文章编号：2096-1618(2022)02-0171-06

自然场景下交通标识文本检测与识别算法研究

胡高丽， 文成玉
(成都信息工程大学通信工程学院，四川 成都 610225)

摘要：针对自然场景下交通标志牌文本粘连、字体复杂、大小形状不一、难以分行，导致交通文本标识率低的问题，提出一种基于 PSENet+CRNN 的改进交通文本检测识别算法。检测算法以 PSENet 为基础网络，采用特征增强模块 FEM 来增加模型的接受域，并改进空洞卷积的特征金字塔模型来增强多支路深层语义信息的融合能力。文本识别部分在 CRNN 模型中采用 CTC+CenterLoss 实现功能和标签的对齐、解决预测重复、预测漏字时的对齐问题。最终在 CTST-1600 数据集上进行验证，检测准确率达到了92.5%，字符识别率达到了88.9%，与原算法相比，分别提升了识别率4.3%和2.3%。实验结果表明，该方法有效提升了模型的检测与识别精度。

关 键 词：PSENet；CRNN；交通文本；文本检测；字符识别
中图分类号：TP391.4 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2022.02.010

0 引言

文字是承载信息的载体，伴随着计算机视觉技术等飞速发展，许多新兴的应用场景都需要提取图像中的文本信息如获取交通标志上的文本信息用于定位和导航车辆，保证道路交通与行车安全。交通标志文本包含了丰富且有价值的交通信息，交通标志检测与识别作为智能驾驶的关键一环，具有重要研究意义[1-2]。

目前采用深度学习算法作为主流的文本检测与识别技术，一般使用的文字检测模型算法主要有 CTPN、EAST、PSENet 及目标检测网络(YOLO 系列[3])等。CTPN[4]能够有效识别水平文本，但是对不规则文本的检测效果差。EAST[5]能有效地解决竖向及倾斜文本的问题，但对较长的文本识别效果不好。基于分割的 PSENet[6]网络能够检测任意形状的文本信息，很好区分文本边界但是应用于交通标志文本检测时，由于文字大小相差较大，且常常伴有辅助箭头横线等，因此在分辨率减低且字符很小的情况时，检测精度有待提升。

在文字识别方面多采用 CRNN 方法。CRNN[7]是在卷积特征的基础上提取序列特征，并采用 CTC 损失[8]来解决预测结果和标签不一致的问题，但局限于识别规则文本。本文涉及交通文本信息的识别，虽然不存在字体弯曲现象，但常出现形近字识别错误的问题，影响识别效果。

在分析现有算法的基础上，采用 PSENet+CRNN 作为交通文本检测与识别算法。在 PSENet 模型中提出了特征增强模块 FEM，增加模型的接受域，以提高对文本区域的检测能力。并改进空洞卷积的特征金字塔模型，增强多支路深层语义信息的融合能力，提升特征信息的利用率。在 CRNN 模型中采用 CTC+CenterLoss[9]实现功能和标签的对齐，解决预测重复、预测漏字时的对齐问题。先把改进的 PSENet 作为交通文本检测模型，可以精准定位多行文本，返回文本候选框，再通过改进的 CRNN 的文本识别算法完成对候选文本框的识别任务。实验结果表明，本文所改进的算法有效提高了检测识别精度。

1 PSENet 文本检测算法原理及改进方法

1.1 PSENet 算法原理

PSENet 是一种基于像素级别的分割检测方法，不同于锚框的局限性，可以检测任意形状的文本信息，并且采用渐进式的尺度扩展算法，能够很好区分文本边界，有效识别出相邻文本，网络结构如图 1 所示。

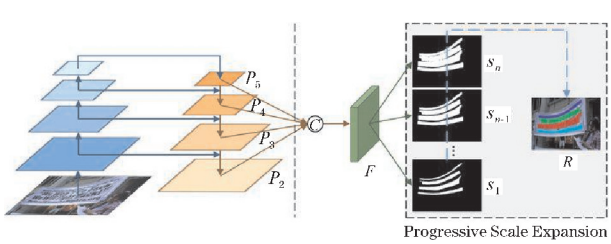

图 1 PSENet 网络结构图

PSENet 网络结构主要包含两部分，文本实例分割部分和渐进式尺度扩展算法部分。文本实例分割部分采用 ResNet[10]和 FPN 结构作为特征提取网络，输入图

收稿日期：2021-09-02

片先通过残差网络进行特征提取,再通过特征金字塔结构获取和融合不同尺度特征图的语义信息,最后对输出结果进行拼接,得到融合特征 F,并对 F 进行卷积上采样得到 n 个不同尺度不同的分割图 $S_1,S_2,\cdots,S_n$。

另一部分采用渐进式尺度扩展算法 PSE,如图 2 所示,它采用广度优先搜索方法,从小尺度的 S_1 开始,得到图片中所有文本实例的最小内核,然后合并 S_2 的分割结果,对 n 个不同比例的分割图进行区域扩展,得到最大尺度的分割图,完成最终预测。

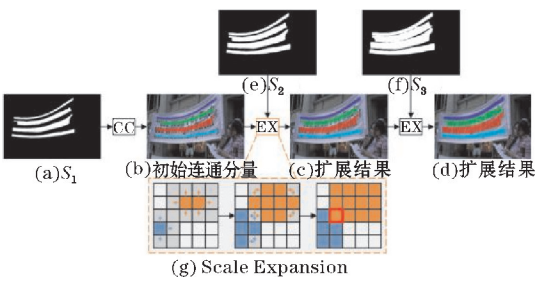

图 2 渐进式尺度扩展算法

1.2 改进 PSENet 模型的网络结构

1.2.1 引入空洞卷积

基于像素级别的图像检测及分割任务,不仅需要抽象的高层语义特征预测结果,也需要细节信息保证分割或定位的精度。为提高网络的特征提取能力,常采池化操作在降低图像尺寸的同时增大感受野,虽然池化提升了特征信息的获取能力,但是会丢失细节信息,带来空间分辨率的下降。

空洞卷积[11]可以在不丢失分辨率、不增加参数量的情况下以指数方式增加感受野,扩大卷积操作后输出的信息范围,并引入扩张率 K 定义卷积操作时卷积核处理数据值的间距,以此获得尺度信息,对需要全局信息或者较长序列的文本信息问题,均有很好的应用效果。

以常见的 3×3 卷积为例,如图 3 所示,图 3(a)对应扩张率为 1 的空洞卷积,即普通卷积;图 3(b)为扩张率为 2 的空洞卷积,实际的卷积核大小还是 3×3,但是感受野已经增大到了 5×5;图 3(c)对应扩张率为 3 的空洞卷积,能得到 7×7 的感受野,是传统卷积操作的 3 个 3×3 卷积叠加和。

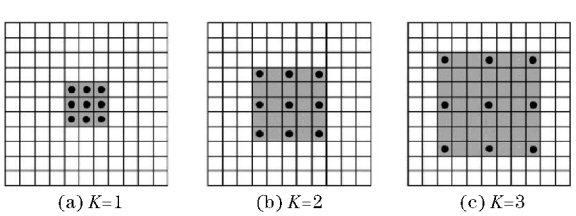

图 3 普通卷积与空洞卷积的对比

采用 PSENet 网络检测长文本或者区域较大的图片信息时,因为受到感受野的限制,出现文本区域中断,没有正确识别完整的文本区域。PSENet 特征提取网络结合 ResNet 残差网络和 FPN 特征金字塔结构[12],融合不同尺度的特征信息,网络结构如图 4 所示。本文选取 ResNet 网络的后 4 个卷积层输出作为 FPN 的输入特征,将参与 FPN 特征融合的这 4 个卷积层中 3×3 的普通卷积改为扩张因子为 2 的空洞卷积,以增大感受野,同时使特征图保留更多目标信息。

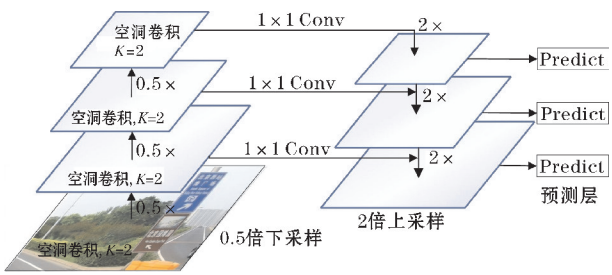

图 4 改进后的特征提取网络

1.2.2 特征增强模块 FEM

提出特征增强模块(),用以增加网络结构的接受域,提高模型对文本区域的检测能力,进而提升网络性能。模型使用 3 种大小不同的卷积核(1×1、3×3、5×5)提取特征,高效利用计算资源,获取更多的语义信息,然后对不同尺寸和感受野的输出特征图进行拼接操作,以达到融合不同特征的目的,提升训练效果。

卷积核尺寸越大,其感受野越大,但缺点是参数量更大,计算复杂度也更高。为了保持在感受野大小不变的同时减少参数量,采用卷积核 $n\times n$ 可以分解成 $1\times n$ 卷积和 $n\times 1$ 卷积的组合方法,使卷积核的参数量从 $n\times n$ 减少到 $2n$,不会增加太多的计算负担[13]。特征增强模块如图 5 所示。

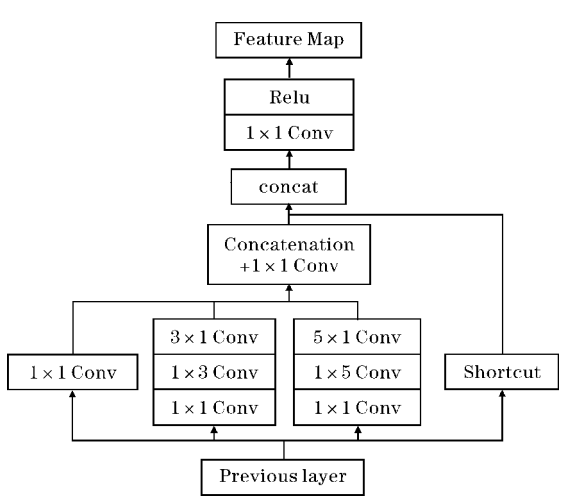

图 5 特征增强模块 FEM

改进的文本检测网络模型如图 6 所示,在骨干特

征提取网络 ResNet 使用空洞卷积的 C2 ~ C5 残差块的最后一个卷积层分别加入特征增强模块 FEM，并通过 FEM 模块末端 1×1 的卷积操作保证其输出特征层的大小和通道数保持不变，再将输出结果通过 2 倍上采样操作恢复到 M2 ~ M5 的特征图大小，形成增加了 FEM 模块的增强型特征金字塔结构。该结构在不增大参数量的基础上增大模型的感受野，提高对多行、多列及长文本的检测能力。最后通过 C 函数将 $P3$、$P4$、$P5$ 分辨率大小分别上采样到 $P2$ 大小，再进行拼接得到融合后的特征图。整合函数 F 如下所示

$$F = C(P_2, P_3, P_4, P_5) = P_2 \parallel Up_{\times 2}(P_3) \parallel Up_{\times 4}(P_4) Up_{\times 8}(P_5) \quad (1)$$

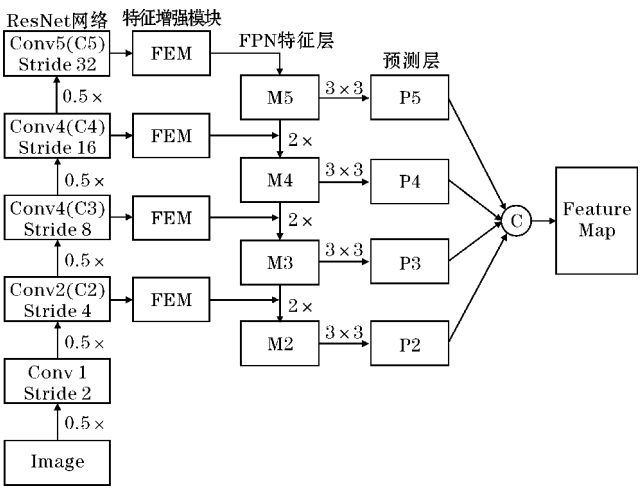

图 6　改进文本检测模型

2　CRNN 文本识别模型

2.1　CRNN 文本识别原理

采用 CRNN 文本识别模型，它结合卷积神经网络 CNN 和循环神经网络 RNN 的特性提取图像特征，可以进行端到端的训练，并借用自然语言处理任务中的序列标注任务的思想，将序列标注算法嵌套在现有的深度卷积网络中，组成完整的支持端到端的梯度反向传播算法，CRNN 网络架构如图 7 所示。

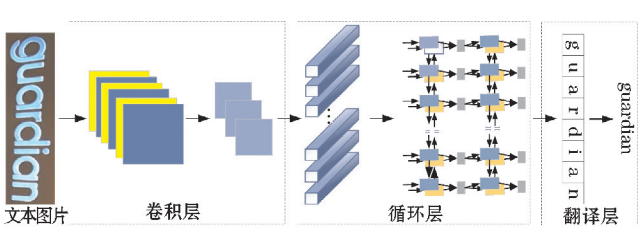

图 7　CRNN 网络架构

CRNN 模型主要由两个部分组成：特征提取部分，由多个卷积层、池化和非线性特性层组成；序列预测部分，由 RNN 和 CTC 模型组成。RNN 部分[14]主要用于学习和建模 CNN 中提到的隐藏状态以及空间特征之间的联系，最后预测初步的序列结果。粗糙的预测序列可能存在字母重复的情况，通过 CTC 模块对 RNN 的序列进行整合，可以对序列进行去重操作。CRNN 接收灰度图或 RGB 彩色图片作为输入，CNN 作为编码器来提取与图片对应的中间层特征。经过变形后整理成 T 个时间步的输入送入随后的解码器 RNN，从而预测初步的序列。初步的序列经过 CTC 整流处理，去除冗余的字符后可以得到最终的预测结果。

2.2　采用 CenterLoss 来解决相近字的问题

直接使用 RNN 进行时序分类时会出现大量的冗余情况，如一个字母可能被识别两次，但是直接采用两个连续字母去掉一个这样的去重处理方法容易造成信息丢失。如文本图片“apple”经过 RNN 网络预测的结果可能是“applee”，简单去重后输出“aple”作为最终结果，显然造成了原始信息丢失。

为解决这一问题，需要用自带空白符号的 CTC 算法，通过 CTC 算法根据条件概率的原理将适当的位置设置为占位符‘_’，上述预测结果可能变成“a_p_p_l_ee”，经过去重处理后再除去所有占位符即可得到最终预测结果 apple。

除去在字符转录的过程中出现冗余的情况，也可能会出现形近字识别错误的问题，影响识别效果，故本文采用 CTC+CenterLoss 实现功能和标签的对齐，解决预测重复，预测漏字时的对齐问题。CenterLoss 最早是用于人脸识别中的损失函数，该损失函数的目的就是更好扩大类间距离，缩小类内距离；而在字符识别中的使用，CenterLoss 可以强化特征之间的差异，能够形近字分类困难的问题。CenterLoss 公式如下：

$$L_C = \frac{1}{2}\sum_{i=1}^{m} \| \boldsymbol{x}_i - c_{y_i} \|_2^2 \quad (2)$$

其中，$\boldsymbol{x}_i$ 为特征向量，c_{y_i} 表示类别中心，m 表示 mini-batch 的大小。在具体应用中，需要传入两个变量，一个特征向量 $\boldsymbol{x}$，一个 label，故需要预先训练好一个识别率不错的模型，生成 label 标签，预测每一份的类别数，再进行 CenterLoss 计算。

3　实验及结果分析

3.1　数据集

选择在 CTST-1600[15]数据集上进行实验。CTST-1600 是来自中国交通的文本数据集，如图 7 所示，共有

1600 张图片,其中训练集 1320 张,测试集 280 张。该数据集仅对文本定位区域进行标注,没有相应的文本识别任务标注,故本实验对 CTST-1600 数据集进行文本区域的获取与剪裁,得到文本识别区域,并进行标注得到文本识别的标签信息。如图 8 所示,1320 张图片共获取 6174 条文本信息,可有效训练和验证文本识别网络的鲁棒性。

图 7 CTST-1600 交通文本数据集

图 8 交通文本识别数据集

3.2 实验环境及参数设置

实验平台为 windows10(64bit)操作系统,E5-2620 CPU、24G 内存,NVIDIA Tesla M40 计算卡、24GB 显存,CUDA 版本为 10.2、CUDNN 7.5,使用 PyTorch 框架搭建网络模型。

实验采用基于改进的 PSENet 和 CRNN 的文本检测与识别网络,其中文本检测网络的训练代次数 epoch 设置为 600,初始学习速率为0.001,Batch Size 为 8,采用 SGD 梯度下降算法来优化网络,动量设置为 0.9。文本识别网络 epoch 设置为 1000,初始学习率为 0.0001,Batch Size 为 64,采用 Adam 优化器来更新梯度。

3.3 评价准则

采用精准率 P(Precision)、召回率 R(Recall)和 F_1 分数来衡量模型的检测识别能力,计算公式如下:

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{3}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{4}$$

$$F_1=2\times\frac{P\times R}{P+R} \tag{5}$$

其中,TP 为真正例、FP 为假正例和 FN 为假反例。

3.4 实验及结果分析

为分析不同改进方法的有效性,在相同实验环境和参数配置的情况下,对基于改进的空洞卷积特征金字塔结构、添加特征增强模块以及融合改进后的 PSENet 网络在 CTST-1600 数据集上进行训练和测试,并与原网络进行性能对比分析,实验结果如表 1 所示。

表 1 基于 PSENet 网络改进的算法性能对比 单位:%

模型	精准率	召回率	F_1
PSENet	88.2	85.6	86.9
PSENet+改进 FPN	90.1	87.2	88.6
PSENet+FEM	89.6	90.3	89.9
PSENet+改进 FPN+FEM	92.5	92.8	92.7

从表 1 可以看出,本文对文本检测网络的改进效果得到了有效验证。基于改进的空洞卷积特征金字塔结构相较于原网络提高了精确度、召回率和 F_1 分数,这是因为空洞卷积能够增加局部感受野,特征层将获取更丰富的多尺度语义信息,改进后的特征融合网络将浅层与深层特征信息充分融合,突出了目标信息,从而提高了文本定位的精度。基于融合的特征增强模块的精度虽然有所下降,但是提高了文本的检测能力,召回率从85.6% 提高到90.3%,F_1 分数从86.9% 提高到 89.9%,并且在没有引入复杂结构的前提下,显著增加了接受域,只带少量参数的增加,没有增加网络的计算负担。

为验证本文算法的普适性,选用目前主流的 CTPN、EAST 在数据集 CTST-1600 上进行实验对比,实验结果如表 2 所示。可以看出本文改进算法整体表现优于其他网络模型。

表 2 本文改进算法与其他网络模型的对比 单位:%

模型	精确率	召回率	F_1
CTPN	79.6	70.4	74.7
EAST	84.8	87.5	86.1
PSENet	88.2	85.6	86.9
本文改进方法	92.5	92.8	92.7

图 10 展示了本文改进的文本检测网络在交通文本数据集 CTST-1600 上训练后测试获得的结果。从图 10 的文本定位区域可以看出,改进后的模型生成的文本定位框更精确,误检漏情况较少,基本实现对交通文本检测的精准定位。

图 10 文本检测网络改进前后效果对比图

表3 对比了改进的 CRNN 模型与原始网络对文本识别的影响。由实验结果可以看出,采用 CTC+CenterLoss 可以解决文本对齐以及重叠及相近字等问题,提高了字符正确识别率。改进的 CRNN 算法识别准确率为88.9%,有效提升了文本识别的精度,验证了改进算法的有效性。

表3　改进 CRNN 模型对比试验

模型	精确率/%	FPS
CRNN	86.3	20
改进算法	88.9	19

把基于 PSENet 的文本检测网络与 CRNN 的识别模型统一到一个网络构架中,采用 PSENet 网络进行文本定位,并裁剪出文本区域送入本文改进的 CRNN 模型来验证改进算法的可行性,实验结果如图 11 所示。

图 11　交通文本检测与识别效果图

4　结束语

针对自然场景下交通文本检测的准确性等问题,提出一种基于深度学习的交通文本检测识别算法,有效提高了文本检测与识别的精准度。将 PSENet 网络与特征增强网络相结合,有利于文本区域的特征提取,可以获得更丰富的语义信息;在 FPN 中采用空洞卷积代替普通卷积,能够增大局部感受野,构建丰富的位置信息与语义信息的融合,有效提升交通文本的检测能力。对 CRNN 网络采用 CTC+CenterLoss 来解决预测重复,预测漏字时的对齐问题,提高了文本识别的准确率。实验表明,本文所采用的方法提高了交通文本检测与识别的精度,证实了算法的有效性。

参考文献:

[1] 李益红,陈袁宇.深度学习场景文本检测方法综述[J].计算机工程与应用,2021,57(6):42-48.

[2] 师广琛,巫义锐.像素聚合和特征增强的任意形状场景文本检测[J].中国图象图形学报,2021,26(7):1614-1624.

[3] RedmonJ, DivvalaS, GirshickR, et al. You Only Look Once: Unified, Real-Time Object Detection [J]. Computer Vision & Pattern Recognition, 2016:779-788.

[4] Tian Z, Huang W, He T, et al. Detecting text in natural image with connectionist text proposal network[C]. European conference on computer vision. Springer, Cham, 2016:56-72.

[5] 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 Recogition, 2017:5551-5560.

[6] Li X, Wang W, Hou w, et al. Shape Robust Text Detection with Progressive Scale Expansion Network[J]. Computer Science Computer´ Vision and Pattern Recognition, 2018:1-12.

[7] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(11):2298-2304.

[8] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks [C]. Proceedings of the 23rd international conference on Machine learning. 2006:369-376.

[9] Wen Y, Zhang K, Li Z, et al. A Discriminative Feature Learning Approach for Deep Face Recognition [C]. European Conference on Computer Vision. Springer, Cham, 2016:499-515.

[10] 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. 2016:770-778.

[11] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[EB/OL]. arXiv preprint arXiv:1511.07122, 2015.

[12] 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.

[13] Szegedy C, Liu W, Jia Y, et al. Going deeper with-convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:1-9.

[14] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]. 2013 IEEE international conference on acoustics, speech and signal processing. Ieee, 2013:6645-6649.

[15] He X, Wang R, Li X, et al. HTSTL: Head-and-Tail search network with scale-transfer layer for traffic sign text detection[J]. IEEE Access, 2019, 7:118333-118342.

Research on Algorithms for Text Detection and Recognition of Traffic Signs in Natural Scenes

HU Gaoli, WEN Chengyu

(College of Communication Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Aiming at the problem of low identification rate of traffic text due to the adhesion of traffic sign text, complex fonts, different sizes and shapes, and difficulty in branching in natural scenes, an improved traffic text detection and recognition algorithm based on PSENet+CRNN is proposed. The detection algorithm uses PSENet as the basic network, uses the feature enhancement module FEM to increase the acceptance domain of the model, and improves the feature pyramid model of the hollow convolution to enhance the fusion ability of multi-branch deep semantic information. The text recognition part uses CTC+CenterLoss in the CRNN model to realize the alignment of functions and labels, and solve the problem of predicting repetition and alignment when predicting missing characters. Finally, it was verified on the CTST-1600 data set. The detection accuracy rate reached 92.5%, and the character recognition rate reached 88.9%. Compared with the original algorithm, the recognition rate increased by 4.3% and 2.3%, respectively. Experimental results show that this method effectively improves the accuracy of model detection and recognition.

Keywords: PSENet; CRNN; traffic text; text detection; character recognition