计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100072

基于增强特征金字塔网络的场景文本检测算法

邵海琳[1]　季　怡[1]　刘纯平[1]　徐云龙[2]
1 苏州大学计算机科学与技术学院　江苏 苏州　215006
2 苏州大学应用技术学院　江苏 苏州 215300
(20184227001@stu.suda.edu.cn)

摘　要　场景文本检测有助于机器理解图像内容,在智能交通、场景理解和智能导航等领域应用广泛。现有的场景文本检测算法未充分利用高层语义信息和空间信息,限制了模型对复杂背景像素的分类能力和对不同尺度的文本实例的检测和定位能力。为解决上述问题,提出了一种基于增强特征金字塔网络的场景文本检测算法。该算法包括比率不变特征增强(Ratio Invariant Feature Enhanced,RIFE)模块和重建空间分辨率(Rebuild Spatial Resolution,RSR)模块。RIFE模块作为残差分支,增强了网络的高层语义信息传递,提高了分类能力,降低了误报率和漏检率。RSR模块重建多层特征分辨率,利用丰富的空间信息改进边界位置。实验结果表明,所提算法提升了在多方向文本数据集ICDAR2015、弯曲文本数据集Totaltext以及长文本数据集MSRA-TD500上的检测能力。

关键词:场景文本检测;特征金字塔网络;语义信息;空间信息;边界位置
中图法分类号　TP391

Scene Text Detection Algorithm Based on Enhanced Feature Pyramid Network

SHAO Hai-lin[1],JI Yi[1],LIU Chun-ping[1] and XU Yun-long[2]
1 School of Computer Science and Technology,Soochow University,Suzhou,Jiangsu 215006,China
2 Applied Technology College of Soochow University,Suzhou,Jiangsu 215300,China

Abstract　Scene text detection helps machines understand image content,and is widely used in the fields such as intelligent transportation,scene understanding,and intelligent navigation. Existing scene text detection algorithms do not make full use of high-level semantic information and spatial information,which limits the model's ability to classify complex background pixels and the ability to detect and locate text instances of different scales. In order to solve the above problems,a scene text detection algorithm based on enhanced feature pyramid network is proposed. The algorithm includes a RIFE (ratio invariant feature enhanced) module and a RSR (rebuild spatial resolution) module. As the residual branch,the RIFE module enhances the high-level semantic information transmission of the network,improves the classification ability,and reduces the false positive rate and the false negative rate. The RSR module reconstructs multi-layer feature resolution and uses rich spatial information to improve the boundary location. Experimental results show that the proposed algorithm improves the detection capabilities on the multi-directional text dataset ICDAR2015,the curved text dataset Totaltext,and the long text dataset MSRA-TD500.

Keywords　Scene text detection,Feature pyramid network,Semantic information,Spatial information,Boundary location

1　引言

随着深度学习的发展,场景文本检测性能大幅提升,可以广泛应用于银行卡识别、场景理解、智能助盲系统和无人驾驶等领域。目前,基于深度学习的场景文本检测算法可以分为3类:基于回归的算法、基于分割的算法和基于混合的算法[1]。

常见的基于回归的算法有TextBoxes算法[2]及其系列

到稿日期:2020-11-09　返修日期:2021-05-02　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61972059,61773272,61602332);江苏省高校自然科学基金重点项目(19KJA230001);吉林大学符号计算与知识工程教育部重点实验室项目(93K172016K08);江苏高校优势学科建设工程资助项目
This work was supported by the National Natural Science Foundation of China(61972059,61773272,61602332),Natural Science Foundation of Jiangsu Higher Education Institutions of China(19KJA230001),Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education,Jilin University(93K172016K08) and Priority Academic Program Development of Jiangsu Higher Education Institutions.
通信作者:刘纯平(cpliu@suda.edu.cn)

改进算法。TextBoxes 算法[2]通过修改 SSD(Single Shot Multibox Detector)[3]检测器中卷积核尺寸和检测框的长宽比，以适应长短不一的文本实例。TextBoxes++算法[4]改变了 TextBoxes 算法[2]中用矩形框描述的方式，提出用四边形或旋转矩形表示文本区域的思想，提升了对旋转文本检测的准确性。但这类算法需要设计锚框，在检测任意形状的文本时具有局限性。

基于分割的算法是在图像分割的基础上检测文本。PAN(Pixel Aggregation Network)算法[5]提取并融合不同层级的特征图，通过分割网络预测文本区域、核以及相似向量来适应对弯曲文本的检测，但是检测类文本像素时会造成误报。Richardson 等[6]通过自适应缩放文本实例，利用语义分割网络预测粗糙的文本区域和文本尺度，但易丢失较小的文本实例。DB(Differentiable Binarization)算法[7]利用可微分二值化算法简化后处理过程，提高了场景文本检测的效率，但对网络中的语义信息和空间信息利用不充分，限制了网络的分类能力和定位能力。尽管基于分割的算法在检测任意形状的文本时具有优势，但由于缺乏足够的上下文信息会造成误报或漏检。

基于混合的算法是上述两类算法的结合。Dai 等[8]利用文本特征增强模块和注意力金字塔感兴趣区域池化机制，提高特征的表达能力，最后通过包围框感知上下文文本分割模块修正边界。Chen 等[9]设计了高分辨率主干网络，提升了场景文本检测性能。由于这类算法是回归与分割两类算法优点的结合，整体上提高了场景文本检测能力，但网络结构较为复杂。

已有场景文本检测算法虽然提高了检测能力，但在复杂背景纹理(见图 1(a))、小文本实例(见图 1(c))和文本边界定位精准性(见图 1(e))等方面仍存在不足，图 1(b)、图 1(d)、图 1(f)给出了真值标注结果。

(a)帽子花纹被误报成文本　(b)文本检测真值标注

(c)小尺度文本漏检　(d)小尺度文本真值标注

(e)文本边界定位粗糙　(f)文本边界真值标注

图 1　场景文本检测中存在的问题

Fig. 1　Problems in scene text detection

针对复杂背景像素误报、小尺度文本漏检和边界定位不准确的问题，本文提出一种基于增强特征金字塔网络(Enhanced Feature Pyramid Network，EFPN)的场景文本检测算法，其整体框架如图 2 所示。

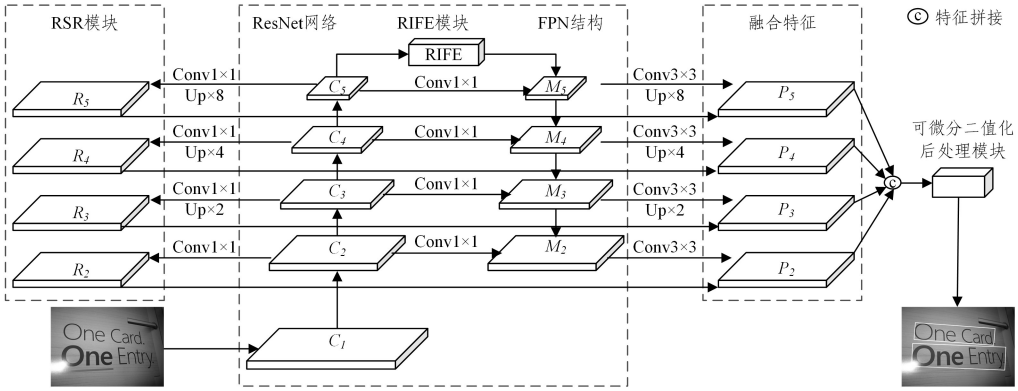

图 2　EFPN 模型结构

Fig. 2　Structure of EFPN model

该算法以 DB 算法[7]为基模型，增加了残差连接 RIFE 模块和恢复原始特征空间分辨率 RSR 模块。本文的贡献如下：

(1)增加残差连接 RIFE 模块。通过充分利用网络高层语义信息，增强特征金字塔结构(Feature Pyramid Network，FPN)[10]自上而下的高层语义信息流动，提高了网络的分类能力，从而降低了基模型中非文本像素造成的误报和小尺度文本实例漏检的现象。

(2)增加重建特征分辨率的 RSR 模块。通过恢复原始特征的空间分辨率，增强了对边界信息的敏感性，提高了定位文本边界的精细度，解决了基模型对文本边界定位不准确的问题。

(3)在多方向文本数据集 ICDAR2015[11]、弯曲文本数据集 Totaltext[12]和长文本数据集 MSRA-TD500[13]上的实验证明了所提算法的合理性和有效性。

2　相关工作

本节重点阐述与提出的 EFPN 场景文本检测算法最为相关的基于分割的场景文本检测算法和 DB 基模型框架。

基于分割的场景文本检测算法一般采用 FCN(Fully Convolutional Network)[14]，FPN[10]等网络提取多尺度特征，给每个像素预测分类标签，适用于检测任意形状文本。

TextSnake 算法[15]用半径、方向各不相同的有序重叠圆盘，沿着文本中心线分布以表示不同形状的文本，提升了弯曲文本的检测效果。LOMO(LOok More than Once)算法[16]通过设计形状表达模块，在多次迭代优化模块的基础上，利用全卷积网络学习的文本区域、文本中心线以及边界偏移属性来描述任意形状的文本。

自然场景文本检测不仅要考虑文本形状变化，还要考虑文本布局方式给检测带来的巨大挑战。PSENet(Progressive Scale Expansion Network)算法[17]利用相邻文本实例最小核之间的距离相对较大的特点，通过渐近扩展尺度后处理算法将最小的文本核逐步扩张至最大的文本核。PSENet 算法[17]采用“先到先得”策略，将像素分配给唯一文本实例，提高了邻近文本的检测能力。CRAFT(Character Region Awareness For Text Detection)算法[18]利用弱监督方式训练一个字符级别的场景文本检测器，通过预测区域分数定位单个字符和预测亲和力分数反映字符与字符间的关系，并依据属于相同文本实例的字符间亲和力分数往往高于属于不同文本的字符亲和力分数这一特点，灵活分割相近的文本。

DB 基模型[7]是 2020 年提出的一个基于分割的算法，其网络结构如图 3 所示。输入图像经过特征提取网络预测分割图和阈值图，根据可微分二值化算法生成近似二值化图，从而区分文本实例和背景区域。传统算法通过设置固定阈值进行文本实例与背景的区分，不能在训练过程中优化。DB 基模型[7]提出的可微分二值化算法，可以在训练过程中自适应地学习图像上每个位置的阈值，提高了对任意形状场景文本的检测效率。但是 DB 基模型[7]通过卷积直接减少最高层特征的通道数目，造成高层语义信息损失，导致类文本像素误报和小尺度文本漏检。同时，其由于多次下采样丢失了空间信息，造成文本实例边界定位粗糙。

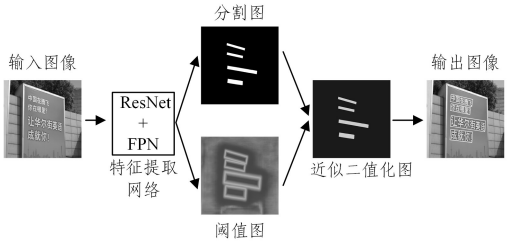

图 3 DB 网络结构

Fig. 3 DB network structure

3 EFPN 模型

鉴于赋予低层特征更多的语义信息或者赋予高层特征更多的空间信息可以提高网络的特征表达能力[19]，因此，以 DB 算法[7]为基本框架，增强语义信息和空间位置信息进行场景文本检测是一个自然的思路，进而本文提出了 EFPN 算法，其结构如图 2 所示。

语义信息增强部分，用 ResNet50 提取原始特征$\{C_2, C_3, C_4, C_5\}$，采用 1×1 卷积降低通道数目进行横向连接，自上而下地传递语义信息构造$\{M_2, M_3, M_4, M_5\}$。顶层特征 C_5 经过 RIFE 模块增强网络的语义信息 M_5。

语义增强特征 M_i 的计算过程如式(1)所示：

$$M_i=\begin{cases}F_{1\times1}(C_5)+RIFE(C_5), & i=5\\ F_{1\times1}(C_i)+Up(M_{i+1},2), & i=2,3,4\end{cases} \tag{1}$$

其中，$F_{1\times1}(C_i)$是对 C_i 进行 1×1 卷积操作，$Up(M_{i+1},2)$表示对 M_{i+1}进行 2 倍双线性上采样。

空间信息增强部分，利用 RSR 模块重建原始特征$\{C_2, C_3, C_4, C_5\}$的空间分辨率。空间信息增强特征 R_i 的计算过程如式(2)所示：

$$R_i=Up(F_{1\times1}(C_i),2^{i-2}),i=2,3,4,5 \tag{2}$$

其中，$Up(F_{1\times1}(C_i),2^{i-2})$表示对 $F_{1\times1}(C_i)$进行 2^{i-2}倍双线性上采样。

融合模块则将 M_i 与 R_i 对应层级特征相加，构造融合特征$\{P_2, P_3, P_4, P_5\}$。融合特征 P_i 的计算如式(3)所示：

$$P_i=Up(F_{3\times3}(M_i),2^{i-2})+R_i,i=2,3,4,5 \tag{3}$$

其中，$F_{3\times3}(M_i)$是对 M_i 进行 3×3 卷积操作。

最后拼接多层融合特征并进行可微分二值化，得到最终的检测结果。

3.1 RIFE 模块

语义上下文信息在场景文本检测中有着重要的作用，研究表明缺乏上下文信息的指导会造成误报[20]。尽管特征金字塔层级结构隐式地学习了上下文信息，但是其最高层语义特征经过横向连接减少通道数后直接自上而下传递，导致网络未完全利用提取的高层特征[21]。

为了降低复杂背景对文本检测的影响，本文提出 RIFE 模块(见图 4)以增强高层语义信息。该模块首先将最高层语义特征图 C_5 输入，经过 3 个并行分支，自适应最大值池化得到 3 个不同尺度的特征图($H_k\times W_k, k=1,2,3$)，并且通过 1×1 卷积将通道数由 2048 维降至 256 维，然后上采样至与 C_5 特征图大小一致，最后将 3 个分支的特征直接相加并经过 Relu 激活函数以增强网络高层语义信息。

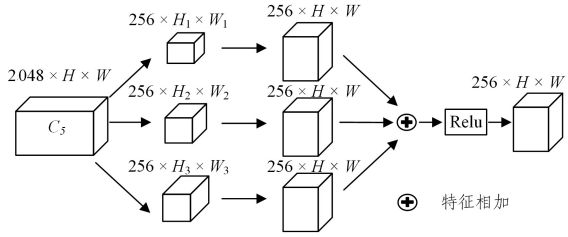

图 4 RIFE 模块的结构图

Fig. 4 Structure of RIFE module

RIFE 模块的计算过程如式(4)所示：

$$RIFE(C_5)=\mathrm{Relu}\left[\sum_{j=1}^{3}Up(F_{1\times1}(Pool(C_5,\alpha_j)),\frac{1}{\alpha_j})\right] \tag{4}$$

其中，$Pool(C_5,\alpha_j)$是对 C_5 进行 α_j 倍下采样，实验中将 α_1，α_2，α_3 分别设置为 0.1，0.2 和 0.3。$F_{1\times1}(Pool(C_5,\alpha_j))$是对 $Pool(C_5,\alpha_j)$进行 1×1 卷积操作，$Up(F_{1\times1}(Pool(C_5,\alpha_j)),\frac{1}{\alpha_j})$对 $F_{1\times1}(Pool(C_5,\alpha_j))$进行$\frac{1}{\alpha_j}$倍双线性上采样。

3.2 RSR 模块

空间信息对定位文本实例边界非常重要，可以有效提升对场景文本的检测能力。因此，EFPN 模型设计了 RSR 模块以充分利用原始空间关系，缩短空间信息传递的距离，以有效增强融合特征的可区分度，提高文本边界的定位准确性。

该模块首先用 1×1 卷积使$\{C_2, C_3, C_4, C_5\}$通道数目均

为 256 维，然后将$\{C_3, C_4, C_5\}$双线性上采样至与 C_2 分辨率一致，即$\{R_2, R_3, R_4, R_5\}$，最后$\{R_2, R_3, R_4, R_5\}$与 FPN 结构中对应层级上采样之后的特征相加，得到$\{P_2, P_3, P_4, P_5\}$融合特征。由于利用了细致的空间信息，因此降低了无用的上下文信息对文本边界定位的影响。

4 实验结果与分析

为了验证所提算法的可行性和有效性，分别进行了消融实验、模型复杂度分析、检测改进性能分析和对比实验。实验在 3 个数据集上展开，并采用准确率、召回率和 F 值来评估算法性能。

4.1 数据集和实验环境

算法验证部分使用了 5 个数据集，其中 SynthText 数据集[22]和 ICDAR2017 MLT 数据集[23]用于预训练模型，ICDAR2015 数据集[11]、Totaltext 数据集[12]、MSRA-TD500 数据集[13]用于模型微调和算法性能验证。

SynthText 数据集[22]大约包含 80 万张图像，由 800 万个合成单词实例组成。EFPN(SynthText)表示采用 SynthText 数据集进行预训练。ICDAR2017 MLT 数据集[23]涉及 9 种语言，包含训练集 7200 张、验证集 1800 张和测试集 9000 张，文本实例用四边形标注。EFPN(IC17 MLT)表示采用 ICDAR2017 MLT 数据集进行预训练。

ICDAR2015 数据集[11]包含 1 500 张图像，其中训练集 1 000 张，测试集 500 张。图像背景比较复杂，文本尺度变化大，图像模糊，所有英文文本实例采用四边形框进行单词级标注。Totaltext 数据集[12]中文本形状是弯曲的，其训练集包含 1 255 张图像，测试集包含 300 张图像，仅包含英文文本，采用单词级标注。MSRA-TD500 数据集[13]规模较小，仅包含 500 张自然场景图像，其中 300 张作为训练集，200 张作为测试集。文本语言包括中文和英文，采用文本行级标注，文本尺度变化大且文本是任意方向的。

实验平台是 NVIDIA GeForce RTX 2080Ti、CUDA10.0 版本，使用 2 个 GPU 进行训练。初始学习率为 0.007，训练采用随机梯度下降优化算法。

4.2 消融实验

为了消除预训练模型对实验结果的影响，在 3 个公开数据集上分别采用了基于 SynthText 数据集的预训练模型和基于 ICDAR2017 MLT 数据集的预训练模型。表 1－表 3 分别列出了 ICDAR2015 数据集、Totaltext 数据集和 MSRA-TD500 数据集上的消融实验结果。考虑到实验环境的不同，分别给出了 DB 基模型[7]在两个预训练模型上的本地复现的实验数据。

表 1 ICDAR2015 数据集上的消融实验

Table 1 Ablation experiments on ICDAR2015 dataset

(单位：%)

算法	基于 SynthText 数据集预训练			基于 ICDAR2017 MLT 数据集预训练		
	准确率	召回率	F 值	准确率	召回率	F 值
基模型(论文)	88.2	82.7	85.4	—	—	—
基模型（本地复现）	89.8	77.3	83.1	88.9	80.3	84.4
基模型＋RIFE	90.0	77.8	83.4	87.6	82.6	85.0
基模型＋RSR	89.4	78.1	83.4	91.2	79.9	85.2
EFPN	89.9	78.4	83.7	89.2	82.0	85.5

表 2 Totaltext 数据集上的消融实验

Table 2 Ablation experiments on Totaltext dataset

(单位：%)

算法	基于 SynthText 数据集预训练			基于 ICDAR2017 MLT 数据集预训练		
	准确率	召回率	F 值	准确率	召回率	F 值
基模型(论文)	87.1	82.5	84.7	—	—	—
基模型（本地复现）	87.6	81.2	84.2	87.6	82.7	85.1
基模型＋RIFE	88.5	81.7	84.9	89.3	82.2	85.6
基模型＋RSR	88.0	81.8	84.8	87.3	84.1	85.7
EFPN	88.5	82.1	85.1	89.5	82.7	85.9

表 3 MSRA-TD500 数据集上的消融实验

Table 3 Ablation experiments on MSRA-TD500 dataset

(单位：%)

算法	基于 SynthText 数据集预训练			基于 ICDAR2017 MLT 数据集预训练		
	准确率	召回率	F 值	准确率	召回率	F 值
基模型(论文)	91.5	79.2	84.9	—	—	—
基模型（本地复现）	89.0	80.8	84.7	93.9	81.4	87.2
基模型＋RIFE	89.8	81.3	85.3	88.3	87.1	87.7
基模型＋RSR	90.5	81.4	85.7	90.4	85.9	88.1
EFPN	91.4	82.0	86.4	92.9	85.1	88.8

从表 1 可以看出，在 ICDAR2015 数据集上，基于 SynthText 数据集进行预训练，加入 RIFE 模块后，准确度、召回率和 F 值分别超过 DB 基模型[7]本地复现结果约 0.2%，0.5%和 0.3%；加入 RSR 模块，召回率和 F 值分别提高约 0.8%和 0.3%；同时加入两个模块，召回率提升了 1.1%，F 值提高了 0.6%。基于 ICDAR2017 MLT 数据集进行预训练，加入 RIFE 模块后，与本地复现结果相比，召回率和 F 值分别提高了约 2.3%和 0.6%；加入 RSR 模块后，准确率和 F 值分别提高了约 2.3%和 0.8%；同时引入两个模块，准确率约提升 0.3%，召回率提升了 1.7%，F 值提高了 1.1%。

从表 2 可以看出，在 Totaltext 数据集上，基于 SynthText 数据集进行预训练，与本地复现结果相比，引入 RIFE 模块后，准确率、召回率和 F 值分别提高了约 0.9%，0.5%和 0.7%；引入 RSR 模块后，准确率、召回率和 F 值分别提高了约 0.4%，0.6%和 0.6%；同时引入两个模块，在 3 个评价指标上均超过本地复现结果约 0.9%。与基模型公开的实验结果相比，EFPN 在准确率和 F 值上分别提高约 1.4%和 0.4%。基于 ICDAR2017 MLT 数据集进行预训练，加入 RIFE 模块后，与相同预训练模型下本地复现结果相比，准确率和 F 值分别提高了约 1.7%和 0.5%；加入 RSR 模块后，召回率和 F 值分别提高了约 1.4%和 0.6%；同时引入两个模块，在召回率相当的情况下，准确率约提升 1.9%，F 值提高了 0.8%。

从表 3 可以看出，在 MSRA-TD500 数据集上，基于 SynthText 数据集的预训练模型，与本地复现结果相比，在引入 RIFE 模块后，准确率、召回率和 F 值分别提高了约 0.8%，0.5%和 0.6%；引入 RSR 模块后，准确率、召回率和 F 值分别提高了约 1.5%，0.6%和 1%；同时引入两个模块，准确率、召回率和 F 值分别提高了约 1.5%，0.6%和 1%。与基模型论文中的结果相比，在准确率相当的情况下，召回率和 F 值分别提高了 2.8%和 1.5%。基于 ICDAR2017 MLT 数据集进行预训练，与同一预训练模型下本地复现结果相比，引入 RIFE 模块后，F 值提高了约 0.5%；引入 RSR 模块后，F 值

提高了0.9%;同时引入两个模块,召回率和F值分别提升了约3.7%和1.6%。

在3个数据集上的消融实验证明了RIFE模块可以较好地引入高层语义信息,增强上下文信息,RSR模块可以有效引入空间信息,提升文本的边界定位。EFPN模型结合两个模块,减少了类文本像素误报和小文本漏检,并修正边界定位,对多方向文本、弯曲文本以及长文本具有鲁棒性。

4.3 模型复杂度分析

为了研究EFPN模型的空间复杂度和时间复杂度,表4统计了基模型、基模型加入RIFE模块、基模型加入RSR模块和EFPN模型的参数量和计算量。基模型加入RIFE模块后参数量和计算量分别仅增加了1.64 MB和0.06 GB,基模型加入RIFE模块后参数量和计算量分别仅增加了0.25 MB和0.8 GB。虽然EFPN模型相比DB基模型[7]增加了1.89 MB的参数量和0.85 GB的计算量,但是降低了类文本像素的误报率、小尺度文本的漏检率以及提高了边界定位精细度,在3个标准数据集上F值均有所提高。

表4 EFPN模型的复杂度

Table 4 Complexity of EFPN model

算法	参数量/MB	计算量/GB
基模型	28.85	37.71
基模型+RIFE	30.49 (↑1.64)	37.77 (↑0.06)
基模型+RSR	29.10 (↑0.25)	38.51 (↑0.8)
EFPN	30.74 (↑1.89)	38.56 (↑0.85)

4.4 模型场景文本检测性能实验结果与分析

为了进一步分析所提模型在类文本像素区域、小尺度文本以及边界定位方面的改进性能,分别进行了类文本像素区域检测、小尺度文本检测和边界定位的实验结果分析。

4.4.1 类文本像素区域的检测性能分析

为了探究EFPN模型对类文本像素区域的检测性能,图5给出了基于SynthText预训练模型在Totaltext数据集上的部分实验结果。从图5(b)和图5(c)可以发现,基模型和本地复现均错误地将类文本像素(王冠图案)分类成文本实例,在图5(d)中,王冠图案被正确分类为背景,表明EFPN模型有效降低了类文本像素的误报率。

(a)真值 (b)基模型 (c)本地复现 (d)EFPN (SynthText)

图5 Totaltext数据集上真值、基模型、本地复现以及EFPN模型的可视化结果

Fig. 5 Visualization results of ground truth, baseline, our reimplement, and EFPN model on Totaltext dataset

4.4.2 小尺度文本检测性能分析

基于MS COCO的分类标准[24],将像素面积不大于32^2的定义为小文本,像素面积不小于96^2的定义为大文本,介于这两个边界之间的定义为中尺度文本。

图6给出了EFPN模型在ICDAR2015数据集上的检测结果示例,图6(a)是真值可视化结果。从可视化的实例中可以看出,提出的EFPN模型在检测小尺度文本上具有优势,如图6(d)所示。图7给出了该模型在ICDAR2015数据集上的统计分析结果,从图7(a)和图7(b)可以看出,提出的EFPN模型在大尺度和中尺度文本上与本地复现的DB基模型[7]性能相当,但是在小尺度文本检测方面,相比本地复现的DB基模型[7]有较为明显的提升,如图7(c)所示。这表明EFPN模型降低了小尺度文本的漏检率,提高了多方向场景文本的检测能力。

(a)真值 (b)基模型 (c)本地复现 (d)EFPN (SynthText)

图6 ICDAR2015数据集上真值、基模型、本地复现以及EFPN模型的可视化结果

Fig. 6 Visualization results of ground truth, baseline, our reimplement, and EFPN model on ICDAR2015 dataset

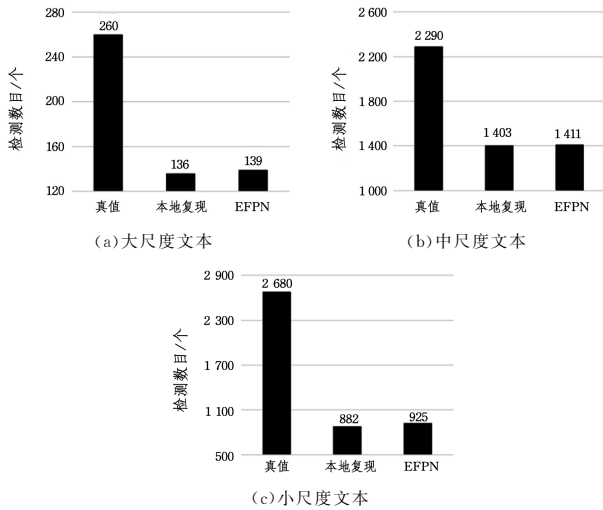

图7 ICDAR2015数据集上基模型与EFPN模型多尺度文本检测的个数

Fig. 7 Number of multi-scale text detection of baseline and EFPN model on ICDAR2015 dataset

4.4.3 边界定位性能分析

图8给出了基于SynthText数据集的预训练模型在MSRA-TD500数据集上的边界定位结果。

(a)真值 (b)基模型 (c)本地复现 (d)EFPN (SynthText)

图8 MSRA-TD500数据集上边界定位可视化结果

Fig. 8 Visualization of boundary location on MSRA-TD500 dataset

图 8(a)是真值可视化结果。从图 8(b)、图 8(c)可以看出，DB 基模型[7]对文本边界像素定位不准确。从图 8(d)可以看出，EFPN 模型检测的文本框边界与真值几乎一致。

为进一步分析 RSR 模块中各特征层对边界定位的作用，自顶层特征图开始，逐层通过 RSR 模块恢复分辨率并直接与相应层级的金字塔特征图融合，实验结果如图 9 所示。在只有 RSR 模块的情况下，当恢复较低层的原始特征图的空间分辨率后，F 值从 84.6%提升至 85.7%，表明 RSR 模块恢复了较低层的原始特征分辨率，增强了融合特征的空间信息，提高了边界定位的准确性，进而提高了对场景文本的检测能力。

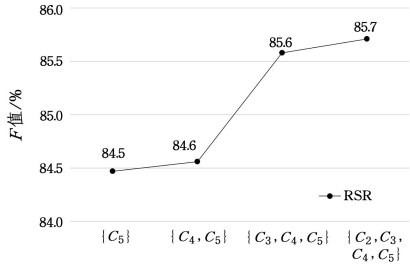

图 9 RSR 模块各层特征对检测结果的影响

Fig. 9 Influence of each feature map of RSR module on detection results

4.5 对比实验

为了验证 EFPN 的有效性，在 3 个数据集上将其与其他算法进行对比，实验结果如表 5－表 7 所列。EFPN(SynthText)和 EFPN(IC17 MLT)分别表示采用 SynthText 数据集预训练和采用 ICDAR2017 MLT 数据集预训练。表中，“*”表示需要使用额外的数据集。

表 5 ICDAR2015 数据集上的结果

Table 5 Results on ICDAR2015 dataset

(单位：%)

算法	准确率	召回率	F 值
CTPN(2016)[25]	74.2	51.6	60.9
EAST(2017)[26]	83.6	73.5	78.2
TextSnake(2018)*[15]	84.9	80.4	82.6
RRPN(2018)[27]	82	73	77
PixelLink(2018)[28]	82.9	81.7	82.3
RRD(2018)*[29]	85.6	79.0	82.2
Corner(2018)*[30]	94.1	70.7	80.7
PAN(2019)*[5]	84.0	81.9	82.9
Dai 等(2019)[8]	86.2	82.7	84.4
PSENet(2019)*[17]	86.9	84.5	85.7
CRAFT(2019)*[18]	89.8	84.3	86.9
Richardson 等(2020)[6]	85.4	83.1	84.2
Liu 等(2020)*[31]	80	58	67
SPN(2019)*[32]	86.6	82.1	84.3
SRPN(2020)[33]	92.0	79.7	85.4
SDnet(2020)[34]	82.9	80.2	81.6
Zhang 等(2020)*[35]	83.2	87.7	85.4
DB(2020)*[7]	88.2	82.7	85.4
EFPN(SynthText)	89.9	78.4	83.7
EFPN (IC17 MLT)	89.2	82.0	85.5

表 6 MSRA-TD500 数据集上的结果

Table 6 Results on MSRA-TD500 dataset

(单位：%)

算法	准确率	召回率	F 值
EAST(2017)[26]	87.28	67.43	76.08
SegLink(2017)*[36]	86	70	77
TextSnake(2018)*[15]	83.2	73.9	78.3
RRPN(2018)[27]	82	68	74
RRD(2018)*[29]	87	73	79
Corner(2018)*[30]	87.6	76.2	81.5
PAN(2019)*[5]	84.4	83.8	84.1
CRAFT(2019)*[18]	88.2	78.2	82.9
SRPN(2020)[33]	84.9	77.0	80.7
Zhang 等(2020)*[35]	81.4	75.3	78.2
DB(2020)*[7]	91.5	79.2	84.9
EFPN(SynthText)	91.4	82.0	86.4
EFPN(IC17 MLT)	92.9	85.1	88.8

表 7 Totaltext 数据集上的结果

Table 7 Results on Total text dataset

(单位：%)

算法	准确率	召回率	F 值
TextSnake(2018)*[15]	82.7	74.5	78.4
PAN(2019)*[5]	89.3	81.0	85.0
Dai 等(2019)*[8]	84.6	78.6	81.5
LOMO(2019)*[16]	87.6	79.3	83.3
PSENet(2019)*[17]	84	78	80.9
CRAFT(2019)*[18]	87.6	79.9	83.6
Wang 等(2019)[37]	80.9	76.2	78.5
TextField(2019)*[38]	81.2	79.9	80.6
SDnet(2020)[34]	82.3	76.5	79.3
DB(2020)*[7]	87.1	82.5	84.7
EFPN(SynthText)	88.5	82.1	85.1
EFPN(IC17 MLT)	89.5	82.7	85.9

在多方向文本数据集 ICDAR2015 上，本文算法与其他算法的对比结果如表 5 所列。EFPN(SynthText)在准确率和 F 值上均超过 CTPN(Connectionist Text Proposal Network)算法[25]、RRPN(Rotated Regional Proposal Network)算法[27]、RRD(Rotation-sensitive Regression Detector)算法[29]等基于回归的算法。EFPN(SynthText)在准确率、召回率和 F 值上分别超过 EAST(Efficient and Accuracy Scene Text)算法[26]约 6.3%，4.9%和 5.7%。Corner 算法[30]会将两个邻近的文本预测为一个文本实例，造成检测不准确。SPN 算法(Short Path Network)[32]对弯曲文本实例的鲁棒性较差。当第一阶段预测的候选区域只包含文本实例的一部分时，SRPN(Scale-based Region Proposal Network)算法[33]在第二阶段无法正确预测整个文本实例的边界。Dai 等[8]不能消除类文本像素的干扰，造成误报。与 EAST 算法[26]、Corner 算法[30]、SPN 算法[32]、SRPN 算法[33]和 Dai 等[8]这些基于混合的算法相比，EFPN 模型充分利用语义信息提高了类文本像素的分类准确性，降低了背景像素对小尺度文本的干扰，利用丰富的空间信息有助于提高文本实例的边界定位能力。

由于基于分割的 PSENet 算法[17]、CRAFT 算法[18]、Richardson 等[6]和 Zhang 等[35]的输入长边分辨率与本文不同，为了保证对比结果的公平性，分别将图像长边分辨率设置为 1440 像素、1920 像素和 2240 像素，图 10 给出了 EFPN 模型在不同分辨率下与其他模型的对比结果。当输入图像长边

分辨率设置为 1 440 像素时，与相同分辨率的 Richardson 等[6]提出的模型相比，EFPN 模型的 F 值高出约 1.9%。在图像长边分辨率设为 1 920 像素时，EFPN 模型的 F 值比 Zhang 等[35]高出 1.5%。在图像长边分辨率设为 2 240 像素时，EFPN 模型的 F 值比 CRAFT 算法[18]和 PSENet 算法[17]分别高出 0.3%和 1.5%。上述结果表明，EFPN 模型在多方向文本数据集上取得了很好的检测结果。

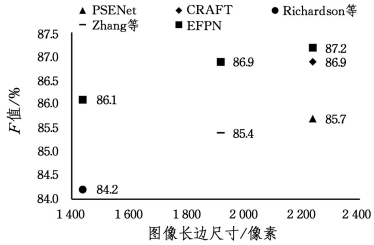

图 10 EFPN (IC17 MLT)与其他模型在 ICDAR2015 数据集上不同分辨率下的比较结果

Fig. 10 Comparison results of EFPN(IC17 MLT) and other models at different resolutions on ICDAR2015 dataset

在长文本数据集 MSRA-TD500 上，本文算法与其他算法的对比结果如表 6 所列。SegLink 算法[36]、RRPN 算法[27]、RRD 算法[29]等基于回归的算法，以及 EAST 算法[26]、Corner 算法[30]、SRPN 算法[33]等基于混合的算法由于感受野不足，在一定程度上限制了检测长文本的能力。EFPN(SynthText)和 EFPN(IC17 MLT)在 3 个评价指标上均超过基于分割的 TextSnake 算法[15]、CRAFT 算法[18]和 Zhang 等[35]算法。PAN 算法[5]模型无法避免复杂背景噪声的干扰，易造成误报。EFPN 模型在长文本数据集上具有鲁棒性，有效降低了背景像素的误报率和长文本的漏检率。在 Totaltext 弯曲文本数据集上将本文算法与其他算法进行比较，结果如表 7 所列。Wang 等[37]提出的模型自适应地确定多边形文本区域顶点个数，采用递归神经网络增加了端到端任务的时间开销。EFPN (SynthText)和 EFPN (IC17 MLT)在 3 个评价指标上均超过 TextSnake 算法[15]、LOMO 算法[16]、PSENet 算法[17]、CRAFT 算法[18]和 TextField 算法[38]等基于分割的算法。EFPN (IC17 MLT)在准确率、召回率和 F 值指标上分别超过 PAN 算法[5]约 0.2%，1.7%和 0.9%。上述结果表明，EFPN 模型在像素级别预测分类标签，可以适应任意形状的弯曲文本，丰富的语义信息提高了网络的分类能力，空间信息有助于确定更精细的边界。

结束语 本文提出了一种基于增强特征金字塔网络的场景文本检测算法。RIFE 模块学习了顶层语义的多尺度信息，增强了特征金字塔自上而下的语义信息传递，提高了类文本像素和文本实例的分类准确性。RSR 模块恢复高层特征分辨率，缩短了空间信息传递的距离，提高了网络的边界定位准确性。通过与其他模型对比，EFPN 模型在 3 个标准公开数据集上均取得了较好的检测结果。为了满足复杂场景下文本检测的需求，需要进一步研究如何检测字符间距较大的行级文本以及如何提高模糊文本的检测效果。

参 考 文 献

[1] RAISI Z, NAIEL M A, FIEGUTH P, et al. Text Detection and Recognition in the Wild: A Review[J]. arXiv: 2006.04305, 2020.

[2] LIAO M, SHI B, BAI X, et al. Textboxes: A fast text detector with a single deep neural network[J]. arXiv: 1611.06779, 2016.

[3] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]// European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[4] 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.

[5] WANG W, XIE E, SONG X, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network [C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 8440-8449.

[6] RICHARDSON E, AZAR Y, AVIOZ O, et al. It's All About The Scale-Efficient Text Detection Using Adaptive Scaling [C]// The IEEE Winter Conference on Applications of Computer Vision. 2020: 1844-1853.

[7] LIAO M, WAN Z, YAO C, et al. Real-Time Scene Text Detection with Differentiable Binarization[C]// AAAI. 2020: 11474-11481.

[8] DAI P, ZHANG H, CAO X. Deep multi-scale context aware feature aggregation for curved scene text detection[J]. IEEE Transactions on Multimedia, 2019, 22(8): 1969-1984.

[9] CHEN M M, XU J H. Scene text detection model based on high resolution convolutional neural networks[J]. Computer Applications and Software, 2020, 37(10): 138-144.

[10] 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. 2017: 2117-2125.

[11] 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). IEEE, 2015: 1156-1160.

[12] 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). IEEE, 2017: 935-942.

[13] YAO C, BAI X, LIU W, et al. Detecting texts of arbitrary orientations in natural images[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 1083-1090.

[14] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[15] LONG S, RUAN J, ZHANG W, et al. Textsnake: A flexible representation for detecting text of arbitrary shapes[C]// Procee-

dings of the European Conference on Computer Vision (ECCV). 2018:20-36.

[16] ZHANG C,LIANG B,HUANG Z,et al. Look more than once: An accurate detector for text of arbitrary shapes[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:10552-10561.

[17] 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. 2019:9336-9345.

[18] BAEK Y,LEE B,HAN D,et al. Character region awareness for text detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:9365-9374.

[19] ZHANG Z,ZHANG X,PENG C,et al. Exfuse:Enhancing feature fusion for semantic segmentation[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018:269-284.

[20] 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. 2019, 33: 9038-9045.

[21] GUO C,FAN B,ZHANG Q,et al. Augfpn:Improving multi-scale feature learning for object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:12595-12604.

[22] 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. 2016:2315-2324.

[23] NAYEF N,YIN F,BIZID I,et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt[C]// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017:1454-1459.

[24] LIN T Y,MAIRE M,BELONGIE S,et al. Microsoft coco:Common objects in context[C]// European Conference on Computer Vision. Cham:Springer,2014:740-755.

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

[26] ZHOU X,YAO C,WEN H,et al. East:an efficient and accurate scene text detector[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017:5551-5560.

[27] MA J,SHAO W,YE H,et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia,2018,20(11):3111-3122.

[28] DENG D,LIU H,LI X,et al. Pixellink:Detecting scene text via instance segmentation[J]. arXiv:1801.01315,2018.

[29] LIAO M,ZHU Z,SHI B,et al. Rotation-sensitive regression for oriented scene text detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:5909-5918.

[30] LYU P,YAO C,WU W,et al. Multi-oriented scene text detection via corner localization and region segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7553-7563.

[31] LIU Y,WEN J. Complex scene text detection based on attention mechanism[J]. Computer Science,2020,47(7):135-140.

[32] CAI Y,WANG W,REN H,et al. SPN:short path network for scene text detection[J]. Neural Computing and Applications, 2019,32(1):1-13.

[33] HE W,ZHANG X Y,YIN F,et al. Realtime multi-scale scene text detection with scale-based region proposal network[J]. Pattern Recognition,2020,98:107026.

[34] QIN X,JIANG J,YUAN C A,et al. Arbitrary Shape Natural Scene Text Detection Method Based on Soft Attention Mechanism and Dilated Convolution[J]. IEEE Access, 2020, 8: 122685-122694.

[35] ZHANG L,LIU Y,XIAO H,et al. Efficient Scene Text Detection with Textual Attention Tower[C]// ICASSP 2020-2020 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). IEEE,2020:4272-4276.

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

[37] WANG X,JIANG Y,LUO Z,et al. Arbitrary shape scene text detection with adaptive text region representation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:6449-6458.

[38] XU Y,WANG Y,ZHOU W,et al. Textfield:Learning a deep direction field for irregular scene text detection[J]. IEEE Transactions on Image Processing,2019,28(11):5566-5579.

SHAO Hai-lin, born in 1995, postgraduate, is a member of China Computer Federation. Her main research interests include scene text detection and so on.

LIU Chun-ping, born in 1971, Ph. D, professor, Ph. D supervisor. Her main research interests include computer vision, image analysis and recognition, in particular in domains of visual saliency detection, objection and scene understanding.

(责任编辑:李亚辉)