

基于实例分割的手写藏文文本行检测方法研究

贡保加¹ 安见才让¹
GONG Baojia Anjian Cairang

摘要

针对目前手写藏文文本检测数据稀少、文本行检测精度较低等问题,构建了一个脱机手写藏文文本图像数据集,并提出了一种改进 Mask R-CNN 的手写藏文文本行检测算法。为了更高效利用网络中不同尺度的特征,将特征提取网络中的 FPN 替换为 ASFF,可以动态学习特征中可用信息并且自适应融合多个级别的特征图,能够增强特征图中具有区分性的特征,过滤不同级别间的矛盾信息。然后在 Mask 分支添加 CAM 注意力机制,调整对每个通道的关注度,使得网络能够更好地聚焦于重要的特征。实验结果表明,改进的 Mask R-CNN 算法提高了对手写藏文文本行的检测精度,相比传统的 Mask R-CNN 算法 mAP 值提高了 6.9%。

关键词

实例分割; 手写藏文; 文本检测; 特征融合

doi: 10.3969/j.issn.1672-9528.2025.05.005

0 引言

文字在日常生活中扮演着重要的“角色”,与人们的生活紧密相连。无论是描摹周边的优美景致,阐释复杂的科学理论,还是记录生活中的琐碎细节,均需借助文字这一有效工具来实现。尤其在信息化时代,伴随互联网技术的飞速发展,大量文字信息不再仅依托于传统纸质媒介,而是以图像或视频等形式留存^[1]。

文档电子化是一种高效的信息保存方式,可使纸质文本摆脱纸张易损、难以长期保存的困境,实现更长久的留存。在此背景下,借助技术手段对图像、视频中的文字进行检测与识别的需求日益迫切,这有助于人们更全面、深入地挖掘其中蕴含的信息。

1 相关工作

印刷体电子文档中文字的检测方法已经发展较为成熟,相关技术在准确性、效率等方面都达到了一定水平。然而,当把目光聚焦到手写文本检测领域时,会发现这一领域仍然存在着许多问题等待解决。与印刷体文字相比,手写文字的独特性十分显著,其具有更强的随意性和不可控性。不同书写者因书写习惯、书写力度、审美偏好等因素的差异,所呈现出的字体笔迹各不相同,在姿态形状上,有的字体端正工整,有的则潇洒飘逸;在笔画粗细长短方

面,也各有千秋、差异明显。除此之外,手写文字还常面临上下行文字粘连、文本背景复杂以及文本行出现倾斜等情况,这些复杂的现象无疑给手写文本的行检测工作增加了很大的难度。

值得一提的是,在中文和英文手写文本行检测领域,经过众多研究者长时间的探索与实践,如今已有一些比较成熟的方法和技术,在应对各自语言手写文本的特点和常见问题上发挥积极有效的作用。但对于藏语这类少数民族语言来说,情况则有所不同。因藏语本身属于低资源语言,无论是在语料数据的积累,还是在相关研究的投入力度等方面,都相对薄弱,这导致其在水写文本检测这一领域的发展速度较为缓慢,目前仍处于不断探索和持续完善的阶段。

2008 年,肖人岳等人^[2]提出了一种文本检测算法,以游程分析为基础,通过对文字区域内空白部分的填充,截断文字行间过长的笔画,最终根据 3 个文本特征准则判断出文字行间的区域。2011 年,王丹等人^[3]提出了一种利用水平梯度差的文本行检测算法,对文本行的特点进行分析。在局部窗口求解最大梯度差并合并文本行区,通过滤除非文本区域来消除字符边界的变化,通过水平梯度差计算文档图像,最终以行块形式显示文档图像。2014 年,周鹏飞等人^[4]提出了一种基于模糊 BP 神经网络的算法,通过构造 3 个明确的模糊分类提高检测和分割结果的正确率。以上方法对简单的图像有着较为良好的检测效果。但是过多的人工制定的特征规则降低了算法本身在复杂图像上的鲁棒性。当深度学习技术在目标检测领域取得了显著成绩后,文本

1. 青海民族大学智能科学与工程学院 青海西宁 810007

[基金项目] 青海民族大学 2023 年度校级本硕博(学生)项目“基于深度学习的脱机手写藏文识别研究”(09M2023004)

行检测问题也有了新的发展。2016年, Tian等人^[5]提出了CTPN算法。CTPN由Faster R-CNN改进而成, 是以检测框为目标的检测网络, 其模型结构类似于Faster RNN, 但加入了支持任意尺寸图像输入的LSTM层, 可以在卷积层中直接定位文本行。同年, Zhang等人^[6]首次将全卷积神经网络(FCN)应用到文本检测任务上。2017年, Liao等人提出了Textboxes算法, 该算法通过修改卷积核和检测框的形状以适应文本区域多变的形状。2017年, Zhou等人^[7]提出的EAST算法是基于回归的文本行检测算法的代表, EAST取消了目标检测中间的候选区域聚合操作, 直接通过全卷积网络对文字区域坐标进行预测, 降低了传统算法的时间复杂度, 实现了文本行的实时检测。目前关于手写藏文文本行的检测研究较少, 更多的是对于藏文古籍文献的文本检测研究。本文使用基于实例分割的Mask R-CNN算法进行手写藏文的文本行检测。

2 Mask R-CNN 模型介绍

Mask R-CNN是由He等人^[8]在Faster R-CNN^[9]目标检测算法基础上提出的一种灵活而简便的目标实例分割算法框架, 可以同时完成检测任务和分割任务。Mask R-CNN采用了可以解决深度网络带来的梯度消失和梯度爆炸问题的ResNet残差网络来取代了VGG网络进行特征提取。Mask R-CNN模型为解决原RoI Pool层中两次量化操作引起的检测框区域误差, 使用基于双线性插值法的RoI Align层解决了其中的位置偏移问题。进而能更加精确地完成目标检测任务。Mask R-CNN还引入了FCN(fully convolution networks)层来实现图像的语义分割。通过FCN层Mask R-CNN可以独立完成类别预测任务和掩膜预测任务, 针对特征图中每一个感兴趣的候选区域产生一个二值掩膜, 从而避免了不同类别之间掩膜预测的冲突问题。在未使用数据增强方法情况下Mask R-CNN的检测效果就优于当时所有的单阶段检测算法, 如图1所示。

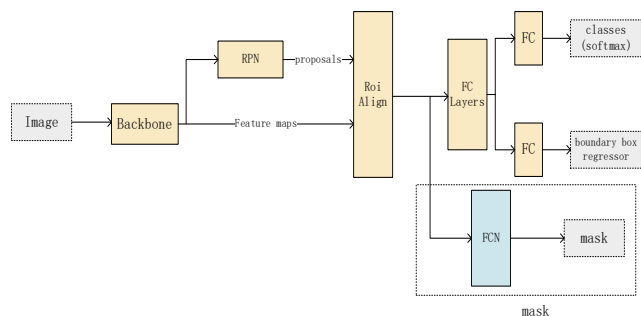


图1 Mask R-CNN 模型结构

3 模型改进策略

藏文作为一种低资源语言, 在数据方面无法与其他主

流的语言相比, 本文研究的手写藏文文本检测而言, 收集的图像数据质量参差不齐, 文本内容大部分以小目标为主, 同时还存在检测目标之间存在互相粘连和互相重叠, 上述此类问题导致在原模型的检测过程中卷积网络特征提取不够精细, 模型难以准确捕捉并识别小目标的细节特征; 目标之间存在遮挡或重叠时, 模型会因为目标特征变得模糊或混淆而无法准确区分和定位目标区域。针对以上问题, 本文对Mask R-CNN模型进行了改进和优化, 主要通过使用更深层次的网络结构、引入多尺度注意力机制以及优化边界框损失函数。改进旨在提高模型对版面布局检测的上下文信息感知能力, 同时抑制无用特征信息, 使模型更专注于目标信息。

3.1 Backbone 改进

为充分利用手写藏文文本行检测和分割任务中不同尺度的特征, 本文用ASFF取代了Backbone中的FPN层, 更好地利用了卷积层输出的不同尺度的特征图, 如图2所示。

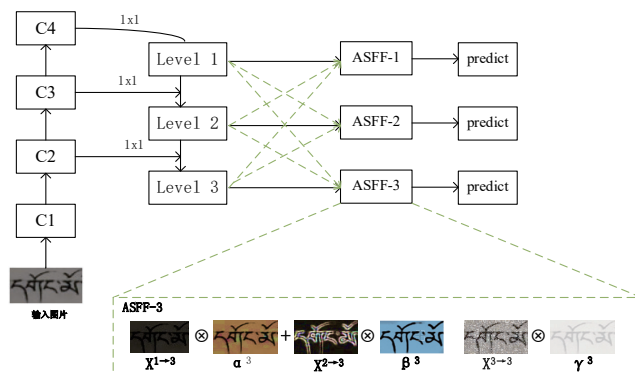


图2 引入ASFF的Backbone

ASFF(adaptive structure feature fusion)网络是一种用于目标检测的神经网络模型^[10]。旨在提高多尺度特征融合的效果, 从而改善目标检测的准确性和性能。ASFF网络引入了自适应结构特征融合模块, 用于融合多个不同层次的特征图。传统的目标检测网络通常只使用单一层次的特征图, 而ASFF网络通过动态地选择和融合多层特征图, 使得网络能够更好地捕捉目标的多尺度信息。

3.2 Mask 分支改进

掩码分支(Mask Branches)是在类标签分支和边界框识别分支的基础上, 通过添加一个新的分支与现有的用于边界框识别的分支并行来预测目标掩码Mask。但在进行目标掩码预测时, Mask分支并不能直接从特征图中准确关注到掩码分割所需要的有用的信息, 增加有用特征信息的权重, 为解决这一问题, 对特征图中的有用信息给予更多的权重, 本文在Mask分支上额外引入了通道注意力机制, 从而提高了手写

藏文文本行检测和分割的准确性。通道注意力机制（channel attention mechanism, CAM）通过为每个通道分配一个不同的权重，来增强重要通道的表达能力，减少不重要通道的影响。这种机制的核心思想就是通过调整对每个通道的关注度，使网络能够更好地聚焦于重要的特征^[11]。

4 数据集构建

在计算机视觉领域的目标检测任务中一个高质量的数据集起着至关重要的作用。数据集可以提供各类图像和对应目标物体的位置以及类别的标注信息^[12]。研究者可以将这些标注的数据去训练和微调目标检测模型，从而让模型可以识别和定位不同类别的对象。现阶段还没有公开且开源的脱机手写藏文文本数据集，实验所用所有图像数据为人工拍摄和采集。最初搜集的文本图像数量为 740 张手写藏文图像，因拍摄者的拍摄设备、拍摄技巧原因筛选掉了其中一部分质量不佳的图片，由于文本书写者中有部分中小学生，因书写文本过于潦草，导致许多文字无法辨认，又筛选掉了其中一部分图片。最终经过人工裁剪和处理后数据集中共有图像 523 张。数据集中包含上下行文字粘连、背景和文字边界模糊、手写文本行倾斜和弯曲等问题。

4.1 数据标注

数据集中的图像均来自甘肃和青海地区的藏族学生书写，手写文本图像多数为学生日常作业和笔记，少部分为专门摘抄的书本内容。所有图像经过人工筛选和裁剪，预处理之后再行统一标注。

数据标注采用的工具为 labelme，labelme 是一个基于 Python 编写的图形图像注释工具，可以对图像进行多种形状和样式的标注，其标注的图像广泛应用于目标检测、图像分割等任务。图像标注时标签分为 background 和 line，其中 background 表示图像背景，line 表示图像中被标注的文本行，图 3 为 labelme 图像界面文本行标注图像，图 4 是经过标注后的图像，明确表示了图像的掩码信息和标签信息。其标注后的信息可保存为 json 和 txt 文件，可转换为 coco 数据集格式和 voc 数据格式。

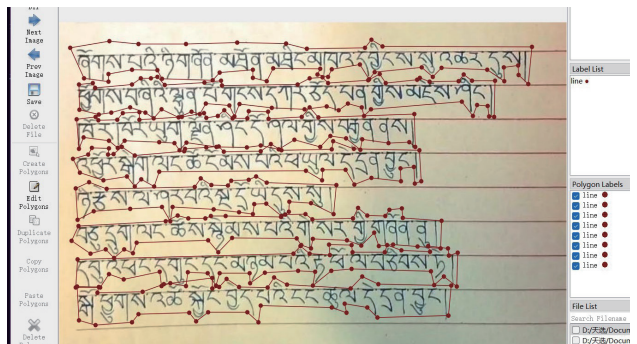


图 3 图像标注实例

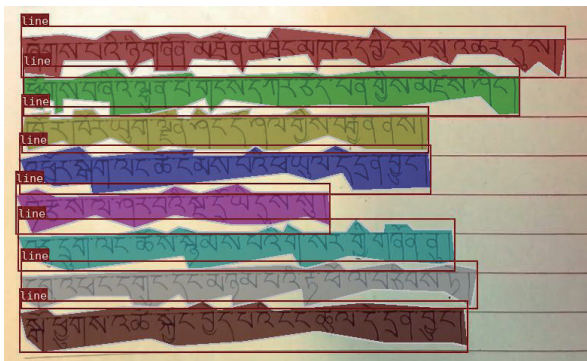


图 4 文本行标注结果

4.2 数据增强

为保证输入模型的数据质量可以提升模型的表现，并让模型对输入的数据有更好的泛化能力。通过图像预处理将原始图片转换成适合输入到模型中的格式。首先将图像的大小和格式进行统一，本文将手写藏文文本图像缩放到固定的尺寸 1 280 px × 720 px，并且按照顺序对图像进行命名。在实际使用场景中，手写藏文因文字的 shape、大小、笔画粗细、书写风格等方面具有多样性。不同的书写风格、习惯和技巧都会导致藏文字的外观差异。由于书写的自由度较高，手写文字可能存在变体和变形。同一个藏文实际在不同的作者之间也存在差异，例如笔画的长度、倾斜角度等。所以，为更好地扩充数据集，同时还原真实使用场景，本文使用基于几何类型的数据增强对数据集进行预处理。

基于几何类型的数据增强是一种常用的数据增强技术，通过图像的几何变换来生成新的训练样本^[13]。几何类型的数据增强可以通过应用平移、旋转、缩放、翻转等几何变换操作，生成大量的新样本。可以显著增加训练数据的规模，提供更多的样本用于模型的训练，从而减轻过拟合问题。因此，本文采用左旋转、右旋转、翻转、水平移动、垂直移动等数据增强方式对数据集进行增强，以此模拟真实世界中的各种变形和扭曲情况，减轻过拟合问题使模型具备更好的鲁棒性。

5 实验结果

在本文的实验中，使用原始数据集中的 523 个样本。为扩充数据集并增加样本的多样性，采用了上述提到的 5 种数据增强方式。经过数据增强后，数据集的样本总数增加到 2 615 个。为了进行模型的训练和评估，将数据集按照 8:2 的比例进行划分，其中训练集包含 2 092 张样本，验证集包含 523 张样本。

将改进后的脱机手写藏文文本行检测模型与 Faster R-CNN、YOLOv5 以及 Mask R-CNN 经典的检测模型进行对比分析，本文在自建的脱机手写藏文文本行检测数据集上进行验证，实验结果如表 1 所示。

表 1 同模型的对比结果

模型	<i>P</i>	<i>R</i>	mAP
Faster R-CNN	0.63	0.654	0.614
YOLOv5	0.652	0.687	0.621
Mask R-CNN	0.69	0.706	0.745
本文方法	0.836	0.725	0.814

根据表 1 所示，发现 Mask R-CNN 模型在精度、召回率和平均精度均值方面表现优于 Faster R-CNN 模型、YOLOv5 模型。本文基于改进 Mask R-CNN 模型的方法在精度、召回率和平均精度均值分别提升了 14.6%、1.9% 和 6.9%。

通过消融实验验证相关改进措施在脱机手写藏文图像上的检测影响，以 Mask R-CNN 为基准模型，分别引入本文所提出的改进机制，对相应的模块评估，如表 2 所示，表中“√”代表在 Mask R-CNN 基础上引入的相应模块。

表 2 消融实验

序号	Mask R-CNN	ASFF	CAM	<i>P</i>	<i>R</i>	mAP
1	√			0.69	0.706	0.645
2	√	√		0.745	0.715	0.715
3	√		√	0.72	0.708	0.697
4	√	√	√	0.836	0.725	0.814

通过表 2 可以直观地看到 Mask R-CNN 在引进 ASFF 模块和 CAM 模块之后各项数值都有了明显的提高。单纯引入 ASFF 模块的 *P* 值、*R* 值、mAP 值分别提高了 5.5%、0.9%、6.1%；引入 CAM 模块的 *P* 值、*R* 值、mAP 值分别提高了 3%、2%、2.4%；同时引入两个模块的 *P* 值、*R* 值、mAP 值分别提高了 14.6%、1.9%、13.1%。

6 总结

本文以 Mask R-CNN 为基线模型，通过引入 CAM 注意力机制模块和 ASFF 模块，显著提升了模型在手写藏文文本行检测任务上的性能。通过与其他经典基线模型进行对比消融实验，验证了本文改进后的文本行检测模型的优越性能。该模型精确率（Precision）为 83.6%、召回率（Recall）为 72.5%、平均精度（mAP）为 81.4%。基本满足了手写藏文文本行检测任务的需求。

参考文献：

- [1] 徐莹. 基于实例分割方法的场景文本检测算法研究 [J]. 现代计算机, 2022,28(6):73-77.
- [2] 肖人岳, 秦慕婷. 一种复杂文本图像中快速文本行检测算法 [J]. 科学技术与工程, 2008,8(23):6253-6257.
- [3] 王丹, 王希常, 杨侠. 一种基于梯度差的文档图像文本行

检测算法 [J]. 微型机与应用, 2011,30(18):32-34.

- [4] 周鹏飞, 张月琴. 基于模糊 BP 神经网络的图像分割方法研究 [J]. 计算机工程与设计, 2014,35(7):2423-2426.
- [5] TIAN Z, HUANG W L, HE T, et al. Detecting text in natural image with connectionist text proposal network[EB/OL].(2016-09-12)[2024-11-12].<https://doi.org/10.48550/arXiv.1609.03605>.
- [6] ZHANG Z, ZHANG C Q, SHEN W, et al. Multi-oriented text detection with fully convolutional networks[EB/OL].(2016-04-16)[2024-11-16].<https://doi.org/10.48550/arXiv.1604.04018>.
- [7] ZHOU X Y, YAO C, WEN H,et al. EAST:an efficient and accurate scene text detector[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2017:5551-5560.
- [8] HE K M, GKIOXARI G, DOLLAR P,et al. Mask R-CNN[C]//2017 IEEE International Conferenceon Computer Vision(ICCV). Piscataway:IEEE,2020:386-397.
- [9] REN S Q, HE K M, GIRSHICK R,et al. Faster R-CNN:towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis&machine intelligence, 2017, 39(6): 1137-1149.
- [10] 陈志旺, 李宗轩, 吕昌昊, 等. 基于 HRNet 和 ASFF 的特征融合目标检测算法 [J]. 控制与决策, 2024,39(10):3207-3215.
- [11] 王维兰, 胡金水, 魏宏喜, 等. 少数民族文字文本分析与识别的研究进展 [J]. 中国图象图形学报, 2024,29(6):1685-1713.
- [12] 杨业辉, 范菁, 钟声. 基于注意力机制改进 YOLOv5 的茶叶病害检测方法 [J/OL]. 云南民族大学学报 (自然科学版), 1-10[2025-02-10].<http://kns.cnki.net/kcms/detail/53.1192.N.20250312.1834.004.html>.
- [13] 李萍. 基于深度学习的离线手写汉字文本行识别与笔迹鉴定研究 [D]. 呼和浩特: 内蒙古大学, 2024.

【作者简介】

贡保加（2000—），男，甘肃甘南人，硕士，研究方向：藏文信息处理及应用。

安见才让（1969—），男，青海西宁人，硕士，教授，研究方向：藏文信息处理和分布式网络。

（收稿日期：2025-03-20 修回日期：2025-05-06）