JIG
JOURNAL OF IMAGE AND GRAPHICS
主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021
07
VOL.26
ISSN1006-8961
CN11-3758/TB

中国多媒体大会(ChinaMM2020)专刊

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第26卷第7期（总第303期）

2021年7月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

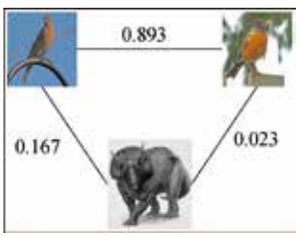
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2021年7月 第26卷 第7期（总第303期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次



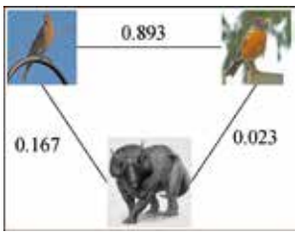
类别语义相似性监督的小样本图像识别(第1594页)

像素聚合和特征增强的任意形状场景文本检测(第1614页)

时序特征融合的视频实例分割(第1692页)

Volume 26, Number 7
Published July 16, 2021

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Few shot image recognition based on class semantic similarity supervision(P1594)

Arbitrary shape scene-text detection based on pixel aggregation and feature enhancement (P1614)

Video instance segmentation based on temporal feature fusion (P1692)

E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2021)07-1614-11

论文引用格式:Shi G C and Wu Y R. 2021. Arbitrary shape scene-text detection based on pixel aggregation and feature enhancement. Journal of Image and Graphics,26(07):1614-1624(师广琛,巫义锐. 2021. 像素聚合和特征增强的任意形状场景文本检测. 中国图象图形学报,26(07):1614-1624)[DOI:10.11834/jig.200522]

像素聚合和特征增强的任意形状场景文本检测

师广琛,巫义锐*

河海大学计算机与信息学院,南京　211100

摘　要:**目的**　获取场景图像中的文本信息对理解场景内容具有重要意义,而文本检测是文本识别、理解的基础。为了解决场景文本识别中文字定位不准确的问题,本文提出了一种高效的任意形状文本检测器:非局部像素聚合网络。**方法**　该方法使用特征金字塔增强模块和特征融合模块进行轻量级特征提取,保证了速度优势;同时引入非局部操作以增强骨干网络的特征提取能力,使其检测准确性得以提高。非局部操作是一种注意力机制,能捕捉到文本像素之间的内在关系。此外,本文设计了一种特征向量融合模块,用于融合不同尺度的特征图,使尺度多变的场景文本实例的特征表达得到增强。**结果**　本文方法在3个场景文本数据集上与其他方法进行了比较,在速度和准确度上均表现突出。在ICDAR(International Conference on Document Analysis and Recognition)2015数据集上,本文方法比最优方法的F值提高了0.9%,检测速度达到了23.1帧/s;在CTW(Curve Text in the Wild)1500数据集上,本文方法比最优方法的F值提高了1.2%,检测速度达到了71.8帧/s;在Total-Text数据集上,本文方法比最优方法的F值提高了1.3%,检测速度达到了34.3帧/s,远远超出其他方法。**结论**　本文方法兼顾了准确性和实时性,在准确度和速度上均达到较高水平。

关键词:目标检测;场景文本检测;神经网络;非局部模块;像素聚合;实时检测;任意形状

Arbitrary shape scene-text detection based on pixel aggregation and feature enhancement

Shi Guangchen, Wu Yirui*

School of Computer and Information, Hohai University, Nanjing　211100, China

Abstract: **Objective** Text can be seen everywhere, such as on street signs, billboards, newspapers, and other items. The text on these items expresses the information they intend to convey. The ability of text detection determines the level of text recognition and understanding of the scene. With the rapid development of modern technologies such as computer vision and internet of things, many emerging application scenarios need to extract text information from images. In recent years, some new methods for detecting scene text have been proposed. However, many of these methods are slow in detection because of the complexity of the large post-processing methods of the model, which limits their actual deployment. On the other hand, the previous high-efficiency text detectors mainly used quadrilateral bounding boxes for prediction, and accurately predicting

收稿日期:2020-08-27;**修回日期**:2021-03-01;**预印本日期**:2021-03-08

*通信作者:巫义锐　wuyirui@hhu.edu.cn

基金项目:国家重点研发计划项目(2018YFC0407901);国家自然科学基金项目(61702160);中央高校基本科研业务费专项资金资助(B200202177);江苏省自然科学基金项目(BK20170892)

Supported by: National Key Research and Development Program of China (2018YFC0407901); National Natural Science Foundation of China (61702160); Fundamental Research Funds for the Central Universities(B200202177); Natural Science Foundation of Jiangsu Province, China(BK20170892)

arbitrary-shaped scenes is difficult. **Method** In this paper, an efficient arbitrary shape text detector called non-local pixel aggregation network (non-local PAN) is proposed. Non-local PAN follows a segmentation-based method to detect scene text instances. To increase the detection speed, the backbone network must be a lightweight network. However, the presentation capabilities of lightweight backbone networks are usually weak. Therefore, a non-local module is added to the backbone network to enhance its ability to extract features. Resnet-18 is used as the backbone network of non-local PAN, and non-local modules are embedded before the last residual block of the third layer. In addition, a feature-vector fusion module is designed to fuse feature vectors of different levels to enhance the feature expression of scene texts of different scales. The feature-vector fusion module is formed by concatenating multiple feature-vector fusion blocks. Causal convolution is the core component of the feature-vector fusion block. After training, the method can predict the fused feature vector based on the previously input feature vector. This study also uses a lightweight segmentation head that can effectively process features with a small computational cost. The segmentation head contains two key modules, namely, feature pyramid enhancement module (FPEM) and feature fusion module (FFM). FPEM is cascadable and has a low computational cost. It can be attached behind the backbone network to deepen the characteristics of different scales and make the network more expressive. Then, FFM merges the features generated by FPEM at different depths into the final features for segmentation. Non-local PAN uses the predicted text area to describe the complete shape of the text instance and predicts the core of the text to distinguish various text instances. The network also predicts the similarity vector of each text pixel to guide each pixel to the correct core. **Result** This method is compared with other methods on three scene-text datasets, and it has outstanding performance in speed and accuracy. On the International Conference on Document Analysis and Recognition (ICDAR) 2015 dataset, the F value of this method is 0.9% higher than that of the best method, and the detection speed reaches 23.1 frame/s. On the Curve Text in the Wild (CTW) 1500 dataset, the F value of this method is 1.2% higher than that of the best method, and the detection speed reaches 71.8 frame/s. On the total-text dataset, the F value of this method is 1.3% higher than that of the best method, and the detection speed reaches 34.3 frame/s, which is far beyond the result of other methods. In addition, we design parameter setting experiments to explore the best location for non-local module embedding. Experiments have proved that the effect of embedding the non-local module is better than non-embedding, indicating that non-local modules play an active role in the detection process. According to the detection accuracy, the effect of embedding non-local blocks into the second, third, and fourth layers of ResNet-18 is significant, while the effect of embedding the fifth layer is not obvious. Among the methods, embedding non-local blocks in the third layer has the best effect. We designed ablation experiments on the ICDAR 2015 dataset for the non-local and feature-vector fusion modules. The experimental results prove that the superiority of the non-local module does not come from deepening the network but from its own structural characteristics. The feature vector fusion module also plays an active role in the scene text-detection process, which combines feature maps of different scales to enhance the feature expression of scene texts with variable scales. **Conclusion** In this paper, an efficient text detection method for arbitrary shape scene is proposed, which considers accuracy and realtime. The experimental results show that the performance of our model is better than that of previous methods, and our model is superior in accuracy and speed.

Key words: object detection; scene text detection; neural network; non-local module; pixel aggregation; real-time detection; arbitrary shape

0 引　言

文本在现实生活中处处可见,物品上的文字表达了人们想传递的信息。对文本的检测能力决定了对文本的识别和对场景的理解水平。随着计算机视觉和物联网等现代技术的高速发展,许多新兴的应用场景都需要提取图像中的文本信息,比如获取路牌中的指路信息为自动驾驶的汽车指引方向,门牌号识别实现无人送货等。一些检测场景文本的新方法被提出,例如 TextSnake(Long 等,2018)达到较高的检测水平,但其后处理方法过于庞大复杂,因而检测速度较慢,这从根本上限制了其在现实中的应用。而 gliding vertex (Xu 等,2021)等高效文本检测方法主要是以四边形边界框进行预测,在检测弯曲文本时会产生偏斜。

为了解决上述问题,本文提出一种高效的任意

形状场景文本检测器:非局部像素聚合网络(non-local pixel aggregation network),可用于检测多方向、任意形状的场景文本,并可以在速度和性能之间取得良好的平衡。本文方法使用特征金字塔增强模块和特征融合模块进行轻量级特征提取,为了弥补轻量级网络提取特征能力不足的缺陷,为骨干网络嵌入非局部模块以增强其提取特征的能力;此外,本文提出了特征向量融合模块,用于增强多尺度场景文本的特征表达,使其检测的准确性得到提高。如图 1 所示,与其他对比方法相比,本文方法的性能与速度均处于领先水平。

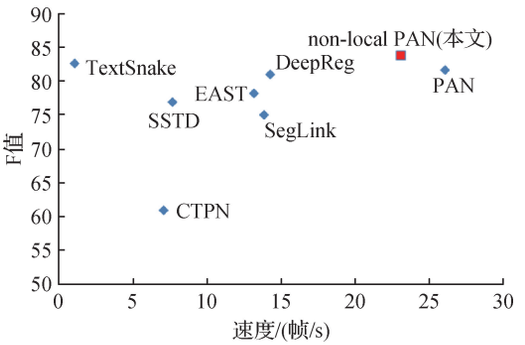


图 1 各模型检测速度和准确度的对比图

Fig. 1 Comparison chart of detection speed and accuracy of each model

1 相关工作

1.1 实时场景文本检测

实时场景文本检测需要使用一种快速方法来生成高质量的文本预测结果。EAST(an efficient and accurate scene text detector)(Zhou 等,2017)直接使用 FCN(fully convolutional network)(Long 等,2015)来预测得分图和相应坐标,然后使用非极大值抑制得到输出结果。EAST 的整个流程非常简洁,因此可以做到实时检测。MCN(Markov clustering network)(Liu 等,2018)将文本检测问题表达为基于图的聚类问题,并在不使用非极大值抑制的情况下生成边界框,使得 MCN 可以在 GPU(graphics processing unit)上完全并行化。但是,这些方法是专为四边形文本检测而设计的,对任意形状场景文本的预测结果非常不理想。

1.2 任意形状的场景文本检测

2017 年,CTW(Curve Text in the Wild) 1500(Liu 等,2017)和 Total-Text(Chng 和 Chan,2017)等任意形状场景文本数据集的出现,带来对任意形状场景文本的研究热潮。

为了检测弯曲的场景文本,Lyu 等人(2018)提出了一种 Mask TextSpotter,巧妙地细化了 Mask RCNN(region-convolutional neural networks),利用字符级标签同时检测和识别字符和实例掩码。该方法显著提高了面向点定位或曲线场景文本的性能。然而,字符级标签的成本极其昂贵,所以该方法难以落实到实际应用中。Liao 等人(2021)对该方法进行改进,显著减轻了对字符级标签的依赖。该方法依赖于区域生成网络,在一定程度上限制了检测速度。

Qin 等人(2019)提出使用 RoI(region of interest)掩膜来聚焦弯曲的文本区域,但其结果很容易受到离群像素的影响。另外,分割分支增加了计算负担,拟合多边形过程也带来了额外的计算负担。Liu 等人(2019)提出了一种基于金字塔掩模的场景文本检测算法(pyramid mask text detector, PMTD)。该检测算法不再预测文本实例的二值掩膜,而是对每个像素进行回归操作,从而使得生成的文本实例掩膜具有更丰富的信息。

2 非局部像素聚合网络

2.1 整体架构

如图 2 所示,non-local PAN(pixel aggregation network)遵循基于分割的方法流程来检测场景文本实例。为了提高效率,骨干网络必须是轻量级网络。但是,轻量级骨干网络的表示能力通常较弱。因此,本文为骨干网络添加了非局部模块,用于增强其提取特征的能力。如图 3 所示,采用 ResNet-18(He 等,2016)作为 non-local PAN 的骨干网络,并将非局部模块嵌入其第 3 层的最后一个残差块之前。此外,本文还设计了特征向量融合模块用于融合不同层次的特征向量,以增强不同尺度的场景文字的特征表达。使用一种轻量的分割头,有效地以较小的

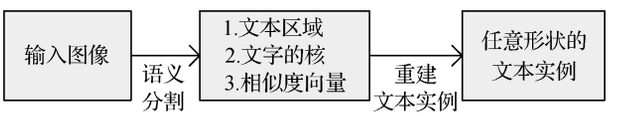


图 2 Non-local PAN 的总体流程

Fig. 2 The overall process of non-local PAN

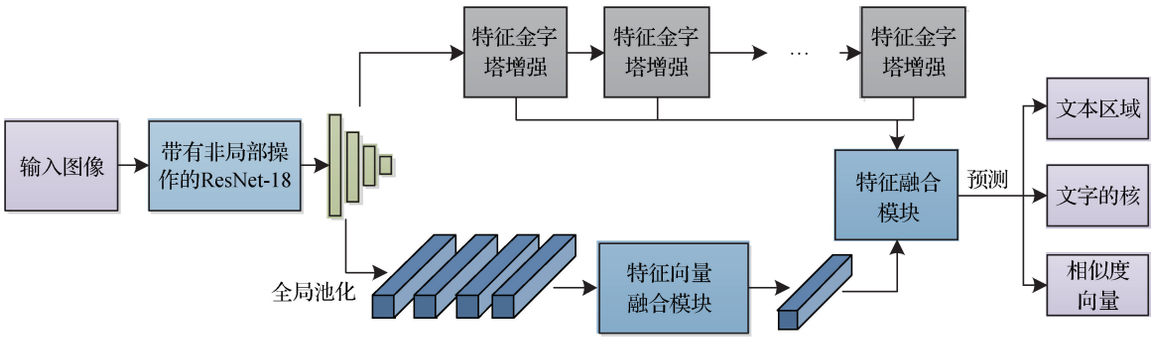


图 3　non-local PAN 的整体架构

Fig. 3　The overall architecture of non-local PAN

计算成本对特征进行处理。该分割头包含两个关键模块，即功能金字塔增强模块（feature pyramid enhancement module，FPEM）和特征融合模块（feature fusion module，FFM）。FPEM 是可级联的，并且计算成本较低，可以将其附着在骨干网络后面，以加深其不同尺度的特征，使其更具表现力。之后，FFM 将不同深度 FPEM 产生的特征融合到最终的特征中进行分割。non-local PAN 用预测出的文本区域来描述文本实例的完整形状，并预测文本的核以区分不同的文本实例。网络还预测每个文本像素的相似度矢量，用于指引每个像素聚合到正确的核中。

2.2　非局部模块

捕捉大范围内数据相互之间的依赖关系是一个很重要的问题。一般方法通常使用较大的卷积核来捕捉图像中较远距离的像素之间的关系。然而，传统的卷积神经网络只是在其时间或空间的很小的邻域内进行捕捉，却很难捕获到更远位置的数据的依赖关系。

非局部网络（non-local network）（Wang 等，2018）可以很好地捕捉到较远位置的像素点之间的依赖关系。定义在深度神经网络中的非局部操作为

$$\boldsymbol{y}_i = \frac{1}{C(\boldsymbol{x})}\sum_{\forall j} f(\boldsymbol{x}_i,\boldsymbol{x}_j)g(\boldsymbol{x}_j) \tag{1}$$

式中，$\boldsymbol{x}$ 表示输入信号，$\boldsymbol{y}$ 表示输出信号，其尺度大小与 $\boldsymbol{x}$ 相同。$f(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 用来计算 i 位置像素和所有可能关联的 j 位置像素之间的内在关系。在本文中，将 $f(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 实现为嵌入高斯核函数，具体计算为

$$f(\boldsymbol{x}_i,\boldsymbol{x}_j) = \mathrm{e}^{\theta(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)} \tag{2}$$

式中，$\theta(\boldsymbol{x}_i) = \boldsymbol{W}_\theta \boldsymbol{x}_i$ 和 $\varphi(\boldsymbol{x}_j) = \boldsymbol{W}_\varphi \boldsymbol{x}_j$ 是两个待学习的嵌入空间。该累加后的结果由因子 $C(\boldsymbol{x})$ 归一化，计算为

$$C(\boldsymbol{x}) = \sum_{\forall j} f(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{3}$$

$g(\boldsymbol{x}_j)$ 用于计算输入信号在 j 位置的特征值，计算为

$$g(\boldsymbol{x}_j) = \boldsymbol{W}_g \boldsymbol{x}_j \tag{4}$$

式中，$\boldsymbol{W}_g$ 是要学习的权重矩阵。在设计网络时，上式被实现为空间中的 1 × 1 卷积操作。

由以上公式，实现上述非局部操作，非局部模块的结构如图 4 所示，其中，“ × ”表示矩阵乘法。

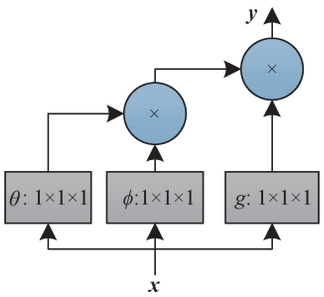


图 4　非局部模块的细节

Fig. 4　Details of non-local modules

将非局部模块嵌入骨干网络 ResNet-18 时，一般插入到不同阶段最后一个残差块之前，并且通过实验发现在第 2 层、第 3 层和第 4 层上嵌入非局部模块的效果好，而在第 5 层嵌入非局部模块的效果不明显。本文方法中，非局部模块被嵌入到 ResNet-18 的第 3 层。嵌入非局部模块的 ResNet-18 第 3 层网络结构如图 5 所示，“ + ”表示逐元素求和。

2.3　特征向量融合模块

为了融合不同尺度的特征图，增强尺度多变的场景文字的特征表达，本文提出了特征向量融合模块。特征向量融合模块由 4 个特征向量融合块串联

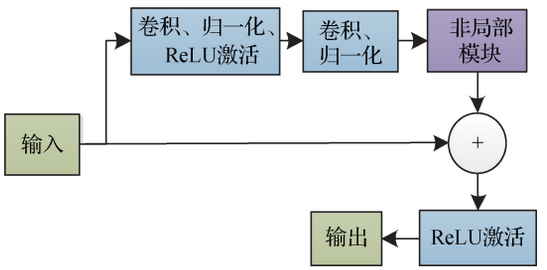


图 5　ResNet-18 第 3 层的网络结构图

Fig. 5　The third layer network structure diagram of ResNet-18

而成。

因果卷积(van der Oord 等,2016)是特征向量融合块的核心成分,因果卷积首先被 WaveNet(van der Oord 等,2016)用于生成原始音频。在 WaveNet 中,因果卷积还进行了扩张操作,以实现用较少的网络层数覆盖较大的感受野。而在本文方法中,输入的特征向量只有 4 个,无需使用扩张卷积。因果卷积的输出长度与其输入长度相同,并且当前输出仅取决于当前点及之前的输入信息。将不同层次的特征向量依次输入到特征向量融合模块中,因果卷积会将之前输入的特征向量信息用于本次特征向量的处理。通过训练,可以使最后一个特征向量对应的输出融合所有特征向量的信息。

特征向量融合块的结构如图 6 所示,输出数据 $\boldsymbol{x}^{(i+1)}$ 的长度与其输入数据 $\boldsymbol{x}^{(i)}$ 相同。特征向量融合块公式化为

$$\boldsymbol{t} = Causal(\boldsymbol{x}^{(i)}) \odot \sigma(Causal(\boldsymbol{x}^{(i)})) \tag{5}$$

$$\boldsymbol{x}^{(i+1)} = \boldsymbol{t} + Conv(\boldsymbol{x}^{(i)}) \tag{6}$$

式中,$Causal(\cdot)$ 是因果卷积函数,$\sigma(\cdot)$ 是 Sigmoid 函数,$\boldsymbol{x}^{(i)}$ 是第 i 个特征向量融合块的输入,⊙代表逐元素相乘。

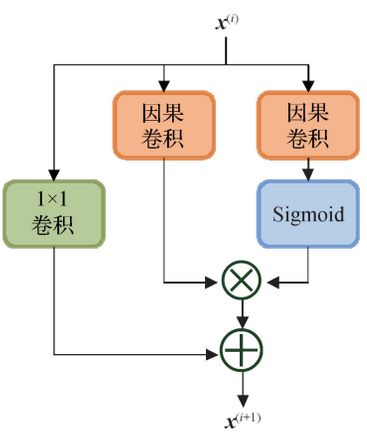


图 6　特征向量融合块的结构细节

Fig. 6　Structural details of the feature vector fusion block

2.4　特征金字塔增强和特征融合

FPEM 能够通过融合低级和高级信息来增强不同尺度的特征。FPEM 是可级联的模块,随着级联层数的增加,不同尺度的特征图会得到更充分的融合,特征图的感受野也随之增大。此外,因为 FPEM 是通过可分解卷积构建的,其计算开销非常小,仅为 FPN(feature pyramid networks)(Lin 等,2017)的 1/5 左右。

图 7 所示的 U 形模块由扩大尺度增强和缩小尺度增强两个阶段组成。扩大尺度增强作用在输入的特征图上,分别对边长为 32、16、8、4 个像素的特征图进行迭代增强。在缩小尺度增强阶段,输入是由扩大尺寸增强生成的特征金字塔,增强过程与扩大尺寸增强相反。缩小尺寸增强的输出即为 FPEM 的最终输出。

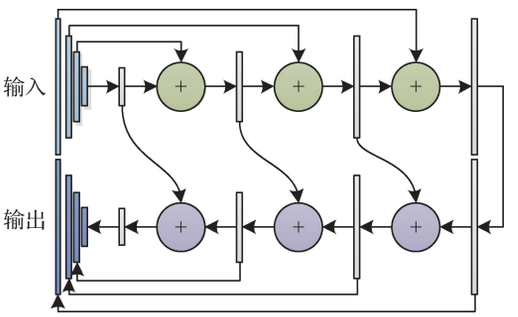


图 7　特征金字塔增强模块的结构细节

Fig. 7　Structural details of the feature pyramid enhancement module

特征融合模块用于融合各个模块输出的不同层次的特征,包括不同深度的特征金字塔输出的特征图以及特征向量融合模块输出的特征向量。如图 8 所示,通过逐元素加法组合相应比例的特征图,并与将特征向量进行上采样得到的特征图一起进行上采样操作,并级联为仅具有 5 × 128 通道的最终特征图。

文本区域保持了文本实例的完整形状,但是紧密放置文本实例的文本区域通常是重叠的,所以需要使用核区分文本实例(见图 9(a))。文字的核并不是完整的文本实例,如果要重建完整的文本实例(见图 9(b)),需要将文本区域中的像素合并到核中。本文采用了一种可学习的算法,即像素聚合(pixel aggregation),以指导文本像素被分类到正确的核。

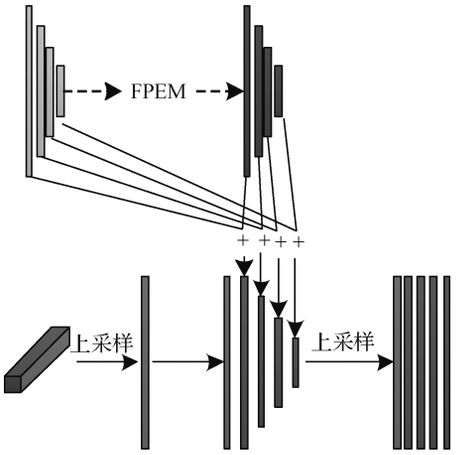


图 8　特征融合模块的结构细节

Fig. 8　Structural details of feature fusion module

2.5　像素聚合

在像素聚合中，借鉴了聚类的思想，从预测的核中重建完整的文本实例。将文本实例的核视为聚类中心，文本像素是要聚类的样本。为了将文本像素聚合到相应的内核，同一文本实例的文本像素与核之间的距离应较小。

在训练阶段，使用聚集损失 L_{agg} 和判别损失 L_{dis} 来评判像素聚合的效果，并用以训练。

在测试阶段，使用预测的相似性矢量将文本区域中的像素引导到相应的核。像素聚合的步骤如下：

1）在核的分段结果中找到连接的组件，每个连接的组件都是一个单独的核。

(a) 文本的核　　(b) 文本实例

图 9　对文本的核进行文本实例重建

Fig. 9　Reconstruct the text instance of the core of the text((a) the kernel of text; (b) the instance of text)

2）对于每个内核，有条件地将其相邻文本像素合并到预测文本区域中，使其相似度向量的欧几里得距离小于某个阈值。

3）重复步骤 2），直到没有符合条件的邻近文本像素。

2.6　损失函数

本文方法的损失函数可以表示为

$$L = L_{tex} + \alpha L_{ker} + \beta(L_{agg} + L_{dis}) \tag{7}$$

式中，L_{tex} 是文本区域的损失函数，L_{ker} 是核的损失函数，L_{agg} 是衡量文本实例中的像素和其对应核的损失函数，L_{dis} 是分辨不同文本实例的核的一个损失函数。α 和 β 用来平衡 L_{tex}、L_{ker}、L_{agg} 和 L_{dis} 的重要程度，L_{tex} 是模型的最终结果，重要程度最高，L_{ker} 用于评价核的分割结果，重要程度仅次于 L_{tex}，而 L_{agg} 和 L_{dis} 重要程度较低。按照其重要程度，α 和 β 分别设为 0.5 和 0.25。

考虑到文本和非文本像素在数量上非常不均衡，可以采用 dice loss(Milletari 等，2016)来监督文本区域的分割结果 P_{tex} 与核的分割结果 P_{ker}，因此 L_{tex} 和 L_{ker} 计算为

$$L_{tex} = 1 - \frac{2\sum_i P_{tex}(i)\, G_{tex}(i)}{\sum_i P_{tex}(i)^2 + \sum_i G_{tex}(i)^2} \tag{8}$$

$$L_{\text{ker}} = 1 - \frac{2\sum_i P_{\text{ker}}(i)\, G_{\text{ker}}(i)}{\sum_i P_{\text{ker}}(i)^2 + \sum_i G_{\text{ker}}(i)^2} \tag{9}$$

式中，$P_{\text{tex}}(i)$ 和 $G_{\text{tex}}(i)$ 分别指分割结果的第 i 个结果以及有标注的文本区域的准确性；类似地，$P_{\text{ker}}(i)$ 和 $G_{\text{ker}}(i)$ 分别指预测结果的第 i 个像素值以及核的准确性。

L_{agg}的作用是保证同一文本实例的核和文本实例内其他像素点之间的距离在一定范围内，计算为

$$L_{\text{agg}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\boldsymbol{T}_i|}\sum_{p\in\boldsymbol{T}_i}\ln(D(p,\boldsymbol{K}_i)+1) \tag{10}$$

式中，N是图像中文本实例的数量，$\boldsymbol{T}_i$表示第i个文本实例，$\boldsymbol{K}_i$是该文本实例的核。$D(p,\boldsymbol{K}_i)$代表文本实例$\boldsymbol{T}_i$内的像素p到相应的核$\boldsymbol{K}_i$的距离。

L_{dis}是用于不同文本实例的核的损失，其作用是保证任意两个核之间的距离不至于太小，其计算为

$$L_{\text{dis}} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\ln(D(\boldsymbol{K}_i,\boldsymbol{K}_j)+1) \tag{11}$$

式中，$D(\boldsymbol{K}_i,\boldsymbol{K}_j)$代表核$\boldsymbol{K}_i$与核$\boldsymbol{K}_j$之间的距离。该公式实际上是对每一个文本实例的核，分别计算与其他核的距离，然后进行累加。

3 实验与分析

本文方法基于 Pytorch 框架实现，在实验时使用一块 GPU 显卡（Nvidia 1080Ti）进行训练和测试。采用随机梯度下降算法进行优化，训练批大小为16，初始学习率设为 0.001，迭代训练 500 次，然后将学习率设为 0.000 01，迭代训练 100 次。

3.1 数据集

ICDAR（International Conference on Document Analysis and Recognition）2015 数据集是 ICDAR 发布的场景文本检测数据集。ICDAR 2015 由 1 000 幅训练图像和 500 幅测试图像组成，该数据集是以 4 个顶点的边界框的形式标注的。

CTW 1500 是一个具有挑战性的曲线文本数据集，由 1 000 幅训练图像和 500 幅测试图像组成。该数据集使用 14 个点标注的十四边形来表示曲线文本实例。

Total-Text 是一个用于曲线文本检测的数据集。该数据集包括水平文本实例、多方向文本实例和曲线文本实例，由 1 255 幅训练图像和 300 幅测试图像组成。

本文方法在各数据集上的检测结果如图 10 所示。

3.2 参数设置实验

为探究非局部模块嵌入位置的不同对结果的影响，本文在 ICDAR 2015 上对非局部模块的作用和嵌入的位置设计了对比实验，实验结果如表 1 所示。

由表 1 可知，嵌入非局部模块之后的效果要优于未嵌入模块的效果，说明非局部模块在检测过程中发挥了积极作用。按照检测精准程度来看，将非局部块嵌入 ResNet-18 的第 2 层、第 3 层和第 4 层的效果显著，而嵌入第 5 层的效果不明显。其中，将非局部块嵌入第 3 层的效果最好。按照检测时间来看，检测时间会随着嵌入位置的后移而延长，这是因为 ResNet-18 越往后其规模越大，对其进行非局部操作的复杂度越高。基于此实验结果，在本文方法中，非局部模块被嵌入至 ResNet-18 的第 3 层的最后一个残差块之前。

3.3 模型对比实验

本文方法与近年出现的其他方法在多个数据集上进行了对比。表 2 展示了多种方法在 ICDAR 2015 上的性能对比。由实验结果可知，PAN（pixel aggregation network）已经在准确度上达到较高水平，在速度上更是远远超过其他方法。而本文提出的 non-local PAN 的 F 值达到了 83.8%，已经超越了 PAN，与其他方法相比，本文方法可以实现较高的性能，并以更快的速度（23.1 帧/s）运行。

LOMO（look more than once）（Zhang 等，2019）在以上几种模型中准确率最高，但速度最慢。LOMO 的网络复杂，后处理方法烦琐，远不如 nonlocal PAN 的轻量分割头和像素聚合方法简单。由于非局部模块、轻量分割头和像素聚合等方法的共同作用，non-local PAN 在保证速度远远优于 LOMO 的基础上，还能在准确率上接近 LOMO 的水平。

本文方法与其他方法在弯曲文本数据集 CTW 1500 和 Total-Text 上的性能对比分别如表 3 和表 4 所示。

如表 3 和表 4 所示，本方法在 CTW 1500 和 Total-text 数据集上的综合性能已超过其他方法，F 值分别达到了 81.3% 和 86.3%。在检测速度上，PAN

(a) ICDAR 2015　(b) CTW 1500　(c) Total-Text

图 10　本文方法在多个数据集上的检测结果

Fig. 10　The detection results of the proposed method on multiple datasets((a) ICDAR 2015;(b) CTW 1500;(c) Total-Text)

表 1　非局部模块嵌入 ResNet-18 的不同位置时在 ICDAR 2015 上的检测结果对比

Table 1　Comparison of results on ICDAR 2015 when the non-local block is embedded in different positions of ResNet-18

嵌入位置	召回率/%	准确率/%	F 值/%	检测速度/(帧/s)
无	83.1	80.2	81.6	**26.1**
res2	84.2	81.1	82.6	25.2
res3	**84.5**	**81.4**	**82.9**	24.5
res4	84.1	81.5	82.7	23.1
res5	83.3	80.5	81.8	20.3

注:res2、res3、res4 和 res5 分别代表 ResNet-18 的第 2 层、第 3 层、第 4 层和第 5 层。F 值用于综合考虑召回率和准确率。加粗字体为每列最优值。

表 2　不同模型在 ICDAR 2015 数据集上的结果对比

Table 2　Comparison of different models on ICDAR 2015 dataset

方法	召回率/%	准确率/%	F 值/%	速度/(帧/s)
EAST(Zhou 等,2017)	73.5	83.6	78.2	13.2
DeepReg(He 等,2017b)	80.0	82.0	81.0	14.3
SegLink(Shi 等,2017)	76.8	73.1	75.0	13.9
SSTD(He 等,2017a)	73.9	80.2	76.9	7.7
TextSnake(Long 等,2018)	84.9	80.4	82.6	1.1
ATRR(Wang 等,2019b)	83.3	90.4	86.6	15.4
CRAFT(Baek 等,2019)	**84.3**	89.8	86.9	12.5
LOMO(Zhang 等,2019)	83.5	**91.3**	**87.2**	3.4
PAN(Wang 等,2019a)	81.9	84.0	82.9	**26.1**
本文	82.7	85.1	83.8	23.1

注:加粗字体为每列最优值。DeepReg 为 deep direct regression,SSTD 为 single shot text detector, ATRR 为 arbitrary text detection with adaptive text region representation, CRAFT 为 character region awareness for text detection。

的检测速度最快,本文方法次之,但远超过其他检测方法。

3.4　消融研究

为验证本文方法中不同模块在检测过程中发挥的作用,针对不同模块设计了消融实验,所有消融实验均在 ICDAR 2015 数据集上进行。

针对非局部模块,设计了 3 组实验:第 1 组实验不对骨干网络嵌入任何模块;第 2 组实验嵌入普通卷积模块;第 3 组实验嵌入非局部模块。第 1、2 组实验与第 3 组实验形成对照,分别探究删除、替换非

表 3　不同模型在 CTW 1500 数据集上的结果对比

Table 3　Comparison of different models on CTW 1500 dataset

方法	召回率/%	准确率/%	F 值/%	速度/(帧/s)
EAST(Zhou 等,2017)	49.1	78.8	60.4	21.2
SegLink(Shi 等,2017)	40.0	42.3	40.8	10.7
SSTD(He 等,2017a)	73.9	80.2	76.9	7.7
TextSnake(Long 等,2018)	67.9	85.3	75.6	5.6
ATRR(Wang 等,2019b)	80.2	80.1	80.1	22.5
CRAFT(Baek 等,2019)	**81.1**	86.0	**83.5**	19.7
LOMO(Zhang 等,2019)	69.6	**89.2**	78.4	4.4
PAN(Wang 等,2019a)	77.4	82.7	79.9	**84.2**
本文	78.9	83.8	81.3	71.8

注:加粗字体为每列最优值。

表 4　不同模型在 Total-Text 数据集上的结果对比

Table 4　Comparison of different models on Total-Text dataset

方法	召回率/%	准确率/%	F 值/%	速度/(帧/s)
EAST(Zhou 等,2017)	36.2	50.0	42.0	19.8
SegLink(Shi 等,2017)	23.8	30.3	26.7	9.1
TextSnake(Long 等,2018)	74.5	82.7	78.4	4.7
ATRR(Wang 等,2019b)	76.2	80.9	78.5	25.4
CRAFT(Baek 等,2019)	79.9	87.6	83.6	21.6
LOMO(Zhang 等,2019)	75.7	88.6	81.6	4.4
PAN(Wang 等,2019a)	81.0	89.3	85.0	**39.6**
本文	**82.9**	**89.9**	**86.3**	34.3

注:加粗字体为每列最优值。

局部模块对结果的影响。若第 3 组实验效果优于第 1、2 组的实验效果,则说明非局部模块在文本检测流程中发挥了不可或缺的作用。

为保持控制单一变量原则,普通卷积操作嵌入的位置与非局部模块嵌入位置相同,均嵌入到骨干网络 ResNet-18 第 3 层的最后一个残差块之前。

本文实验的运行环境以及参数设置均与原实验相同。不同嵌入模块的实验结果对比如表 5 所示。由实验结果可知:在准确度上,嵌入普通卷积模块的效果不如嵌入非局部模块的效果,甚至比未嵌入任何模块的效果还要差一些。在检测速度上,嵌入普通卷积模块要比嵌入非局部模块快一些,但比未嵌入模块要慢。这说明非局部模块不是普通卷积可以代替的,非局部模块表现出的优越性并非来自加深了网络,而是来源于它自身的结构特性。另一方面,ResNet-18 中嵌入普通卷积之后,在一定程度上破坏了残差网络的结构,使得检测效果甚至不如未嵌入任何模块。在模块的算法复杂度上,普通卷积操作比非局部操作略简单,所以其检测速度比嵌入非局部模块时略快。

表 5　不同嵌入模块效果对比

Table 5　Comparison of effects of different embedded modules

嵌入模块	召回率/%	准确率/%	F 值/%	检测速度/(帧/s)
无	80.2	83.1	81.6	**26.1**
普通卷积模块	79.5	82.2	80.8	25.2
非局部模块	**82.7**	**85.1**	**83.8**	23.1

注:加粗字体为每列最优值。

此外,本文设计了消融实验验证特征向量融合模块的可行性。表 6 展示了有无特征向量融合模块对最终检测结果的影响。由实验结果可知,特征向量融合模块在场景文字检测过程中发挥了正向的积极作用,融合了不同尺度的特征图,增强尺度多变的场景文字的特征表达。

表 6　有无特征向量融合模块网络检测结果对比

Table 6　Comparison of results with and without feature vector fusion module

特征向量融合模块	召回率/%	准确率/%	F 值/%	检测速度/(帧/s)
无	81.4	84.5	82.9	**24.5**
有	**82.7**	**85.1**	**83.8**	23.1

注:加粗字体为每列最优值。

4　结　论

本文提出了非局部像素聚合网络,该网络能够实现对任意形状场景文本的实时性检测。针对文本字符的特征,引入了非局部模块,将其嵌入到像素聚合网络的骨干网络中,使其能够捕捉到像素之间的内在关系,大大增强了提取特征的能力。此外,本文设计了一个特征向量融合模块,用于融合不同尺度的特征图,增强尺度多变的场景文字的特征表达。本文方法基于轻量级网络构建,并通过多个模块进

行特征增强,在检测速度和检测精度上都达到了较高水平。

本文提出的场景文本检测模型仍存在一些可改进之处。比如:本文方法用到了非局部操作,非局部操作是一次全局卷积操作,即卷积核的大小与特征图的大小相等,当特征图较大时,非局部操作的计算量也会非常大。如何简化模型结构还能保持捕捉远距离像素之间的关系,是该模型未来需要优化的方向之一。

参考文献(References)

Baek Y, Lee B, Han D, Yun S and Lee H. 2019. Character region awareness for text detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9357-9366 [DOI: 10.1109/CVPR.2019.00959]

Chng C K and Chan C S. 2017. Total-text: a comprehensive dataset for scene text detection and recognition//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Kyoto, Japan: IEEE: 935-942 [DOI: 10.1109/ICDAR.2017.157]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He P, Huang W L, He T, Zhu Q L, Qiao Y and Li X L. 2017a. Single shot text detector with regional attention//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3066-3074 [DOI: 10.1109/ICCV.2017.331]

He W H, Zhang X Y, Yin F and Liu C L. 2017b. Deep direct regression for multi-oriented scene text detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 745-753 [DOI: 10.1109/ICCV.2017.87]

Liao M H, Lyu P Y, He M H, Yao C, Wu W H and Bai X. 2021. Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2): 532-548 [DOI: 10.1109/TPAMI.2019.2937086]

Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S J. 2017. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]

Liu J C, Liu X B, Sheng J, Liang D, Li X and Liu Q J. 2019. Pyramid mask text detector[EB/OL]. [2020-11-07]. http://arxiv.org/pdf/1903.11800.pdf

Liu Y L, Jin L W, Zhang S T and Zhang S. 2017. Detecting curve text in the wild: new dataset and new solution[EB/OL]. [2020-11-07]. http://arxiv.org/pdf/1712.02170.pdf

Liu Z C, Lin G S, Yang S, Feng J S, Lin W S and Goh W L. 2018. Learning Markov clustering networks for scene text detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6936-6944 [DOI: 10.1109/CVPR.2018.00725]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Long S B, Ruan J Q, Zhang W J, He X, Wu W H and Yao C. 2018. TextSnake: a flexible representation for detecting text of arbitrary shapes//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 19-35 [DOI: 10.1007/978-3-030-01216-8_2]

Lyu P Y, Liao M H, Yao C, Wu W H and Bai X. 2018. Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 71-88 [DOI: 10.1007/978-3-030-01264-9_5]

Milletari F, Navab N and Ahmadi S A. 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation//Proceedings of the 4th International Conference on 3D Vision. Stanford, USA: IEEE: 565-571 [DOI: 10.1109/3DV.2016.79]

Qin S Y, Bissaco A, Raptis M, Fujii Y and Xiao Y. 2019. Towards unconstrained end-to-end text spotting//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4703-4713 [DOI: 10.1109/ICCV.2019.00480]

Shi B G, Bai X and Belongie S. 2017. Detecting oriented text in natural images by linking segments//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3482-3490 [DOI: 10.1109/CVPR.2017.371]

van der Oord A, Dieleman S, Zen H G, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A and Kavukcuoglu K. 2016. WaveNet: a generative model for raw audio[EB/OL]. [2020-11-07]. http://arxiv.org/pdf/1609.03499.pdf

Wang W H, Xie E Z, Song X G, Zang Y H, Wang W J, Lu T, Yu G and Shen C H. 2019a. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea(South): IEEE: 8439-8448 [DOI: 10.1109/ICCV.2019.00853]

Wang X B, Jiang Y Y, Luo Z B, Liu C L, Choi H and Kim S. 2019b. Arbitrary shape scene text detection with adaptive text region representation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6442-6451 [DOI: 10.1109/CVPR.2019.00661]

Wang X L, Girshick R B, Gupta A and He K M. 2018. Non-local neural

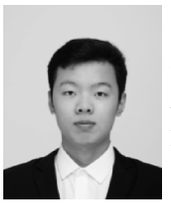
networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7794-7803 [DOI: 10.1109/CVPR.2018.00813]

Xu Y C, Fu M T, Wang Q M, Wang Y K, Chen K, Xia G S and Bai X. 2021. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(4): 1452-1459 [DOI: 10.1109/TPAMI.2020.2974745]

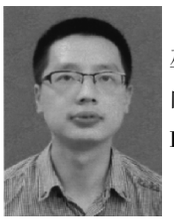
Zhang C Q, Liang B R, Huang Z M, En M Y, Han J Y, Ding E R and Ding X H. 2019. Look more than once: an accurate detector for text of arbitrary shapes//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10544-10553 [DOI: 10.1109/CVPR.2019.01080]

Zhou X Y, Yao C, Wen H, Wang Y Z, Zhou S C, He W R and Liang J J. 2017. EAST: an efficient and accurate scene text detector//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2642-2651 [DOI: 10.1109/CVPR.2017.283]

作者简介



师广琛,1998 年生,男,硕士研究生,主要研究方向为计算机视觉。
E-mail:shiguangchen@hhu.edu.cn



巫义锐,通信作者,男,副教授,主要研究方向为计算机视觉、模式识别与智慧水利。
E-mail: wuyirui@hhu.edu.cn