边界框作为检测结果。一般来说基于回归文本检测方法的文本检测方法可以大致分为两类：基于锚点（Anchor）回归的方法和基于直接回归（Anchor Free）的方法。

基于锚点回归（Anchor）的方法

基于锚点回归的文本检测方法大多都是基于何凯明等人在 2015 年提出的物体检测方法 Faster R-CNN 改进得到的。其中比较有名的基于锚点回归的文本检测方法有水平文本检测方法 CTPN、TextBoxes 等，多方向文本检测方法 TextBoxes++、RRPN 等。

CTPN 针对文本自身的特点对 Faster R-CNN 中的锚点进行了重新设计。众所周知的是，文本检测的难点在于文本的长度是不固定，文本可以很长，也可以很短．如果采用物体检测的方法，如何生成好的文本提案（Text Proposal）会为一个棘手的问题。针对这个问题，CTPN 的作者提出了竖直锚点（Vertical Anchor）的方法。具体的做法是只预测文本在竖直方向上的位置，水平方向的位置不做预测。与 Faster R-CNN 中的锚点不同，竖直锚点的宽度都是固定的，论文中的大小是 16 个像素，而高度则从 11 像素到 273 像素(每次除以 0.7)变化，总共 10 个锚点。然后利用卷积神经网络（CNN）用来提取深度特征，循环神经网络（RNN）用来序列的特征识别，并且预测一系列的文本提案。这些密集的文本提案在经过 NMS 算法处理之后，利用预先定义的规则将这些小文本段连接起来，得到文本实例。由于基于规则的连接方式缺乏自适应性，CTPN 比较适合检测行级的水平的文本，不适合对单词级别的文本进行检测，因此 CTPN 经常被用来做视频中字幕的文本检测任务。

在场景文字检测中一个最常见的问题便是倾斜文本的检测，通用的物体检测算法只能检测水平的目标，所以物体检测算法无法直接应用于倾斜文本的检测。针对这个问题，RRPN 的作者提出了旋转锚点（R-Anchor），并依据锚点的角度特征重新设计了 IoU 的计算方法，提出了影响深远的旋转 NMS 算法（Rotate NMS）以及旋转池化（RROI）算法等算法。R-Anchor 的锚点由 3 个尺寸，3 个比例以及 6 个角度组成：3 个尺寸分别是 8，16，32；3 个比例分别是 1:2，1:5，1:8 ；6 个角度分别是$-\frac{\pi}{6},0,\frac{\pi}{6},\frac{\pi}{3},\frac{\pi}{2},\frac{2\pi}{3}$。 但是引入旋转角度使得锚点的数量成倍增加，带来了巨大的计算开销，因此 RRPN 的检测速度是很慢的。

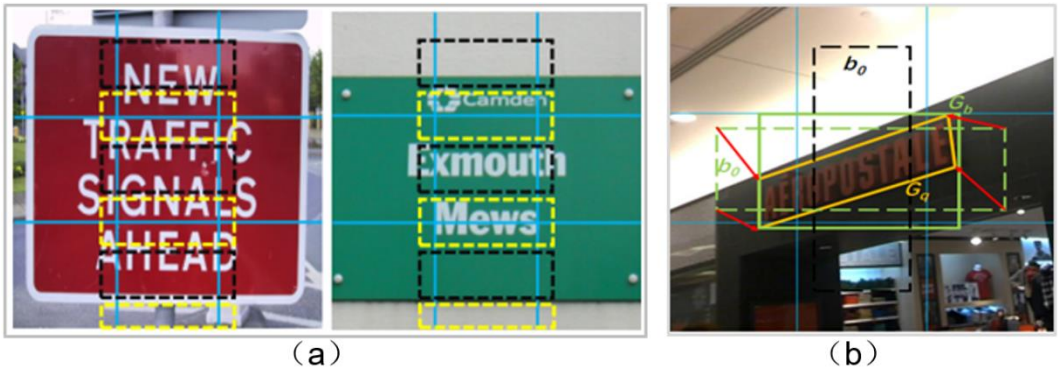

图 3-2　左图为锚点的偏移的示意图；右图为 TextBoxes++中文本框示意图。

TextBoxes 借鉴 SSD 的网络的网络设计，并做了一些改进将其用于文字检测。和 SSD 相比，TextBoxes 依据文本实例的成带状分布这一特点，修改了锚点默认的长宽比，分别为[1 2 3 5 7 10]。这么做的目的是为了适应文本长度比较长，宽度比较短的特性，也就是在 TextBoxes 中的锚点是长条形。此外 TextBoxes 将分类的卷积核的大小修改为 1×5 ，而 SSD 中卷积核的大小为 3×3，这样做的目的是更适合文本行的检测，避免引入非文本噪声。为了提高文本行检测的效果，TextBoxes 增加了文本识别模块。因为采用了不一样尺寸的锚点，这些尺寸都是细长形的，这样可能导致锚点在水平方向密集在垂直方向上稀疏，从而导致检测不准确。为了解决上述问题，TextBoxes 给每个锚点加上了垂直偏移。和 SDD 一样，TextBoxes 也只能检测水平目标检测，无法检测倾斜的文本行。

为了进一步解决倾斜文本的检测问题，TextBoxes 的作者对 TextBoxes 进行了扩展和增强，提出了 TextBoxes 的增强版 TextBoxes++。因此增强版的 TextBoxes++可以检测多角度的文本。类似于 TextBoxes，TextBoxes++也给每个锚点加上垂直偏移。

基于直接回归（Anchor Free）的方法

基于锚点回归的文本检测方法大多计算复杂度很高，所以就有一些学者提出了一些非锚点（Anchor Free）设计的文本检测方法，这些方法也被称为基于直接回归的文本检测方法。在这些方法中，有一些方法受到了广泛的关注和引，如 EAST，DDR 等。

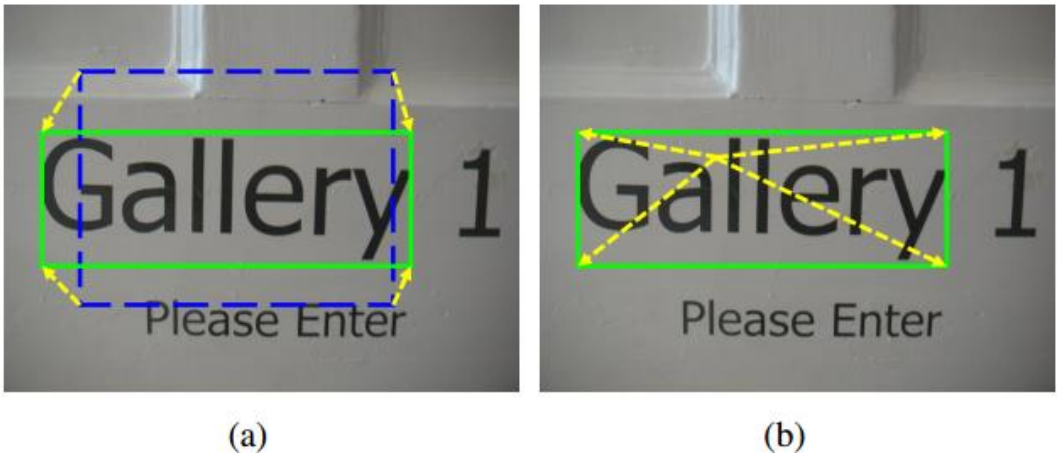

图 3-3　左图为基于 Anchor 类的方法，它们通过回归到锚点的偏移来获取文本的外接四边形；右图为基于直接回归的文本检测方法，它们通过直接预测文本像素到文本外接四边四个顶点的偏移或者到四个边界的距离来确定文本的外接四边形。

DDR 首次提出直接回归的概念，在特征金字塔（FPN）网络结构中直接学习四个点相对于中心点（特征图上的某个点）的偏移量，并用 Scale & shift 方案来缩小要学习目标的值范围。本文提出的方法在定位场景图片中的文本上非常有效。以 ICDAR2015 场景文本数据集基准进行测试，该方法实现了 81%的 F 值，并且明显优于以往的方法。

EAST 的作者认为传统的文本检测方法和一些基于深度学习的文本检测方法，大多是多阶段的，在训练时需要对多个阶段进行调优，这将会影响最终的模型效果，而且非常耗时。所以，论文的作者提出了端到端的文本检测方法，消除中间多个阶段 (如候选区域聚合，文本分词，后处理等)，直接预测文本行。EAST 通过将文字区域进行适当地收缩来解决密集文本实例之间的粘连问题，通过预测文本每个文本像素到上下左右边界的四个距离和文本的旋转角度来定位文本的外接四边形，整个模型结构十分简单，同时速度较快。 但是由于像素到文本边界的距离取值范围比较大，而且分布不平衡，所以 EAST 模型会出现对长文本和大文本检测不精确的问题。

基于分割的文本检测方法

基于分割的文本检测方法多是受语义分割和实例分割类的方法的启发，将每个文本当成一个目标实例，然后将其利用一些方法将其与背景区分开来。 由于分割类的方法都是基于像素级的预测，并不依赖文本的形状这个先验知识。因此

和基于回归类的方法相比，基于分割的文本检测方法对文本的形状具有更好的适应性，对曲形文本检测也更加鲁棒。所以近些年，基于分割的文本检测方法受到了极大的关注。其中比较有名的文本检测方法有 PixelLink, LSAE, CSE, TextField, MSR, DB, PAN, PSENet。

PixelLink 是比较早将实例分割的思路应用到文本检测上，也因此该方法并没有在论文里报告它在曲形文本检测数据集上的性能，但该方法应该是可以直接用来处理曲形文本的检测问题。PixelLink 主要基于 CNN 网络，分别做文本，非文本分类预测和像素的 8 个方向是否连接预测这 2 个任务。然后通过求取文本连通域的最小外接矩形获取四边形的文本边界。最后再通过噪声滤除和并查集合并出最终的文本框。

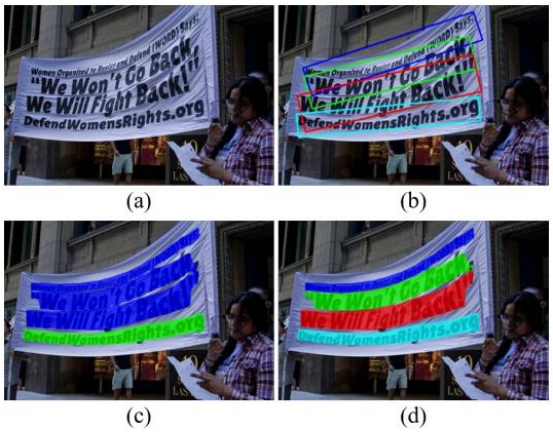

图 3-4 PSENet 中提出的两大挑战。（a）是原图；（b）是基于四边形或旋转矩形的检测结果；（c）基于像素分割的方法没能能区分非常邻近的文本实例的结果。

PSENet 是旷视提出的一种任意形状文本检测算法，论文指出任意形状文本检测存在两大挑战；第一个挑战是现有的文本检测是基于四边形或旋转矩形的文本检测方法很难将任意形状的文本(特别是形状文本)进行包闭操作；第二个挑战是大多数基于像素分割的方法不能很好地区分非常邻近的文本实例。针对这两个挑战，该论文提出了基于语义分割的单文本实例的预测方法，它采用了前向渐进式尺度扩展的方法用来区分邻近的文本实例，可用于检测任意方向的文本。PSENet 沿用了特征金字塔网络结构（简称 FPN），并在此基础上增加了特征融合和渐进式尺度扩展的方式来实现自然场景中文本行的检测。

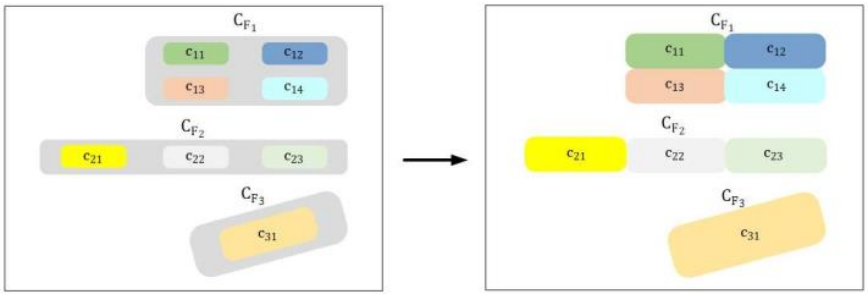

图 3-5　LSAE 中的像素聚类过程。

LSAE 的主要思想是将文本检测当作一种实例分割，采用了基于分割框架来进行检测。具体的做法是将每个文本行看成一个连通区域，为了更好地区分不同文本实例（即距离很近的文本或者是很大很长的文本），提出了将图像像素映射到嵌入特征空间中，在该空间中，属于同一文本实例的像素会更加接近彼此，反之不同文本实例的像素将会远离彼此。特征提取主干网络采用的是 ResNet-50，接着使用两个对称的特征金字塔结构进行特征融合，一个用于后续的嵌入分支（Embedding branch），另一个用于后续的分割分支（文本行的前景图，包括全文本行前景图和向内收缩后的文本行的前景图）。通过权重共享，使得两个任务优势互补。网络输出包括嵌入特征图和文本行的前景掩膜图，然后经过后处理得到最终的预测文本。

CSE 将文本检测表示成一个条件区域扩展问题，它首先在文本区域内初始化种子，然后通过区域扩展逐步检索目标对象；通过参数化的条件空间扩展机制对种子与对象其余部分之间的空间依赖性进行建模，然后有选择地提取种子所指示的文本区域，具有较高的区域精度。该方法可以充当第二阶段的文本提取器，可以无缝集成到现有的对象检测工作流中，具有种子位置任意性和空间选择性强的特点，减少了与以前的检测器的耦合，从而提供了灵活、可靠的边界预测。

TextField 针对任意形状检测，采用实例分割的思路，提出一种对于分割点的新的表示方法 TextField，旨在解决文本区域的粘连问题。TextField 是一个二维的向量场， 它表示每个文本像素点到离自己最近的边界点的向量。用一个 VGG 加 FPN 结构的网络来学习 TextField 的方向分布图和模长分布图，然后这两张图上做关于超像素、合并、形态学等后处理来得到文本实例。

MSR 通过网络来回归文字的边界像素点来得到文本区域。在多尺寸网络中利用 FPN 的上采样把多个不同尺寸得到的结果进行融合，然后在融合的特征上进行文本分类和边界框回归， 回归部分直接预测点与最近的边界点的坐标 x 的偏移和坐标 y 的偏移，思路清晰且易实现。

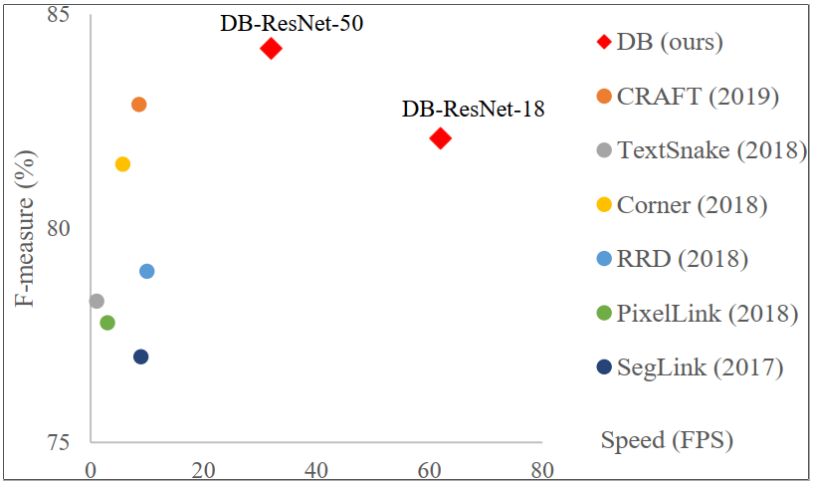

图 3-6　MSRA-TD500 数据集上几种场景文本检测方法的准确性和速度的比较。

DB 是华中科技大学白老师组的研究成果，这篇提出的方法虽然很简单，但是在现有 5 个文本检测数据上（包括曲形文本检测数据集）在检测精度与速度上均表现不凡。基于分割的方法其中关键的步骤是其后处理部分，这步中将分割的结果转换为文本框或是文本区域。DB 也是基于分割的，但是通过提出可微分的二值化模型（Differenttiable Binarization Module）来简化分割后处理步骤（加了一个变的预测），并且可以设定自适应阈值来提升网络检测性能。

PAN 是旷视研究人员发表在 ICCV2019 上的一个工作，论文中介绍的算法是 PSENet（发表在 CVPR2019 上）的改进版。文章中采用 ResNet-18 作为主干网络，试图改善文本检测的速度，但是该轻量级的主干网络感受野较小，表达能力也不足，因此提出了特征金字塔增强模块（FPEM）和特征融合模块（FFM）。FPEM 以较小的开销融合不同尺度的信息，且可通过级联多个 FPEM 补偿 ResNet-18 的深度。而 FFM 用来融合 FPEM 得到的特征。PAN 通过得到的相似性向量对文本像素进行聚类。该方法在文本检测速度和精度上取得了很好的平衡。

基于组件连接的文本检测方法

基于组件连接的文本检测通常先检测独立的文本组件或者字符，然后经过连接或者归并等后处理过程生成最终的检测结果。由于基于组件的文本检测方法，能够很好地适应曲形文本的检测问题，因此它也成为了目前任意形状文本检测方法中一大主流方法。其中比较有名的方法有 CTPN，SegLink, TextSnake, CRAFT，SegLink++等。除此之外还有一些端到端的文本检测方法的检测部分是基于文本组件连接的检测方法，例如 CharNet,　TextDragon 以及 CRAFTS。

CTPN 的文本组件是通过基于锚点回归来预测的，在本文中已经将其分类为基于锚点回归的方法， 其实 CTPN 也可以被分类为基于组件连接的方法。CTPN 将文本实例划分成许多等宽的文本组件，然后利用人工定义的连接规则连接这些组件。由于人工定义的连接关系缺乏灵活性，所以 CTPN 只能检测一些水平的文本，无法解决多方向和曲形文本的检测问题。

SegLink 模型先将每个单词切割成更易检测的有方向的局部片段(segment)和可以指明相邻局部片段是否连接的 Link。然后在基于 SSD 的改进版网络结构(全卷积网络结果)上同时预测不同尺度的 Segments 和 Links。然后通过规则将所有的片段进行连接，得到最终的文本行，这样做的好处是可以检测任意长度文本行。

图 3-7　SegLink++中的组件归并及轮廓提取。

SegLink++的作者通过改进原先的 SegLink 模型得到了一个新的模型， 并把它命名为 SegLink++。新的方法（SegLink++）能够很好的解决文本检测任务中类似于商品信息等图片具有密集且任意形状的文本框的问题。和 SegLink 相比，SegLink++的作者引入了两种新的 Link 关系，一种是吸引关系（Attractive link），一种是排斥关系（Repulsive link）。这两种关系，一种是将属于相同文本区域的文本片段相连，一种是将属于不同文本区域的文本片段相拒。除此之外，作者还提出了一种实例感知损失（Instance-aware loss），将后处理加入到优化中，从而进一步提升文本检测的性能。

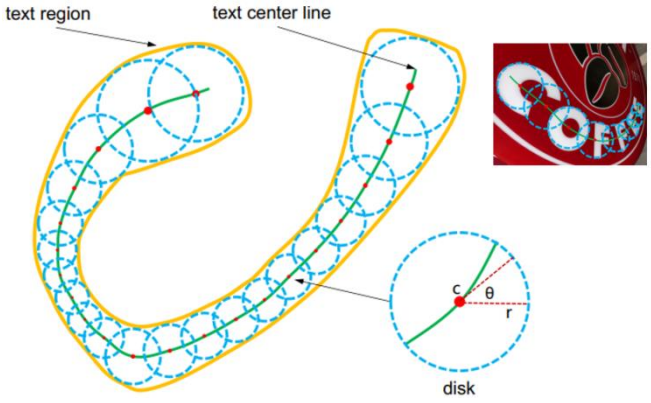

图 3-8　TextSnake 中的圆盘。

TextSnake 提出一个新的文本表示方法，用一个有序的圆盘序列来表示文字，先用卷积神经网络检测文本区域、文本中心线、以及每个点的圆盘半径、方向，然后利用文本区域掩码和中心线掩码得到文字的实例分割区域。在每个文字实例上，交替进行点中心化和点扩展，得到文本中心点序列。最后结合圆盘半径，得到文本区域的 TextSnake 表示并进行组合得到最终的文本区域。

CRAFT 是 Youngmin Baek 等人在 2019 年 CVPR 上提出基于组件连接的文本检测模型，Youngmin Baek 等人认为对与曲形文本或者文本检测来说，比较好的方式就是先检测出单个字符，然后通过字符链接形成单词或者句子。但是目前的文本检测公开数据集最多只标注到单词级别，缺乏字符级别的标注。因此 Youngmin Baek 在这篇文章中提出了一个基于弱监督方法的字符级别的文本标注工具，利用这个弱监督的字符集文本标注工具将现有的文本检测数据集转换成字符级标注。CRAFT 的第二个贡献点在于，对于训练标签生成，采用与以往二值化分割图的生成方式不同，用高斯热度图来生成文本区域得分（Region Score）和连接关系得分（Affinity Score）。其中文本区域得分用来检测每个字符区域，连接关系得分用来判断字符区域之间的连接关系。

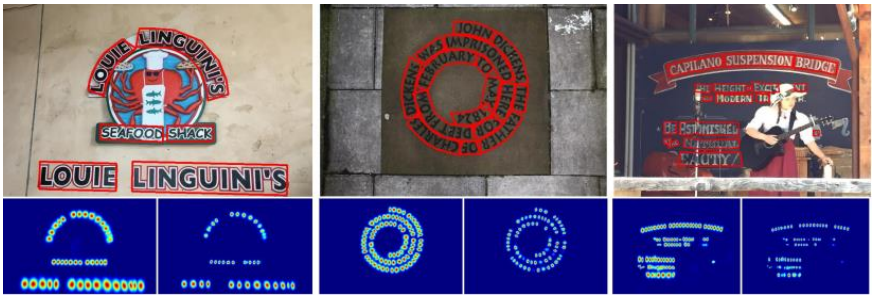

图 3-9　CRAFT 检测结果图。

基于关键点预测的文本检测方法

基于关键点预测的方法通过预测文本实例轮廓上的关键点来检测任意形状的文本。但是这种方法存在很明显的确定，那就是关键点的定义是个很大的难题。因此这种方法在文本检测中用的比较少，目前的方法有 ATRR, CTD, SLPR 等。

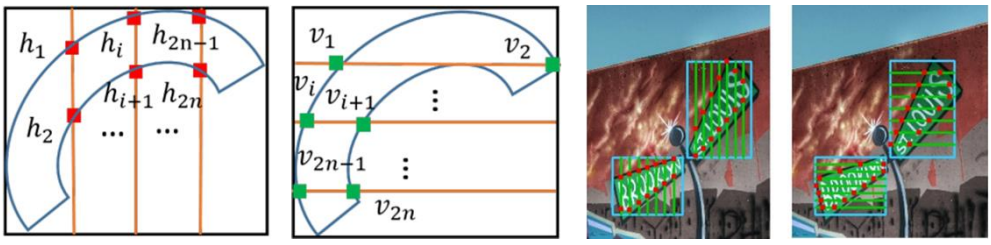

图 3-10　SLPR 关键点预测示意图。

SLPR 针对任意形状检测，采用物体检测思路，基于 R-FCN 框架，增加了沿 x 和 y 轴均匀划线与多边形交点的纵和横坐标的回归（14 个点，仅回归 x 或 y 坐标，水平 7 个维度和竖直 7 个维度的各 2 个点的回归），最后把点串起来得到多边形。

CTD 和 SLPR 的检测思路基本一致，CTD 基于 Faster R-CNN 进行修改，除了学习文本和非文本分类，多边形的边界框的回归，还增加了 14 个点的回归（14 个点，回归 x 和 y 坐标）。

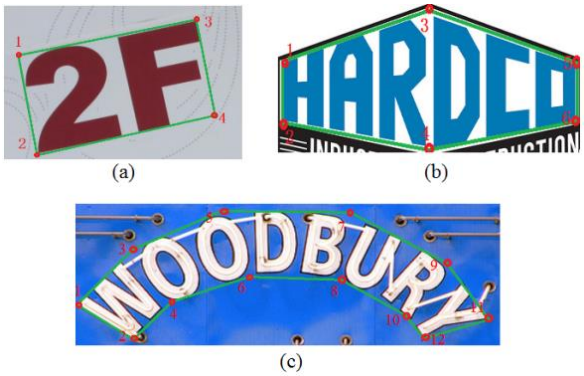

图 3-11　ATRR 中关键点的定义。

ATRR 文中的主要思想是给定一张输入图像，首先由一个文本候选区域提取网络(Text-RPN)来提取文本，然后通过一个修正网络(Refinement network)来验证和改进这些候选区域，在这个过程中，一个基于自适应文本区域表示的循环神经网络（RNN）结构被用于文本区域修正， 每次预测一对边界点直到没有新的点。

基于检测和识别一体的文本检测方法

基于检测和识别一体的文本检测方法，通常将用一个端到端可训练的模型同时解决文本检测和文本识别两个问题。通常人们认为利用文本识别的结果和文本识别模型学到的语义关系，可以帮助人们获得更好地检测结果，大量的的实验也证实了这一观点。端到到的文本，检测方法拥有很久远的研究历史，早期的端到端的文本检测方法有 TextBoxes T，extBoxes ++, FOST, 不过这些早期的方法都是针对四边形文本设计的。近年来，有不少针对曲形的端到端的文本检测方法被提出， 例如 Mask TextSpotter, Boundary, TextDragon, CharNet, CRAFTS,Text Perceptron,, ABCNet 等。越来越多的检测和识别一体的任意形状文本检测方法的出现，也在一定程度上反映出了未来任意形状文本检测的发展趋势。

Mask TextSpotter 是华中科技大学白翔老师组的研究，到目前为止已经有三

个版本了，最新的 Mask TextSpotter v3 也在不就前放到了 arXive 网站上。这篇文章主要受到物体检测方法 Mask R-CNN 的启发，引入语义分割的思想进行端到端的训练，提出了一种可以对各种曲面文字进行检测的框架。并具体对 10 个数字加 26 个字母共 36 个字母的识别能力。

图 3-12　两种文本区域表示方法的说明。

Boundary 是华中科技大学白翔老师课题组发表在 AAAI2020 上的研究。该文提出了用边界点来表示任意形状文本的方法，解决了自然场景图像中任意形状文本的端到端识别问题。如图 3-12 所示，现有方法用外接四边形框来表示文本边界（图 3-12，（a）），通过 RoI-Align 来提取四边形内的特征（图 3-12，（b）），这样会提取出大量的背景噪声，影响识别网络。作者认为边界点能够描述精准的文本形状，消除背景噪声所带来的影响；通过边界点可以将任意形状的文本矫正为水平文本，有利于识别网络，识别分支则通过反向传播来进一步优化边界点的检测。该方法在 Total-Text 数据上取得了已发表论文中的最好检测结果。

TextDragon 的命名受到了 TextSnake 名称的启发。在 TextDragon 中，文本检测器设计为使用一系列四边形来描述文本的形状，这些四边形可以处理任意形状的文本。为了从特征图中提取任意文本区域，作者提出了一个新的可微分运算符 RoISlide，它是连接任意形状的文本检测和识别的关键。基于通过 RoISlide 提取的功能，引入了基于 CNN 和 CTC 的文本识别器，使框架免于依赖字符的标注。TextDragon 在两个曲形文本数据集 CTW-1500 和 Total-Text 上取得了当年最先进的性能。

CharNet 是一个融合文本检测和文本识别任务的端到端的文本处理算法。该算法解决了双阶段文本检测与识别算法中 ROI Pooling 层对识别精度的影响，并且可以直接输出图片中文本和字符所在的位置和相应的字符标签，是一个单阶段检测算法。此外，改论文还提出了一个迭代字符检测算法用来更好的将在合成数据中训练好的模型应用到真实场景中的数据中，这些方法使得 CharNet 可以很好的处理多方向和曲线文本得检测和识别，并在 3 个标准的文本检测和识别数据集上面获得较大的性能提升。

Text Perceptron 通过基于分割的文字检测方法得到文本的轮廓点，提出形状矫正模型对弯曲文本进行矫正之后在输入识别模型，得到识别结果。整个过程是端到端的，识别模型的梯度可以回传检测模型，利用检测的梯度和识别梯度共同优化识别模型与检测模型。文中的算法包含了平时较为常见的插件，虽然最终的性能不是特别好，但是思路值得借鉴和学习。

ABCNet 的最大亮点在于使用了贝塞尔曲线来构成文本框，而不是使用矩形文本框，这样一来模型可以预测任意形状的文本，而且并不会增加太多的参数。此外，本文设计了一个新颖的 BezierAlign 层，用于精确地提取任意形状文本实例的卷积特征。通过 BezierAlign 层将识别分支连接到整体结构之中，是的整个模型成为一个端到端的文本检测和识别一体的模型。

CRAFTS 在文本检测方法 CRAFT 加入了文本识别分支，组成了一个端到端的文本检测模型，该文章发表在 ECCV2020 上。CRAFTS 利用字符区域注意力来定位文本中每个字符的位置进行文本实例定位和辅助文本字符识别。区域注意力有助于识别模型更好地关注字符中心点，并且识别损失向检测器模块的传播会增强字符区域的定位。CRAFTS 在端到端的检测和识别任务上性能优于 CharNet、Mask TextSpotter 等网络。

经典文本检测方法总结

论文题目	*模型*	*方法*	*时间*	*检测文本类别*	*备注*
Tian et al.[1]	CTPN	回归	ECCV 2016	水平文本	
Liao et al. [8]	TextBoxes	回归	CVPR 2017	水平文本	
Shi et al.[2]	SegLink	回归	CVPR 2017	水平+弯曲文本	
Zhou et al.[3]	EAST	回归	CVPR 2017	水平+旋转文本	
Liao et al. [9]	TextBoxes++	回归	IEEE 2018	水平+旋转文本	
Zhu et al. [10]	SLPR	回归	arXiv 2018	水平+弯曲+不规则文本	
Lyu et al. [11]		回归+分割	CVPR 2018	水平+旋转文本	
Liao et al. [12]	RRD	回归	CVPR 2018	水平+旋转文本	
Yang et al. [13]	IncepText	回归+分割	IJCAI 2018	水平+旋转文本	
Yue et al. [14]	Guided CNN	回归+分割	BMVC 2018		
Liu et al. [15]	MCN	分割	CVPR 2018	水平+旋转文本	
Long et al. [16]	TextSnake	回归	ECCV 2018	水平+弯曲+不规则文本	
Chu et al. [17]	Border	回归	ECCV 2018	水平+旋转文本	
Wang et al.[20]	ITN	回归	CVPR 2018	水平+旋转文本	
Zhao et al. [24]	Elite Loss	分割	2018	水平+旋转文本	
Liu et al.[22]	CSE	回归	CVPR 2019	水平+弯曲+不规则文本	
Wang et al.[4]	PSENet	分割	CVPR 2019	水平+弯曲+不规则文本	
Tian et al.[5]	LSAE	分割	CVPR 2019	水平+弯曲+不规则文本	
Wang et al.[6]	ATRR	回归	CVPR 2019	水平+弯曲+不规则文本	
Zhang et al. [26]	LOMO	回归+分割	CVPR 2019	水平+弯曲+不规则文本	
Baek et al.[7]	CRAFT	分割	CVPR 2019	水平+弯曲+不规则文本	检测+识别
Wang et al.[18]	PAN	分割	ICCV 2019	水平+弯曲+不规则文本	
Liao et al. [27]	MaskTextSpotter	分割	TPAMI2019	水平+弯曲+不规则文本	
Liao et al. [23]	DBNet	分割	AAAI 2019	水平+弯曲+不规则文本	
Liu et al. [19]	SBD	回归	arXiv 2019	水平+旋转文本	检测+识别
Zheng et al. [21]	SR-Deeptext	分割	PR 2019	水平+旋转文本	
Liu et al. [25]	ABCNet	回归	CVPR 2020	水平+旋转+不规则文本	
Zhang et al.[28]	DRRG		CVPR 2020	水平+旋转+不规则文本	
Wang et al.[29]	ContourNet		CVPR2020	水平+旋转+不规则文本	
Yiqin et al.[30]	FCENet		CVPR2021		
He et al.[31]	MOST		CVPR2021		
Zhu et al.[32]	TextMountain		PR2021		
Amanpreet et al.[33]	TextOCR		CVPR2021		
Hao et al. [34]	STR-TDSL		CVPR2021		
Zhang et al.[35]	TextBPN		ICCV2021		
Dai et al.[36]	PCR		CVPR2021		

参考文献

[1] Tian Z, Huang W, He T, et al. Detecting text in natural image with connectionist text proposal network. European conference on computer vision(ECCV), 2016: 56-72
[2] Shi B, Bai X, Belongie S. Detecting Oriented Text in Natural Images by Linking Segments. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 3482-3490
[3] Zhou X, Yao C, Wen H, et al. EAST: an efficient and accurate scene text detector. CVPR, 2017: 2642-2651.
[4] Wenhai W, Enze X, et al. Shape Robust Text Detection with Progressive Scale Expansion Network. In CVPR 2019.
[5] Zhuotao Tian, Michelle Shu, et al, Learning Shape-Aware Embedding for Scene Text Detection. In CVPR, 2019.
[6] Xiaobing Wang, Yingying Jiang, et al, Arbitrary Shape Scene Text Detection with Adaptive Text Region Representation. In CVPR, 2019.
[7] Youngmin Baek, Bado Lee, et al. Character Region Awareness for Text Detection. In CVPR 2019.
[8] Liao M, Shi B, Bai X, et al. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. AAAI. 2017: 4161-4167.
[9] Liao M, Shi B , Bai X. TextBoxes++: A Single-Shot Oriented Scene Text Detector. IEEE Transactions on Image Processing, 2018, 27(8):3676-3690.
[10] Zhu Y, Du J. Sliding Line Point Regression for Shape Robust Scene Text Detection. arXiv preprint arXiv:1801.09969, 2018.
[11] Pengyuan Lyu, Cong Yao, Wenhao Wu et al. Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation. In CVPR 2018.
[12] Minghui L, Zhen Z, Baoguang S. Rotation-Sensitive Regression for Oriented Scene Text Detection. In CVPR 2018.
[13] Qiangpeng Yang, Mengli Cheng et al. IncepText: A New Inception-Text Module with Deformable PSROI Pooling for Multi-Oriented Scene Text Detection. In IJCAI 2018.
[14] Xiaoyu Yue et al. Boosting up Scene Text Detectors with Guided CNN. In BMVC 2018.
[15] Zichuan Liu, Guosheng Lin, Sheng Yang et al. Learning Markov Clustering Networks for Scene Text Detection. In CVPR 2018.
[16] Long, Shangbang and Ruan, Jiaqiang, et al. TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes. In ECCV, 2018.
[17] Chuhui Xue et al. Accurate Scene Text Detection through Border Semantics Awareness and Bootstrapping. In ECCV 2018.
[18] Wenhai Wang et al. Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network. In ICCV 2019
[19] Yuliang Liu et al. Exploring the Capacity of Sequential-free Box Discretization Network for Omnidirectional Scene Text Detection

[20] Fangfang Wang et al. Geometry-Aware Scene Text Detection with Instance Transformation Network. In CVPR 2018
[21] Yuqiang Zheng,Yuan Xie,Yanyun Qu,Xiaodong Yang,Cuihua Li,Yan Zhang. Scale robust deep oriented-text detection network[J]. Pattern Recognition,2019
[22] Zichuan Liu et al. Towards Robust Curve Text Detection with Conditional Spatial Expansion. In CVPR2019
[23] Minghui Liao et al. Real-time Scene Text Detection with Differentiable Binarization. In AAAI2020
[24] Xu Zhao et al. Elite Loss for scene text detection. Neurocomputing 333: 284-291 (2019)
[25] YuLiang Liu et al. ABCNet: Real-time Scene Text Spotting with Adaptive Bezier-Curve Network. In Proc. IEEE Conf. Comp. Vis. Pattern Recogn. (CVPR) 2020
[26] Chengqian Zhang et al. Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes. CVPR 2019: 10552-10561
[27] Minghui Liao et al. Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes. TPAMI 2019
[28] Shi-Xue Zhang et al. Deep Relational Reasoning Graph Network for Arbitrary Shape Text Detection. CVPR 2020
[29]Yuxin Wang et al. ContourNet: Taking a Further Step Toward Accurate Arbitrary-Shaped Scene Text Detection. CVPR 2020
[30]Yiqin Zhu et al. Fourier Contour Embedding for Arbitrary-Shaped Text Detection. CoRR abs/2104.10442 (CVPR 2021)
[31] Minghang He, Minghui Liao, Zhibo Yang, Humen Zhong, Jun Tang, Wenqing Cheng, Cong Yao, Yongpan Wang, Xiang Bai: MOST: A Multi-Oriented Scene Text Detector with Localization Refinement. CoRR abs/2104.01070 (CVPR 2021)
[32] Yixing Zhu, Jun Du: TextMountain: Accurate scene text detection via instance segmentation. Pattern Recognit. 110: 107336 (2021)
[33] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, Tal Hassner: TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. CVPR (2021)
[34] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, Wenyu Liu: Scene Text Retrieval via Joint Text Detection and Similarity Learning. CVPR (2021)
[35] Shi-Xue Zhang, Xiaobin Zhu, Chun Yang, Hongfa Wang, Xu-Cheng Yin: Adaptive Boundary Proposal Network for Arbitrary Shape Text Detection. ICCV (2021)
[36] Pengwen Dai, Sanyi Zhang, Hua Zhang, Xiaochun Cao:
Progressive Contour Regression for Arbitrary-Shape Scene Text Detection. CVPR 2021: 7393-7402