ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 2,Mar. 2022,pp. 415－425
DOI:10. 16337/j. 1004-9037. 2022. 02. 014

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

局部与全局双重特征融合的自然场景文本检测

李云洪，闫君宏，胡　蕾

（江西师范大学计算机信息工程学院，南昌 330022）

摘　要：自然场景中文本的形状、方向和类别等变化丰富，场景文本检测仍然面临挑战。为了能够更好地将文本与非文本分隔并准确定位自然场景图像中的文本区域，本文提出一种局部与全局双重特征融合的文本检测网络，通过跳跃连接的方式实现多尺度全局特征融合，对恒等残差块进行改进实现局部细粒度特征融合，从而减少特征信息丢失，增强对文本区域特征提取力度，并采用多边形偏移文本域与文本边缘信息相结合的方式准确定位文本区域。为了评估本文方法的有效性，在现有经典数据集ICDAR2015和CTW1500上进行了多组对比实验，实验结果表明在复杂场景下该方法文本检测的性能更加卓越。

关键词：文本检测；跳跃连接；细粒度特征融合；全局特征融合；多边形偏移文本域

中图分类号：TP391　　**文献标志码：**A

Natural Scene Text Detection Based on Local and Global Dual-feature Fusion

LI Yunhong，YAN Junhong，HU Lei

（School of Computer Information Engineering，Jiangxi Normal University，Nanchang 330022，China）

Abstract：The shape，direction and category of text in natural scenes are varied，and scene text detection is still a challenge. In order to better separate text from non-text and accurately locate the text area in natural scene image，this paper proposes a text detection network that fuses local and global features. Multi-scale global feature fusion is realized through jump connection，and the constant residual block is improved to realize local fine-grained feature fusion，thereby reducing the loss of feature information and enhancing the strength of feature extraction in text regions. The combination of polygon offset text field and text edge information is used to local text region accurately. In order to evaluate the effectiveness of the method in this paper，multiple sets of comparative experiments are conducted on the existing classic data sets ICDAR2015 and CTW1500. The experimental results show that the method has better performance in text detection in complex scenes.

Key words: text detection; jump connection; fine-grained feature fusion; global feature fusion; polygon offset text field

引　　言

文本作为人类沟通的主要媒介之一，经常出现在自然场景图像中，例如商场商标、街道路标、车牌

基金项目：国家自然科学基金（61662033）；江西省教育厅科学技术研究（GJJ210326）。
收稿日期：2021-04-25；**修订日期：**2021-07-28

和票据等,文本信息对理解和解析场景内容有极其重要的作用,自然场景文本识别[1]一直深受研究者关注,而准确有效的文本检测是文本识别的前提。相较于文档类文本检测,背景多文字少、遮挡、文本类似块、文体形态各异、大小排列不一、方向不同、弯曲、艺术体、镜面反光等因素导致自然场景文本检测仍然面临严峻挑战。

传统的文本检测算法多采用自底向上方式进行文本检测,大致可分为两类:基于连通区域的算法和基于滑动窗口的算法。基于连通区域的文本检测算法多从图像边缘检测开始,根据文本的低级属性(大小、颜色和形状等)形成多个连通区域,然后对连通区域进行处理、合并生成最后的文本框,较为经典的算法有Matas等[2]提出的最大稳定极值区域(Maximally stable extremal regions,MSER)算法,Epshtein等[3]提出的笔画宽度变换(Stroke width transform,SWT)算法。基于滑动窗口的文本检测算法最早出现在目标检测中,Zitnick[4]提出的Edge Boxes算法,在一幅图像上形成一个特定大小窗口,从左上角开始以特定步长扫描,寻找文本出现的区域,并对区域进行评分,根据分数高低来确定候选框。传统的文本检测算法对上下文信息较为依赖,一些类似文本纹理的干扰会导致严重的误检与漏检。

近年来深度学习技术在目标检测中取得了显著成效[5],可将文本视为被检测目标。由于文本定位需覆盖整个字符区域,而场景文本没有规律的边缘界限,导致很多现有的目标检测算法在文本检测中不能直接使用,很多研究者针对场景文本检测进行了技术的迁移与改进。目前深度学习技术下文本检测算法大致可分为两类:(1)基于文本框回归的算法,使用四边形表征文本区域,在文本方向多样、长短不一等情况下,该方法存在一定局限性;(2)基于文本分割的算法,将文本与非文本进行分割,不需要考虑文本的长短与方向。典型的文本框回归算法有Tian等[6]提出的联接文本提议网络(Connectionist text proposal network,CTPN)算法,该算法对基于区域的快速卷积神经网络(Faster region-based convolutional neural network,Faster RCNN)[7]做了改进,考虑了水平文本的长短不确定性,用碎片框进行文本区域定位,利用循环神经网络(Recurrent neural network,RNN)的语义信息,通过长短期记忆网络(Long short-term memory,LSTM)合并文本的碎片区域生成最终的文本框。Shi等[8]提出的SegLink算法,可以检测任意角度的文本,在CTPN的思想上融入了单次多框检测器(Single shot multibox detector,SSD)[9]算法思路,在SSD中引入角度因子,检测出包含方向的多个候选框,拼接属于同一个文本的候选框得到最终文本框。典型的文本分割算法有Deng等[10]提出的Pixellink算法,在对文本或者非文本像素进行分离预测的基础上,预测文本像素的8个方向上是否存在连接,通过判断连通区域得到最终的文本框。Long等[11]提出的TextSnake算法,首先在分割结果中确定文本中心线,然后围绕中心线采用不同大小和连接角度的圆盘覆盖文本区域,从而提高不规则文本检测性能。Wang等[12]提出的渐进多尺度扩展网络(Progressive scale expansion network, PSENet)算法,通过精确查找多个尺度的内核,对紧密相连的文本进行准确定位与检测,该方法很大程度上解决了紧密相连文本的问题。从实验结果分析,这些方法在文本区域所占比例较小或者具有不规则艺术体的场景图像中会出现严重的漏检与误检。

本文在PSENet(Resnet-50)[12]的基础上提出一种局部与全局双重特征融合的网络模型(Local and global network,LAGNet),选择ResNet-50[13]作为骨干网络,对恒等残差块进行改进,实现局部细粒度特征融合(Fine-grained locally feature fusion,FLFF);然后在特征金字塔网络(Feature pyramid networks,FPN)[14]结构中采用跳跃连接的方式,实现多尺度全局特征融合(Multi-scale global feature fusion,MGFF),从而增强特征提取的性能;最后将多边形偏移文本域与真实文本边缘信息结合,对文本进行准确定位从而实现文本的检测。

1 LAGNet网络模型

本文提出的自然场景文本检测网络模型LAGNet如图1所示,采用FPN结构的ResNet-50作为核

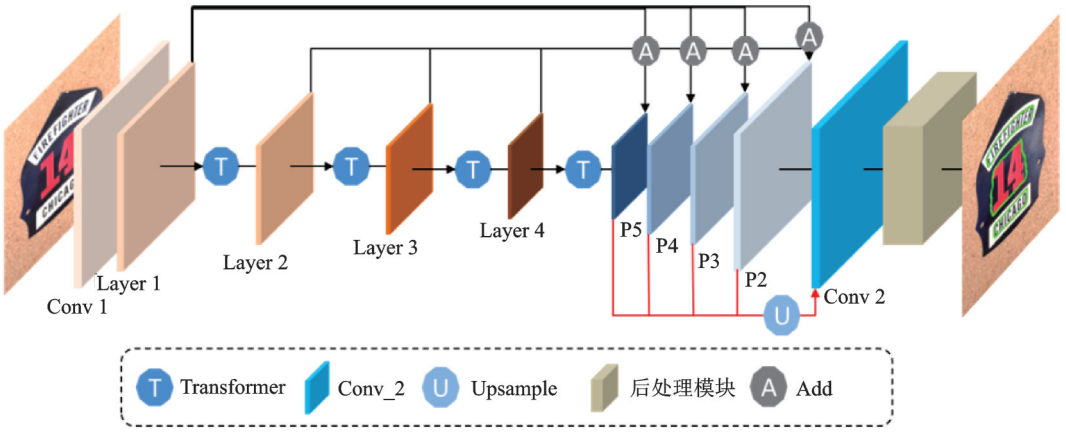

图1 LAGNet网络模型

Fig.1 LAGNet network model

心网络,主要包含Layer 1~Layer 4构成的Down-top分支,P5~P2构成的Top-down分支,并在每一个Layer层后引入Transformer,实现图片尺寸缩放与通道降维以对不同尺度大小的图片进行卷积,从而利用浅层特征区分显著文本、利用深层特征区分较小文本。在Layer层的恒等残差块中,用多分支卷积替换单分支卷积,实现局部特征融合;将Layer中的特征信息通过跳跃连接的方式传递并采用Add的形式融入到P5~P2,实现全局特征融合。通过浅层与深层的特征信息全局共享,增强了网络模型对各类文本检测的鲁棒性。P5~P2中的特征映射上采样(Upsample)到原图尺寸并输入到Conv 2,得到一个文本实例区域、一个多边形偏移文本域、一个文本边缘信息,经过后处理模块形成最终的检测结果。其中Conv_2由Concat与n个Conv-BN-ReLU层和Conv-Sigmoid层组成,Concat指将不同卷积层的特征通道融合,Add指在保证通道数相等的情况下将卷积结果逐元素叠加。

2 功能模块

2.1 特征提取网络

现有经典算法在对文本区域所占比例较小的场景图像中会出现大量的漏检误检情况,本文的网络模型从特征提取模块入手,对ResNet-50的恒等残差块进行改进,尽可能地保留底层语义信息,以提高对小文本的检测性能。

在Layer 1的恒等残差块中引入密集残差块思想,如图2所示,将每一层的卷积结果保留并传递给之后的每一层,并通过Add形式进行特征融合,从而使局部特征信息通过深度级联聚合传递到整个网络,经过Transformer1对图像尺寸与通道数进行处理并传递给Layer 2。此处采用Add的融合形式是因为Add的计算量比Concat低很多。

Layer 2和Layer 3的恒等残差块(图3)引入细粒度特征信息融合思想,采用分割-转换-合并的结构实现细粒度特征信息融合。具体为,将初始残差块中3×3卷积分割成n条分支同步进行,即图3中a1~a4(n=4),采用Add形式对各分支卷积结果进行融合(图3中B),将B传递到1×1卷积。在卷积后采用Add形式将前两层保留的特征信息进行融合,实现细粒度的特征信息提取,不仅扩大了感受野,同时提升了卷积的表达能力。

Layer 4的恒等残差块(图4)采用传统的卷积块模式,在直接映射中加入了一个1×1卷积,对最小尺度的文本进行特征提取并堆叠。核心网络ResNet-50经过残差块调整后构成的Down-top分支参

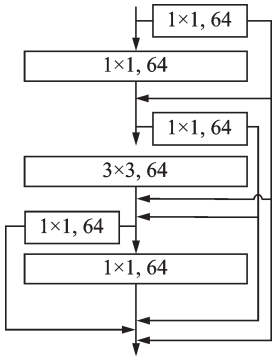

图2 Layer 1恒等残差块结构图

Fig.2 Identity residual block structure diagram of Layer 1

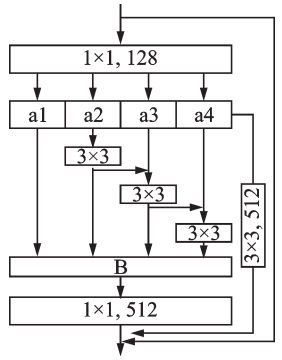

图3 Layer 2和Layer 3恒等残差块结构图

Fig.3 Identity residual block structure diagram of Layer 2 and Layer 3

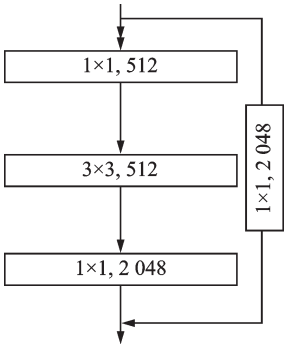

图4 Layer 4恒等残差块结构图

Fig.4 Identity residual block structure diagram of Layer 4

数如表1所示,从表1中可以很直观地看出输入图片的尺寸在每一阶段的变化情况以及所进行的操作。

表1 Down-top网络参数

Table 1 Down-top network parameters

名称	输出大小	卷积参数	名称	输出大小	卷积参数
Conv 1	320×320	7×7×64, stride=2	Transformer2	80×80	3×3, stride=2, padding=1
	160×160	3×3maxpool, stride=2, padding=1	Layer 3	80×80	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1\,024\end{bmatrix}\times6$
Layer 1	160×160	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,64\end{bmatrix}\times3$	Transformer 3	40×40	3×3, stride=2, padding=1
Transformer 1	160×160	1×1,256	Layer 4	40×40	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2\,048\end{bmatrix}\times3$
Layer2	160×160	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	Transformer 4	20×20	3×3, stride=2, padding=1

注:本文采用跳跃连接的方式,将Layer 1层提取的特征经过上采样与降采样处理后,通过Add与P5到P2各层特征相融合,同时Layer 2、Layer 3和Layer 4采用同样的方式,将每一层的特征信息均传递给P5到P2,该方式实现多尺度全局特征提取与融合,提高网络特征提取性能。

2.2 后处理模块

现有公共数据集中,不仅存在部分紧密衔接的文本图像,而且存在一些遮挡、覆盖等文本图像。最明显的是文本占有比例严重不均衡问题,有些图像中文本占有比例较大,有些图像中文本占有比例较小。如图5(a)场景中文本占有比例非常少,图5(b)图中文本占有比例相对较大。

现有经典检测算法在进行文本定位时,偏向于图5(a,b)中某一类,为提高模型对场景图像中文本检测的泛化能力,本文采用多边形偏移文本域与文本边缘信息相结合的方式对文本进行检测与分离,

(a) Small text data graphs (b) Large text data graphs

图5 样本示例图

Fig.5 Sample diagram

从多边形偏移文本域的边缘像素向外扩张,以文本边缘信息为最大边界,清晰地分离出多个文本组件,采用多边形非极大值抑制算法[15]丢弃多余检测框,生成最终的文本检测标签。其中,多边形偏移文本域是在文本实例基础上按照一定缩放概率进行收缩,得到一个完全由本文像素组成的文本区域;文本边缘信息是文本实例区域的边界信息,图6给出了示例图。

(a) Initial image (b) Text instance area (c) Polygon offset text filed (d) Text edge map

图6 文本区域示例图

Fig.6 Example text area diagram

为计算多边形偏移文本域,本文采用Vatti裁剪算法[16]将初始文本图像进行裁剪,裁剪比例d_i为

$$d_i=\frac{\text{Area}(T)\times\left(1-r_i^2\right)}{\text{Perimeter}(T)} \tag{1}$$

式中:Area()为面积函数,T为多边形文本实例,r_i为第i个文本偏移域的缩放因子,Perimeter()为周长函数。

缩放因子r_i的计算过程为

$$r_i=1-\frac{(1-m)\times(n-i)}{n-1}\quad m\in(0,1) \tag{2}$$

式中:m为超参数最小缩放比例;n为获取多边形文本偏移域的数量,本文中$n=2$。

为了更加直观地展示多边形偏移文本域的生成,如图7所示,p_i为第i个多边形偏移文本域;d_i为在文本实例基础上进行的偏移距离;p_t为文本实例的边缘信息。

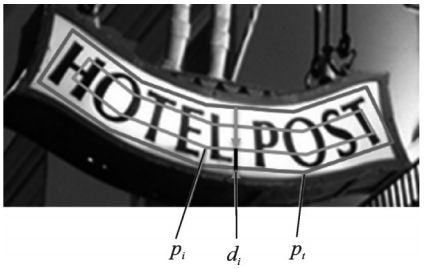

图7 多边形偏移文本域示意图

Fig.7 Diagram of polygon offset text field

2.3 损失函数

本文采用实例分割方式进行文本检测,因此可以当作二分类任务选择损失函数。目前比较受欢迎的损失函数有很多,如交叉熵损失、焦点损失、Dice系数损失等。Dice系数损失源于二分类任务,经过改进被称为Soft dice损失,改进过程中使用了目标掩码,利用目标掩码的大小归一化损失的效果,使得

Soft dice损失很容易从图像中具有较小空间表示的类中学习。而本文损失函数为式(3),损失主要由3部分组成,(1)预测文本边界框损失L_t;(2)生成多边形偏移文本域损失L_d;(3)像素损失L_p,指偏移文本域基于像素向外扩展过程产生的损失,λ_1、λ_2、λ_3为平衡3个损失设定的平衡系数。

$$\mathrm{Loss}=\lambda_1 L_t+\lambda_2 L_d+\lambda_3 L_p \tag{3}$$

文本边界框损失主要用来对场景中文本域非文本进行区分,选择Soft dice损失作为初始损失函数(式(4)),为对样本不均衡进行处理,在损失函数中引入在线难例挖掘(Online hard example mining, OHEM)[17],文本与非文本比例设置为3:1,得到的文本边界框损失为式(5)。

$$\mathrm{Dice\_loss}=\frac{2\times\sum_{x,y}P(x,y)\times G(x,y)}{\sum_{x,y}P^2(x,y)+\sum_{x,y}G^2(x,y)} \tag{4}$$

$$L_t=1-\mathrm{Dice\_loss}\left(P_t(x,y)\times M, G(x,y)\times M\right) \tag{5}$$

式中:$P_t(x,y)$为预测文本框P_t的像素点(x,y),$G(x,y)$为真实标签G的像素点(x,y),M为经过OHEM训练得到的掩码值。

多边形偏移文本域损失,本文也采用Soft dice损失,根据偏移文本域p_i中的像素点进行计算,计算过程为

$$L_d=\frac{2\times\sum_{x,y}P_d(x,y)\times G(x,y)}{\sum_{x,y}P_d^2(x,y)+\sum_{x,y}G^2(x,y)} \tag{6}$$

像素损失采用的是图像语义分割领域常用的逐像素交叉熵损失,见式(7)。

$$L_p=-\sum_{\mathrm{classes}}P_p(x,y)\times\lg\left(P_p(x,y)\right) \tag{7}$$

式中$P_p(x,y)$为扩展像素P_p的坐标(x,y)。

3 实验结果与分析

3.1 基准数据集

为了测试LAGNet的性能,选取国际文档分析与识别大会(International conference on document analysis and recognition,ICDAR)提供的比赛数据集ICDAR2015,该数据集以英文为主,大部分场景是街区、商场和路标等,复杂的背景加上文本的多样性非常具有挑战性。该数据集共有1 500张图,1 000张训练集,500张测试集,标签是4个坐标点顺时针排布。

为了进一步测试LAGNet在弯曲文本上的性能,选取由Liu等[15]构建的具有挑战性的曲线文本检测数据集SCUT-CTW1500,数据集中艺术字体较多,文本连接密集,场景多为广告牌、商标等。该数据集由1 000幅训练图像和500幅测试图像组成,标记方式为14个点的多边形,可以描述任意曲线文本的形状。

3.2 训练细节

训练模型过程中,没有预训练步骤,直接在ICDAR2015、CTW1500等数据集上从头开始训练,使用一块NVIDIA GTX 1080Ti GPU,反向传播采用的是Adam[18]和Adadelta优化算法,其计算梯度为0.9,梯度平方的运行平均值为0.999,权重衰减系数为1E−8。初始的学习率设定为1E−4,在随后的训练中每经过训练批次的1/3更新一次(乘以1E−1)。

深度学习采用Pytorch网络框架,在训练过程中,忽略数据集中的模糊标签,对输入的图片大小归一化处理为640像素×640像素。在测试阶段,参考基础网络PSENet,对于数据集ICDAR2015中的测试图片,大小归一化为2 240像素×2 240像素,最小卷积尺度设为0.4;对于数据集CTW1500中的测试图片,大小归一化为1 280像素×1 280像素,最小卷积尺度设为0.6;分类置信度设为0.9,将大于置信度的像素归为文本像素。

本文对LAGNet网络模型进行训练的时候,使用TensorboardX库对训练集准确率Accuracy与损失(Loss)进行可视化,方便观察网络模型的收敛情况。在常文本数据集ICDAR2015上的训练情况如图8所示,由图8可以看出,经过300个批次以后精度基本达到稳定,准确率与损失变化幅度都变小,为了确保准确性,在经过400批次时对学习率衰减,曲线图并未发生突变,到达600批次时终止训练。在弯曲文本数据集CTW1500上的训练情况如图9所示,参考前面的训练过程,当达到300批次以后模型基本趋于稳定状态,终止训练并采用当前模型进行测试。

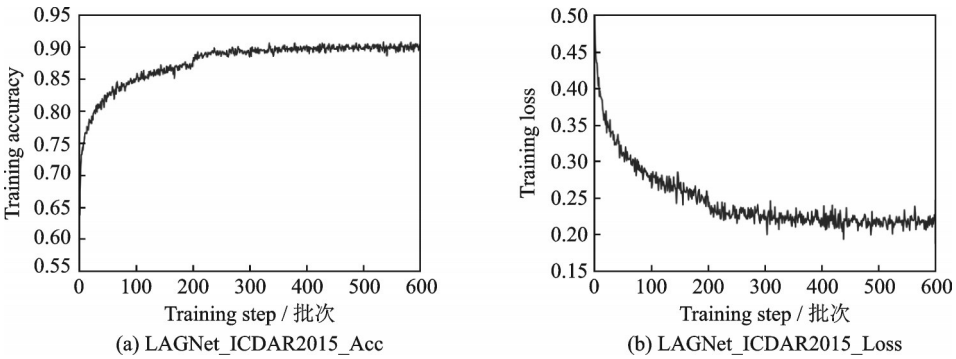

图8 ICDAR2015准确率与损失训练曲线图

Fig.8 ICDAR2015 accuracy and loss training curves

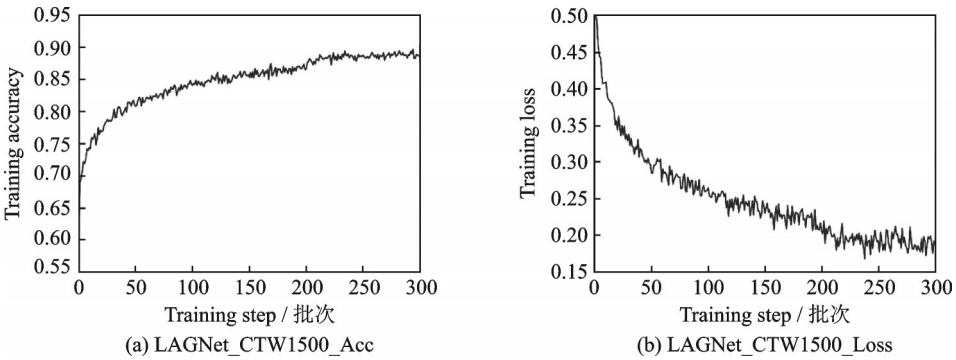

图9 CTW1500精确度与损失训练曲线图

Fig.9 CTW1500 accuracy and loss curves

3.3 实验分析

3.3.1 常文本检测分析

ICDAR2015数据集是非常典型的常文本数据集,数据集中的图片背景极其复杂,包含大量的无关信息还包含多种多样的字体。如表2所示,本文将LAGNet检测模型与目前经典的CTPN[6]、SegLink[8]等检测算法在准确率P、召回率R和F值3个指标上进行评估分析,同时为了验证本文特征提取模块改进的有效性,开展了消融对比实验,其中LAGNet-FLFF是指在LAGNet中只对恒等残差块进行设计,

实现局部细粒度特征融合;LAGNet-MGFF是指在LAGNet中仅实现多尺度全局特征融合。为了直观显示本文方法的有效性,图10展示了基础网络PSENet(Resnet-50)与LAGNet模型的部分测试对比图。

通过表2中的指标分析,本文方法在仅实现LAGNet-FLFF的情况下,与基础网络PSENet(Resnet-50)相比,召回率有所提高,因为该方式加强了对高层语义信息的提取,提高了对小文本的检测性能,减少了漏检误检的情况,但是在准确率上却没有基础网络模型好。在仅实现LAGNet-MGFF的情况下,本文方法的准确率有所提升,因为该方式将特征实现了全局共享,低层语义信息中包含了大量的文本特征,对大文本的检测性能提升很多,而数据集中大文本所占比例相对较大,对应的一些草木、护栏和铁轨等类似文本模块的检测也提升了,导致召回率降低很多。将两个模块进行合并以后,本文方法在ICDAR2015数据集上的准确率达到了84.2%。与CTPN、PSENet的经典算法的评价指标值[12]相比,本文方法准确率低于TextSnake 0.7%,召回率低于PiexlLink 0.2%,但综合指标F值高于TextSnake和PexelLink,同时,相比基础网络PSENet(Resnet-50)F值提高了2.2%。因此本文方法在ICDAR2015数据集上的综合检测性能有所提升。

表2 ICDAR2015数据集检测结果

Table 2 ICDAR2015 data set detection results

方法	准确率P	召回率R	F值
CTPN[6]	74.2	51.6	60.9
SegLink[8]	73.1	76.8	75.0
EAST[22]	83.2	78.3	80.7
PixelLink[10]	82.9	**81.7**	82.3
TextSnake[11]	**84.9**	80.4	82.6
PSENet[12]	81.5	79.7	80.6
LAGNet-FLFF	81.2	80.3	80.7
LAGNet-MGFF	82.5	78.3	80.3
LAGNet	84.2	81.5	**82.8**

图10中,第1、2行为基础网络模型PSENet(Resnet-50)的部分检测效果图。当场景中文本大小不一排列时,所检测的文本不够完整,边缘判断存在缺陷,例如第1行第1张张图所示。当场景中文本比

图10 ICDAR2015部分实验结果图

Fig.10 Some experimental results of ICDAR2015

例不均衡时,会出现漏检的情况,可能是浅层中提取的特征丢失引起的,如第1行第2张与第2行的第1、2图所示。当遇到场景中镜面反光的情况,如第1行第3张与第2行第3张图所示,对文本边缘的定位不准,也出现了漏检的情况。当场景图中文本所占比例非常小或出现遮挡的时候,文本检测出现漏检的现象,如第1行第4张与第2行第4张图所示。第3、4行为本文所提出的LAGNet网络模型对常文本的定位效果图,可以很直观地看出,经过改进后的网络模型在干扰较强的文本检测中有所提升。

3.3.2 弯曲文本检测分析

CTW1500是一个典型的弯曲文本数据集,该数据集中存在大量的艺术体、模糊小文本和类似文本干扰等因素。为了验证本文所提方法在自然场景中弯曲文本检测效率,基于CTW1500,本文将LAGNet检测模型与目前经典的文本检测算法在准确率P、召回率和F值3个指标上进行评估分析,并对本文改进模块进行了消融对比实验,分析结果如表3所示,部分检测效果图如图11所示。

表3 CTW1500数据集检测结果

Table 3 CTW1500 data set detection results

方法	准确率P	召回率R	F值
CTPN[6]	60.4	53.8	56.9
SegLink[8]	42.3	40.0	40.8
EAST[22]	78.7	49.1	60.4
TextSnake[11]	69.9	**85.3**	75.6
PSENet[12]	80.6	75.6	78.0
LAGNet-FLFF	82.3	73.6	77.7
LAGNet-MGFF	**84.5**	70.2	76.7
LAGNet	83.5	75.4	**79.2**

通过表3中的指标分析,在弯曲文本中检测中,本文方法仅实现LAGNet-MGFF的情况下,准确率提升幅度相对较大,达到了84.5%,比以上经典检测算法均高,但是与仅实现LAGNet-FLFF的情况相比,召回率降低了3.4%,当对两个模型合并以后本文方法的召回率达到了79.2%,在所测检测算法中最高。

(a) PSENet

(b) LAGNet

图11 CTW1500部分实验结果图

Fig.11 Some experimental results of CTW1500

图11文本中,第1、2行为基础网络模型PSENet(Resnet-50)的部分检测效果图,从图11中可知,PSENet对一些模糊的艺术体文本检测会出现遗漏或者检测不完全的情况,如第1行第1张与第2行第1、3、4图所示;对铁轨、围栏等一类的干扰因素无法排除,会发生误检,如第1行第2张与第2行第2张图所示;对遮挡的小文本区域定位不准,会出现漏检的情况,如第1行第3张图所示;对出现部分遮挡的文本检测不够完整,如第1行第4张图所示。第3、4行为本文所提出的LAGNet网络模型对弯曲文本的定位效果图,从图11可以看出在边缘模糊文本、强干扰以及部分遮挡小文本等自然场景文本检测中,本文模型的检测效果有一定的提升。

4 结束语

基于PSENet方法,本文提出自然场景文本检测网络LAGNet-MGFF,结合FPN结构思想,实现了多尺度特征全局共享,增强了网络的鲁棒性,同时在后处理模块中将多尺度偏移文本域与文本边缘信息相结合,提高对复杂场景下文本的定位准确性。在数据集ICADAR2015和CTW1500上开展的训练与测试表明,在召回率、F值等指标上,本文的检测模型优于基础网络PSENet(Resnet-50)等方法。

参考文献:

[1] 王德青, 吾守尔·斯拉木, 许苗苗. 场景文字识别技术研究综述[J]. 计算机工程与应用, 2020, 56(18): 1-15.
WANG Deqing, WUSHOUER·Silamu, XU Miaomiao. Review of research on scene text recognition technology[J]. Computer Engineering and Applications, 2020, 56(18): 1-15.

[2] MATAS J, CHUM O, URBAN M, et al. Robust wide-baseline stereo from maximally stable extremal regions[J]. Image and Vision Computing, 2004, 22(10): 761-767.

[3] EPSHTEIN B, OFEK E, WEXLER Y. Detecting text in natural scenes with stroke width transform[C]//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010: 2963-2970.

[4] ZITNICK C L, DOLLÁR P. Edge boxes: Locating object proposals from edges[C]//Proceedings of European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014: 391-405.

[5] 李国和, 乔英汉, 吴卫江, 等. 深度学习及其在计算机视觉领域中的应用[J]. 计算机应用研究, 2019, 36(12): 3521-3529, 3564.
LI Guohe, QIAO Yinghan, WU Weijiang, et al. Review of deep learning and its application in computer vision[J]. Application Research of Computers, 2019, 36(12): 3521-3529, 3564.

[6] TIAN Z, HUANG W, HE T, et al. Detecting text in natural image with connectionist text proposal network[C]//Proceedings of European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, Cham, 2016: 56-72.

[7] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6):1137-1149.

[8] SHI B, BAI X, BELONGIE S. Detecting oriented text in natural images by linking segments[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 2550-2558.

[9] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]// Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, Cham, 2016: 21-37.

[10] DENG D, LIU H, LI X, et al. Pixellink: Detecting scene text via instance segmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, Louisiana, USA: AAAI Press, 2018: 6773 - 6780.

[11] LONG S, RUAN J, ZHANG W, et al. Textsnake: A flexible representation for detecting text of arbitrary shapes[C]// Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, Cham, 2018: 20-36.

[12] WANG W, XIE E, LI X, et al. Shape robust text detection with progressive scale expansion network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 9336-9345.

[13] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 770-778.

[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, USA: IEEE, 2017: 2117-2125.

[15] LIU Y, JIN L, ZHANG S, et al. Curved scene text detection via transverse and longitudinal sequence connection[J]. Pattern Recognition, 2019, 90:337-345.

[16] VATTI B R. A Generic solution to polygon clipping[J]. Communications of the ACM, 1992, 35(7):56-63.

[17] SHRIVASTAVA A, GUPTA A, GIRSHICK R. Training region-based object detectors with online hard example mining [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 761-769.

[18] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]//Proceedings of International Conference on Learning Representations. San Diego, CA, USA, 2015:1-15.

作者简介:

李云洪(1972-),女,硕士,副教授,研究方向:数据智能化处理,E-mail: leeyunhong@126.com。

闫君宏(1993-),男,硕士研究生,研究方向:机器学习、图像处理,E-mail: 214384469@qq.com。

胡蕾(1980-),**通信作者**,女,博士,副教授,研究方向:图像处理、模式识别、数据挖掘等,E-mail:hulei@jxnu.edu.cn。

(编辑:陈珺)