融合局部特征与全局特征的场景文本检测算法

赵晓芹
(中国石油大学(华东) 计算机科学与技术学院,山东 青岛 266580)

摘 要:检测复杂场景下的文本是一项极具挑战性的任务,现有的文本检测方法有将字符作为目标进行检测的,也有将单词作为目标进行检测的。对于单词内部排列较为松散的文本或字符之间间隔较小的文本,基于字符的检测算法容易将一个单词检测为多个单词,或将多个单词检测为一个单词。在这种情况下,基于单词的方法检测精度要更高一点,但是基于字符的方法比基于单词的方法更能准确的检测到文本中的每个符号。鉴于它们各自的优缺点,使用 ResNet 与 FPN 结合的网络,将这两种方法进行整合,充分利用文本的底层特征与高层特征。在检测单词的同时也检测单词中每个字符的信息,将这两种信息优化、融合,从而达到一种更好的检测效果。为了降低标注字符数据集的成本,在实验中加入弱监督的方法,使网络在只有单词标注的数据集上训练也能很好的检测字符。最后在 ICDAR 2013 数据集、ICDAR 2015 数据集和 Total-Text 数据集上验证此方法的有效性。

关键词:文本检测;弱监督;特征融合;ResNet;FPN
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2022)0025-06

Scene Text Detection Algorithm Combining Local and Global Features

ZHAO Xiao-qin
(School of Computer Science & Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Detecting text in complex scenes is a challenging task. Some methods use character as target for detection, and some use word as target for detection. For words with loosely or tightly arranged between characters, character-based method easily detects one word as multiple words, or multiple words as one word. In this case, word-based method has higher accuracy, but the character-based method can detect each character more accurately than the word-based method. In view of these advantages and disadvantages, ResNet+FPN network is used to integrate these methods and make full use of the shallow and deep features of text. The network detects words and characters at the same time, optimizes and merges these two kind of information to achieve better result. In order to reduce the cost of labeling character data sets, weakly supervised learning is added to the experiment, so that the network can detect characters well when training on the data sets with word annotations. Finally, the effectiveness of this method is verified on ICDAR 2013、ICDAR 2015 and Total-Text data sets.

Key words: text detection; weakly supervised learning; features fusion; ResNet; FPN

0 引 言

文字是人类生活中不可或缺的组成部分,承载了丰富的语义信息,对人们理解图像起到了至关重要的作用。相比于发展成熟的文本文档识别技术(OCR)[1],自然场中的文本检测显然更具有挑战性。OCR 更加擅长白纸黑字的识别,而自然场景中的文本存在更多的不确定性,比如场景中的文本存在各种角度,有多种多样的字体,容易受到光照的影响,还存在遮挡和阴影等种种问题。

在深度学习出现之前,传统的文本检测方法一般采用手工特征提取的方式检测文本,比如 SWT(笔画宽度变化)和 MSER(最大极值稳定区域)等,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。传统的场景文本检测技术是基于自上而下的特征生成方法,需要人工设计文本的形状和线条,但是由于自然场景下文本的背景十分复杂,文本也会出现扭曲变形等情况,所以这一技术在复杂的自然场景中对文本的检测效果欠佳。近年来,随着深度学习的发展,复杂场景中的文本检测技术[2]有了很大的进步,越来越多的人专注于研究复杂场景下的文本检测。文献[3]通过对图像进行区域划分的方法减弱复杂背景对检测结果的影响从而提高检测精度;文献[4]提出融合分

收稿日期:2021-11-22
基金项目:国家自然科学基金(61379106),山东省自然科学基金(ZR2013FM036,ZR2015FM011)
作者简介:赵晓芹(1997-),女,硕士研究生,研究方向为自然场景中的文本检测。

类置信度与定位置信度的方法优化检测结果;文献[5]提出将图像中的文本划分为三个区域进行分割检测;也有人通过尝试对图片进行预处理提高检测精度,例如文献[6]。

基于深度学习的检测方法不再需要手工设计文本特征,而是利用神经网络自动学习,并且捕捉像素之间的关系,从而实现更加精确的检测。检测类型也从最开始只能检测水平文本,到检测多方向文本,再到检测不规则文本和复杂场景下的文本。

根据检测方法的不同,可以将近年来常用的文本检测方法大致分为两种类型:基于字符的检测方法和基于文本的检测方法。基于字符的方法是将单个字符作为检测目标,首先检测场景中的字符,然后将检测到的字符按照某种规则组合成单词。基于字符的检测方法虽然能够更准确的定位文本,但是由于字符标注的数据集较少,而且标注字符数据集需要花费非常大的成本,所以这种检测方法并不流行。基于文本的检测方法能直接检测场景中的文本,这种方法比基于字符的方法更加简洁方便,但是对于形状不规则的文本,这种方法很难完整的检测出文本中的每一个字符。该文提出一种新的检测方法,将局部的字符语义与全局文本语义信息相结合,这种方法既能快速定位每个单词的位置,又能准确定位每个字符的位置,同时还采用了弱监督的方法来解决字符级数据不足的问题。

1 相关工作

传统方法在提取特征时,需要不断的进行预处理和后处理的步骤,传统方法主要提取的是底层的图像特征,这些特征几乎无法处理复杂的文本。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。起初,文本检测是在目标检测算法的基础上进行尝试和改进的,所以基于目标检测的文本检测算法就是在骨干网络的基础上,通过增加额外的卷积层得到文本区域,然后对得到的文本区域执行两个子任务:判断文本区域是文本的概率和回归调节正样本的位置。基于目标检测的算法又可以大致分为将文本作为目标进行检测和将文本片段作为目标进行检测两种方式。基于文本的检测方法可以直接检测图中的文本区域,而基于文本片段的检测方法首先需要检测多个包含文本的片段,然后基于某种规则将文本片段组合成文本。

1.1 基于文本的检测方法

由于文本目标与普通目标有着非常大的差异,文本具有更大的长宽比,所以不能直接用目标检测的方法检测文本,TextBoxes[7]、TextBoxes++[8]在目标检测的基础上,通过修改预设矩形框的大小来检测长文本,但是由于文本的长度变化范围比较大,通过修改预设文本框的大小不能兼顾到各种长度的文本。由于文本的形状是不规则的,所以用矩形框表示弯曲程度较大的文本容易引入过多的背景信息干扰检测精度。为了减少背景信息对文本的干扰,文献[9]使用多个点组成的多边形表示文本区域。

虽然使用多边形可以减少背景信息的干扰,但是弯曲程度越高的文本,需要的点也就越多,回归多个点不仅增加了网络的压力,而且每个点需要非常精确的预测,点越多,到最后累积的误差也会越大。为了解决这个问题,ABCNet[10]提出用三阶贝塞尔曲线来表示文本,文本上下边界各取四个点生成新的边界来表示文本区域。对于弯曲比较严重的长文本,使用三阶贝塞尔曲线不足以表示整个文本区域,而且对于包含多个单词的文本而言,做到每个字符的精确检测也是很困难的。

1.2 基于文本片段的检测方法

由于受到感受野的限制,基于文本的检测方法不能检测较长的文本,所以有学者提出基于文本片段的检测方法,CTPN[11]通过预设小的矩形框来检测多个文本片段,然后将检测出的文本片段进行筛选组合。CTPN 判断相邻文本片段是否属于同一个文本时,只是对比两个相邻矩形框之间的距离是否小于 50 个像素,这种方法只适用于内部排列紧密的文本。SegLink[12]在检测文本片段的同时还检测了片段之间的连接关系,并通过连接来判断相邻矩形框是否属于同一文本,该模型同样不能检测内部字符之间隔较大的文本。基于文本片段的检测方法容易将内部排列松散的单词检测成多个单词,对于单词之间排列比较密集的情况,也会出现将多个单词组合为一个单词的情况。

文本片段可以是包含文本的一部分,也可以是完整的字符。近年来,大多数基于文本片段的检测方法都是把字符作为一个检测单元,但是由于字符级文本标注成本较高,相关公开数据集稀少,导致多数检测模型只能在具有文本标注的数据集上进行训练。百度 2017 年发表的 WordSup[13]巧妙地运用弱监督训练,解决了字符标注数据集不足的问题。但是由于 WordSup 只检测每个字符的位置,忽略了字符之间的关系,同样容易导致模型将距离较近的两个单词检测为一个单词,将字符间排列比较松散的一个单词检测为多个单词。

大多数检测方法都是把单词或文本行作为一个检测单元,这种方法更加方便快捷,但是对于各种语言和数学公式,字符是基本的组成部分。不同的文本有不同的结构,比如顺序化的语言文本和结构化的数学公

式,而且在不同的视角和情境中,文本的视觉形状和扭曲程度也不同,但是文本有一个共同的特点:都是由字符组成的。虽然基于文本的检测方法更加快捷,但是基于字符的检测方法比基于文本的方法能更加精准的检测到文本内部的每一个字符,所以在文本检测的发展中,基于字符的检测方法也占据了相当重要的地位。

2　文中算法

由于检测文本的最终目的是为了方便文本识别算法更好的识别文本,如果将整个文本行作为一个检测单元,对于背景复杂的文本来说,在后续识别的过程中很容易出现多字或少字的情况。为了更好的服务于识别算法,应尽量检测更小的文本单元,所以把字符作为目标进行检测。相对于普通目标来说,文本目标的优点是其具有丰富的上下文信息,所以为了能更准确将检测出来的字符组合成文本,还加入了对文本的检测。

网络的训练目标是能够精确检测自然场景中的每个单词以及字符的位置。由于网络既要检测单词,又要检测字符,而基于字符标注的公开数据集相对较少,所以本文使用弱监督训练模型的方法,在单词标注的数据集上训练网络,使网络具有检测字符的能力。

正如前文所提到的,基于字符检测的算法存在各种各样的问题,网络容易将内部字符排列松散的文本检测为多个文本,或者将文本与文本之间间隔较小的多个文本检测为一个文本,为了提高检测精度,我们不直接将检测到的字符连接成文本,而是在检测字符的同时检测单词。通过利用单词的语义信息来更加准确的组合字符。

2.1　网络框架

由于场景中的文本变化多样,与普通目标相比,文本会有更大的尺度变化,所以使用与 Mask TextSpotter[16] 一样的 ResNet+FPN 组成的网络作为骨干网络来提取多个尺度的特征图。网络框架如图 1 所示,主要由四部分组成:提取特征的 FPN[14] 网络;生成文本候选区域的 RPN[15] 网络;对候选区域进行细化的 ROIAlign 和融合字符特征与文本特征的融合分支。图片通过骨干网络提取到特征图后,将其送入 RPN 网络生成一系列的文本候选区域,这些文本候选区域包括字符候选区域和单词候选区域,这些候选区域再被送入检测分支进行细化,细化主要包括预测候选区域的种类和对候选区域进行 bounding box 回归操作。细化部分用 ROIALign 分别提取单词特征和字符特征,并且对单词和字符部分进行检测。最后,将提取到的这两种不同的特征进行融合。

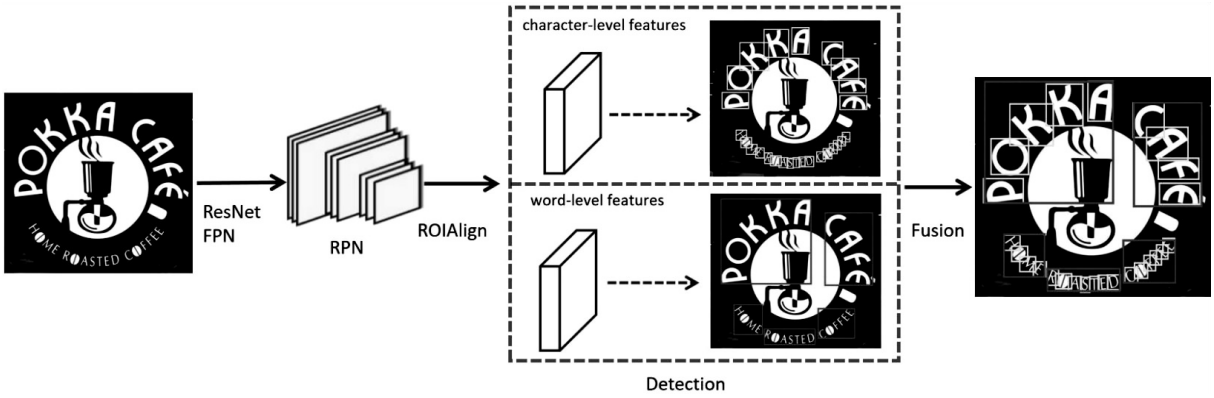

图 1　网络框架

图片通过 FPN 与 RPN 之后生成一系列的字符和单词的候选区域,基于这些生成的文本候选区域,通过 ROIALign 提取了单词特征与字符特征,再通过像素求和的方式将这两种不同的特征进行融合,得到更加丰富的文本特征,然后将融合的特征进行 3 * 3 和 1 * 1 卷积,再将卷积后的特征进行分类和候选框的回归。由于场景中文本变化较大,我们很难通过预设包围框的大小来检测到场景中所有的文本。

受文本检测算法 SegLink 的启发,为了将场景中不同大小的文本更加精确的检测出来,在骨干网络输出的多个尺度的特征图上分别进行了字符和单词的检测和特征提取,具体细节如图 2 所示。由于需要在多个尺度的特征图上提取文本候选区域,为了减轻网络对边界框回归的压力,该文使用最简单的矩形框表示文本区域,文本的位置用矩形框的左上角及右下角的坐标来表示。

2.2　弱监督训练

网络既需要检测字符又需要检测单词,所以训练网络不仅需要单词标注的数据集,还需要字符标注的数据集。由于能够支撑训练网络的字符标注数据集较少,而且标注字符数据集需要花费大量的时间、金钱和精力。受到 WordsUp 的启发,采用弱监督训练的方法来训练我们的网络,使网络具有同时检测单词和字符的能力。所谓的弱监督训练,就是在只有单词标注的数据集训练网络的情况下,使得网络既能检测单词,又能检测字符。

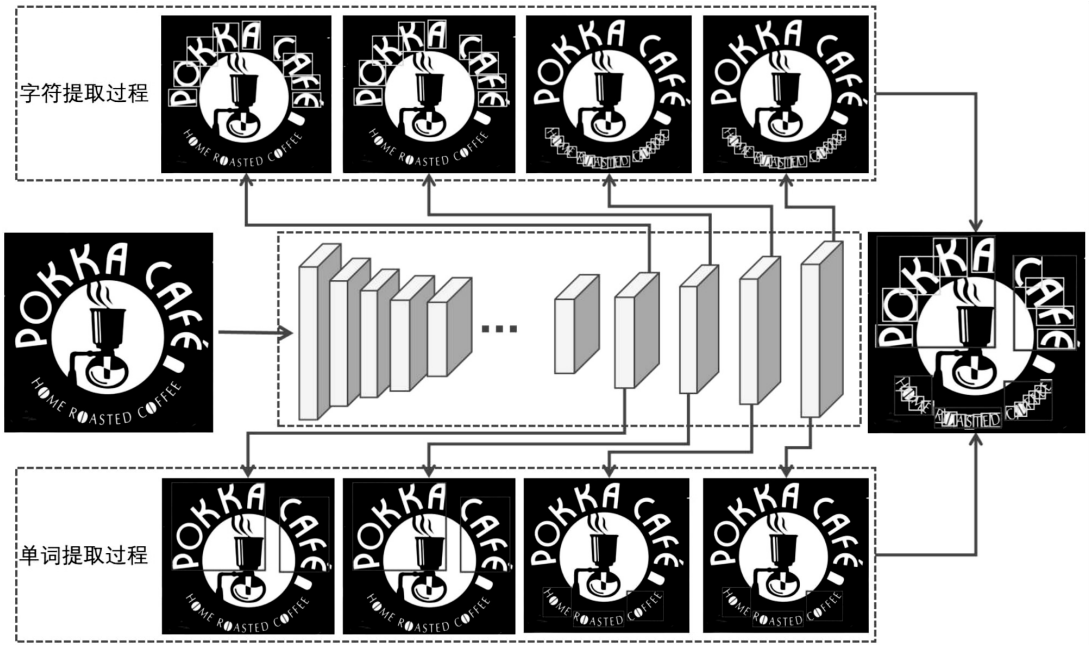

图 2 多尺度特征提取过程

2.3 训练过程

整个网络的训练过程分为三个步骤:模型预训练;弱监督训练;文本检测与微调。

由于合成数据集 SynthText 既有单词标注信息,又有字符标注信息,所以用 SynthText 数据集预训练模型,使该模型在正式训练之前具有检测字符的能力。预训练好的模型分别在 ICDAR 2015 和 Total-Text 这些只有单词标注的数据集上进行弱监督训练,使得训练好的网络既能检测单词又能检测字符。

模型在单词标注的训练数据集上进行弱监督训练时,会检测出一连串的字符,由于合成数据集 SynthText 与真实数据之间存在一定的差异,所以模型在弱监督训练初期对字符的检测并不是百分之百正确的,由于弱监督训练时没有字符的监督信息,此时更容易出现字符检测错误的情况,所以需要对检测出的字符计算一个置信度概率,在计算损失时为其乘以这个置信概率。

假设检测的单词 w 中有 N 个字符(数据集中有单词的信息,所以不需要标注每个字符也能知道每个单词由多少个字符组成),网络对每个单词会检测出 M 个字符,其置信度为公式:

$$S_c(w) = 1 - \frac{|N - M|}{N} \tag{1}$$

将识别出的字符结合原来的单词标注信息,在当前的训练数据集上微调预训练好的模型。由于 ICDAR 2015 和 Total-Text 数据集只有单词标注信息,为了保证模型的有效性,在这两种数据集上进行训练时,还加入了少量带有字符标注信息的合成数据来辅助训练。在训练过程中,只加入带有字符标注信息的合成数据集来辅助训练是远远不够的,因为合成的数据与真实数据之间还是有一定差别的,所以在训练时,需要将对真实文本生成的字符区域与单词标注信息做损失,由于刚开始模型在单词标注的数据集上预测字符的准确性没有保证,所以在计算损失的时候,需要为此处单词实例对应的损失乘以一个置信参数,置信参数计算公式为公式(1)。图片的像素级置信度为公式(2)。由于合成数据有字符标注信息,所以其置信参数为 1。

$$S_c(p) = \begin{cases} S_c(w), p \in w \\ 1, p \notin w \end{cases} \tag{2}$$

其中 w 表示单词实例,即特征图上某像素点若在单词区域内则此处的置信参数为整个单词的置信参数,否则为 1。

3 实验及结果分析

为了验证该方法的有效性,在数据集 ICDAR 2013,ICDAR 2015 和 Total-Text 上对模型进行了评估,并且和一些比较经典的文本检测方法做了比较。

3.1 数据集

ICDAR 2013[17] 数据集是文档分析与识别国际会议于 2013 年举办的场景文本检测竞赛中使用的标准数据集,该数据集只有水平方向的文本,包含 229 张训练图片和 233 张测试图片。

ICDAR 2015[18] 是一个包含多方向文本的数据集,于 2015 年场景文本检测竞赛中提出。与 ICDAR 2013 数据集不同,ICDAR 2015 数据集中的文本主要是场景中附带的文本,没有 ICDAR 2013 中的文本那么规整和清晰。ICDAR 2015 数据集包含 1 000 张训练图片

和 500 张测试图片。

Total-Text[19] 数据集是最大弯曲文本数据集之一,包含 1 255 张训练图片和 300 张测试图。

SynthText[20] 数据集是一个巨大的合成数据集,用来对模型进行预训练,此数据集有 80 万张图片,包含 8 百万个合成单词,既有单词的标注信息,又有字符的标注信息。

3.2　实验设置

该文采用 ResNet50 作为网络的框架,训练过程主要为文本分类和检测框的回归,网络使用交叉熵损失作为文本分类损失,SmoothL1 损失用于检测框回归。在训练过程中,采用随机梯度下降(SGD)来优化网络,权重衰减系数设置为 0.000 1,动量设置为 0.8,batch size 为 10。

3.3　评估指标

该文使用召回率、精确率和 F-Measure 对模型进行了评估。其中召回率是预测正确的正样本数量与正样本总数的百分比;精确率是预测正确的正样本数量与预测为正样本的样本数量的百分比;F-Measure 是对召回率和精确率的加权调和平均数。召回率、精确率和 F-measure 的计算公式如下:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{F} - \text{measure} = \frac{2 * PR}{P + R} \tag{5}$$

其中,TP 表示正样本中被检测为正类的样本数量,FN 表示被检测为负类的正样本数量,FP 表示被预测为正类的负样本数量。

3.4　实验分析与结果对比

用 ResNet+FPN 作为骨干网络提取单词与字符特征,在特征提取过程中,将纹理、颜色等底层文本特征与深层语义信息融合,解决了在检测过程中字符漏检与将背景误检等问题,不仅精确地检测出了每个单词,还将单词中的每个字符也准确检测出来。通过召回率 Recall、精确率 Precision 和 F-Measure 等指标,将文中方法与 CTPN、SegLink、WordSup、Mask TextSpotter 等经典的文本检测方法在 ICDAR2013、ICDAR2015 和 TotalText 等三种类型的数据集上进行了比较。

表 1 所示为文中方法与其他文本检测方法在水平文本数据集是的对比结果。在精确率上,相较于 SPCNet,文中方法提升了 0.3%。但是与 CTPN、SegLink、TextBoxes++等基于 VGG 实现的方法相比,在召回率和精确率都有明显的提高,这就证明了底层信息与高层信息的融合能够提高检测精度。

表 1　不同方法在 ICDAR2013 数据集上的结果

方法	Recall	Precision	F-measure
CTPN	83.0	93.0	88.0
SegLink	83.0	87.7	85.3
WordsUp	87.5	93.3	90.3
TextBoxes++	86.0	92.0	89.0
Mask TextSpotter	88.6	95.0	91.7
SPCNet	90.5	93.8	92.1
文中	89.1	94.1	91.5

表 2 所示为文中方法与其他方法在多方向直线型文本数据集上的对比结果。与 Mask TextSpotter 相比,在 F-measure 上提升了 1.3%。与 SPCNet[21] 相比,在召回率上与其持平,在精确率上稍有提升。

表 2　不同方法在 ICDAR2015 数据集上的结果

方法	Recall	Precision	F-measure
CTPN	52.0	74.0	61.0
SegLink	76.8	73.1	75.0
WordsUp	77.0	79.3	78.2
TextBoxes++	78.5	87.8	82.9
Mask TextSpotter	81.0	91.6	86.0
PSENet	84.5	86.9	85.7
SPCNet	85.8	88.7	87.2
文中	85.8	88.9	87.3

表 3 所示为文中方法与 Mask TextSpotter、PSENet[22]、SPCNet 在弯曲文本数据集上的对比结果。在召回率上,略逊于 SPCNet,在精确率上比 SPCNet 高出 0.1%。该骨干网络采用与 Mask TextSpotter 相同的 ResNet50,但是在召回率和精确率上的表现都优于 Mask TextSpotter。

表 3　不同方法在 TotalText 数据集上的结果

方法	Recall	Precision	F-measure
Mask TextSpotter	55.0	69.0	61.3
PSENet	78.0	84.0	80.9
SPCNet	82.8	83.0	82.9
文中	82.1	83.1	82.6

4　结束语

借鉴了经典检测方法 Mask TextSpotter,使用 ResNet+FPN 作为骨干网络,将浅层文本特征与高层语义信息融合,解决了在特征提取过程中底层信息丢失的情况。同时还加入了弱监督的方法使网络能在只有单词标注的数据集进行训练的情况下,还能很好的检测字符。为了能够更好的检测各种大小的文本,在多个不同尺度的特征图上都进行了单词和字符的检测。

由于图片上背景区域所占比例远远高于文本区域所占比例,大量候选区域的分类和回归会使得整个网络的检测速度有所下降,所以为了尽量减少网络的压力,在检测单词和字符时都使用简单的水平矩形框表示文本。尽管如此,大量候选区域的分类和回归还是会拖慢网络的速度,所以接下来会着重于研究在保持精度不变的情况下如何提高检测速度。

参考文献:

[1] 张婷婷,马明栋,王得玉. OCR 文字识别技术的研究[J]. 计算机技术与发展,2020,30(4):85-88.

[2] 许 肖,顾 磊. 复杂背景下文本检测研究[J]. 计算机技术与发展,2015,25(3):40-44.

[3] 陈梓洋,王宇飞,钱 侃,等. 自然场景下基于区域检测的文字识别算法[J]. 计算机技术与发展,2015,25(7):230-233.

[4] 蒋志鹏,潘坤榕,张国林,等. 基于置信度融合的自然场景文本检测方法[J]. 计算机技术与发展,2021,31(8):39-44.

[5] 李 煌,王晓莉,项欣光. 基于文本三区域分割的场景文本检测方法[J]. 计算机科学,2020,47(11):142-147.

[6] 章 安,马明栋. 基于 Tesseract 文字识别的预处理研究[J]. 计算机技术与发展,2021,31(1):73-76.

[7] LIAO M,SHI B,BAI X,et al. Textboxes:a fast text detector with a single deep neural network[C]//Thirty-first AAAI conference on artificial intelligence. [s. l.]: AAAI, 2017: 4161-4167.

[8] LIAO M,SHI B,BAI X. Textboxes++:a single-shot oriented scene text detector[J]. IEEE Transactions on Image Processing,2018,27(8):3676-3690.

[9] QIN S, BISSACCO A, RAPTIS M, et al. Towards unconstrained end-to-end text spotting[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul:IEEE,2019:4704-4714.

[10] LIU Y, CHEN H, SHEN C, et al. Abcnet: real-time scene text spotting with adaptive bezier-curve network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [s. l.]:IEEE,2020:9809-9818.

[11] ZHI T,HUANG W,TONG H,et al. Detecting text in natural image with connectionist text proposal network[C]//European conference on computer vision. [s. l.]:Springer,2016.

[12] SHI Baoguang,BAI Xiang,BELONGIE S. Detecting oriented text in natural images by linking segments[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu:IEEE,2017:3482-3490.

[13] HU H,ZHANG C,LUO Y,et al. Wordsup:exploiting word annotations for character based text detection[C]//Proceedings of the IEEE international conference on computer vision. Venice:IEEE,2017:4950-4959.

[14] LIN T Y,DOLLÁR P,GIRSHICK R,et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu:IEEE,2017:936-944.

[15] REN S,HE K,GIRSHICK R,et al. Faster R-CNN:towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. [s. l.]:[s. n.],2015.

[16] LYU P,LIAO M,YAO C,et al. Mask textspotter:an end-to-end trainable neural network for spotting text with arbitrary shapes[C]//Proceedings of the European conference on computer vision (ECCV). Munich:Springer,2018:71-88.

[17] KARATZAS D, SHAFAIT F, UCHIDA S, et al. ICDAR 2013 robust reading competition[C]//2013 12th international conference on document analysis and recognition. [s. l.]: IEEE,2013:1484-1493.

[18] KARATZAS D,GOMEZ-BIGORDA L,NICOLAOU A,et al. ICDAR 2015 competition on robust reading[C]//2015 13th international conference on document analysis and recognition (ICDAR). [s. l.]:IEEE,2015:1156-1160.

[19] CH'NG C K,CHAN C S. Total-text:a comprehensive dataset for scene text detection and recognition[C]//2017 14th IAPR international conference on document analysis and recognition (ICDAR). Kyoto:IEEE,2017:935-942.

[20] GUPTA A,VEDALDI A,ZISSERMAN A. Synthetic data for text localisation in natural images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:2315-2324.

[21] XIE E,ZANG Y,SHAO S,et al. Scene text detection with supervised pyramid context network[C]//Proceedings of the AAAI conference on artificial intelligence. [s. l.]: AAAI, 2019:9038-9045.

[22] WANG W,XIE E,LI X,et al. Shape robust text detection with progressive scale expansion network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [s. l.]:IEEE,2019:9336-9345.