计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201000154

结合多粒度特征融合的自然场景文本检测方法

陈 卓 王国胤 刘 群
重庆邮电大学计算智能重庆市重点实验室 重庆 400065
(512619302@qq.com)

摘 要 自然场景下的文本信息通常具有多样性和复杂性的特点。由于采用手工设计特征的方式,传统的自然场景文字检测方法缺乏鲁棒性,而已有的基于深度学习的文本检测方法在各层网络提取特征的过程中存在丢失重要特征信息的问题。文中从多粒度和认知学的角度,提出了一种结合多粒度特征融合的自然场景文本检测方法。该方法的主要贡献是通过对通用特征提取网络的不同粒度特征进行融合,并加入残差通道注意力机制,使得模型在充分学习图像中不同粒度特征信息的基础上,更加关注目标特征信息并抑制无用的信息,提升了模型的鲁棒性和准确率。实验结果表明,相比其他最新的方法,该方法在公开数据集上取得了85.3%的准确率和82.53%的F值,具有更好的性能。

关键词:特征提取;多粒度信息;残差注意力;卷积神经网络
中图法分类号 TP391

Natural Scene Text Detection Algorithm Combining Multi-granularity Feature Fusion

CHEN Zhuo,WANG Guo-yin and LIU Qun
Chongqing Key Laboratory of Computational Intelligence,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

Abstract In natural scenes,text information usually has the characteristics of diversity and complexity. Due to the way of manually designing features,traditional natural scene text detection methods lack robustness,and the existing text detection methods based on deep learning have the problem of losing important feature information in the process of extracting features in each layer of the network. This paper proposes a natural scene text detection method combined with multi-granularity feature fusion. The main contribution of this method is that by combining the features of different granularities in the general feature extraction network and adding the residual channel attention mechanism,the model can pay more attention to the target feature information and suppress useless information on the basis of fully learning the feature information of different granularities in the image,and this method improves the robustness and accuracy of the model. The experimental results show that,compared with other latest methods,the model has achieved 85.3% accuracy and 82.53% F-value on public datasets,and has better performance.

Keywords Feature extraction,Multi-granularity information,Residual attention,Convolutional neural network

1 介绍

图像作为一种重要的信息传播媒介,逐渐成为了日常生活中必不可少的信息载体。但是,由于图像的复杂性和多样性,如何有效提取图像中的各类信息(如人脸、物体、文字等)已经成为当前的研究热点和难点。自然场景图像中的文字具有比其他各类信息更丰富的语义信息,因此准确的文字检测方法有助于充分理解和分析图像中的场景内容。图像文字检测识别技术也被广泛应用于图像搜索、自动驾驶等方向的研究,并被相关公司运用于图文识别、卡证识别、试卷识别、车牌识别等实际场景中。

目前,针对自然场景文本处理的研究工作主要使用两大类方法。第一类是传统的自然场景文本检测方法,该方法又分为基于像素连通域分析的方法和基于滑动检测窗口的方法,这类方法主要依赖于图像的像素和文本的形状、排列、笔画宽度等特征,首先获得文本候选区域,然后采用手动设计的特征对所获得的候选区域进行验证,以此确定图像中的文本信息区域。另一类是基于深度学习的自然场景文本检测方法,该方法通过神经网络模型组合低层特征来形成高层特征以表示属性类别,并设计专用的损失函数让计算机自动并精准学习图像中文字信息的特征。但是,由于目前对深度神经网络模型的鲁棒性要求越来越高,针对自然场景图像中复杂多变的文字检测问题,现有的研究方法在速度和精度上都有很大的提升空间。本文提出了一种结合多粒度特征融合的自然场景文本检测方法(Natural Scene Text Detection Algorithm Combining Multi-granularity Feature Fusion,TDMF),

到稿日期:2020-10-26 返修日期:2021-04-03
基金项目:国家自然科学重点基金项目(61936001)
This work was supported by the Key Program of National Natural Science Foundation of China(61936001).
通信作者:刘群(liuqun@cqupt.edu.cn)

该方法首先通过通用的特征提取网络和图像中多个粒度的信息，并利用残差通道注意力机制减少融合过程中的冗余信息。通过对不同粒度的特征进行关注和融合，有效提高了文本检测的效果。本文的主要贡献如下：1）构建了一个基于卷积神经网络的自然场景文本检测模型，该模型直接生成预测的文本框，再通过非极大值抑制算法输出结果，简化了训练过程，提高了训练速度；2）结合多粒度认知计算，对特征提取网络多个不同粒度信息的特征进行融合，使模型保留图像中不同感受野的目标特征；3）加入残差通道注意力机制，使模型更加关注多粒度特征融合，防止融合过程中的特征信息丢失，提高了模型的预测准确度。

本文第 2 节讨论了相关工作；第 3 节描述了 TDMF 模型架构；第 4 节介绍了实验设置和评估结果的对比；最后总结全文并展望未来。

2 相关工作

与印刷文档图像中的文字检测不同，自然场景下图像中的文字不再是规范或者背景单一的，而是普遍存在着水平、倾斜、弯曲等各种不定的形状和方向，文本检测对象已经从最开始的经过手工挑选的单一对象转变成了自然场景下的多样对象。从自然场景文本检测方法中采用的学习文本特征的模型来看，基于深度学习的自然场景文本检测方法因较高的精度和较快的速度正在逐渐取代原有的基于手工设计特征的传统方法，如卷积神经网络（Convolutional Neural Networks，CNN）以及递归神经网络（Recurrent Neural Networks，RNN）等在自然场景文本检测领域得到了很好的应用，并且由于神经网络模拟了人脑认知事物机理，相比传统方法省去了繁琐低效的人工特征工程，因此其检测性能有了很大的提升。

2.1 基于手工设计特征的传统方法

基于手工特征设计的自然场景文本检测方法主要分为两种：基于连通域分析的方法以及基于滑动检测窗口的方法。这种传统的自然场景文本检测方法主要依赖于图像的像素以及文本的形状、笔画宽度等物理特征，利用设计好的模型提取出文字的候选区域，然后采用手动设计的特征对所获得的候选区域进行验证，以此确定图像中的文本信息区域。

基于像素连通域的分析方法主要采用自下而上的思想来检测文本。依据像素的特征信息，搜索单个像素或单个文本以形成与相邻像素或文本相关联的文本行，可以将这类方法分为边缘检测方法和文本水平检测方法。边缘检测方法可以使用自然场景文本的丰富边角信息来获取文本候选区域，并通过分类器或者分类规则对文本候选区进行分类。Cho 等[1]采用 Canny 边缘检测算子检索图像的边缘信息，提取字符候选区域，然后使用 AdaBoost 级联分类器得到强候选字符和弱候选字符，最后通过滞后文本跟踪找出可靠的检测字符，获得文本候选区域。文本水平检测方法则利用了自然场景图像的像素信息。Neumann 等[2]首先采用最大稳定极值区域提取出图像中的文本候选区域，再通过根据字符的特征（长宽比、分段高度、紧密度、孔数、字符颜色一致性、背景颜色一致性等）训练好的分类器，最终得到检测的文本区域。由于这类算法对像素特征产生的文本连通域具有依赖性，连通区域的分割结果将会直接影响模型的性能，因此基于像素连通域的方法在面对自然场景图像中光照强度、颜色变化等复杂因素时会表现出检测精度低、鲁棒性不足的情况。

与像素连通区域的分析方法相反，基于滑动检测窗口的方法采用自上而下的思想来检测文本。这种类型的方法使用滑动检测窗口扫描整个图像以获得图像的全部特征。Tian 等[3]提出的统一文本检测系统采用结合滑动检测窗口和快速级联的方法来检测候选字符，采用 6 个像素信息（像素紧密度、水平和垂直梯度以及二阶梯度等）来加速特征的提取过程。该方法在每个像素位置计算特征，并将其串联为用于增强学习的特征向量，最后采用一个最小成本的流网络对提取的特征进行置信度计算，输出文本检测区域。Wang 等[4]提出一种基于特定单词的自然场景文本检测算法，首先通过滑动检测窗口获得单个候选的文本，然后根据多个相邻的候选文本间的特征计算出可能的文本组合的评分，最后从评分的集合中求出最相近的文本组合作为最终的文本检测区域。基于滑动检测窗口的方法采用人为设计的手工特征来区分目标区域和背景区域，复杂的描述特征加大了模型的计算强度，进而导致检测效率不尽人意。另外，自然场景图像中的文字排列不规律，如横行、竖列、斜体及弯曲字体等，滑动检测窗口方法无法对这些不规则排列实现较好的检测效果。

2.2 基于深度学习的自然场景文本检测方法

由于传统的自然场景文本检测方法在确定文本区域的过程中包含手工设计的特征分类，这种方式不仅包含大量的人工特征工程，而且在检测性能上具有比较明显的缺点。深度学习作为神经网络的深层次发展，通过神经网络模型组合低层特征从而形成高层特征来表示属性类别，让计算机自动并精准学习图像中文字信息的特征。根据目前使用的广泛程度，基于深度学习的自然场景文本检测方法大致可以分为两类：基于区域建议的文本检测方法和基于图像分割的文本检测方法。

基于区域建议的文本检测算法通过回归文本框的方式来获得图像中的文本信息。Tian 等[5]提出的 CTPN 由 Faster-RCNN 改进而来，该方法首先用 3×3 的滑动窗口提取特征网络的最后一层卷积图的特征，随后将这些序列化的向量输入双向 LSTM 中，并通过全连接层输出文本区域的预测得分、anchor 的 y 轴坐标和边细化偏移量，最后模型输出序列化固定宽度的文本建议区域。Shi 等[6]提出的 SegLink 使用改进版的 SSD 来解决多方向的文字检测问题。该方法首先将图像输入到改进的 SSD 网络中，并输出字符级或者单词级的检测框信息以及不同检测框的连接信息，最后使用融合算法得到文本或者单词的检测区域。Xu 等[7]提出的分层检测方法先采用 CNN 提取特征，然后使用最大稳定极值区域获得种子文本，并依据种子文本来定位其他退化的文本区域，最后采用随机森林结合文本行的上下文信息精细地分类文本候选区域。Wang 等[8]提出的 ContourNet 首先利用 Adaptive-RPN 模块根据主干网络的输出回归生成高精度候选框，然后通过 LOTM 算法来解耦候选框中的水平和竖直方向的文本轮廓，最后计算生成更精确的任意形状文本框。基于图像分割的文本检测方法通常利用全卷积网络等方式进行像素级的文本标注，该类方法可以很好避免图像中文本的排列、文本框形状等

的影响，但后续特征提取的处理过程具有较高的复杂性。Yang 等[9]提出的模型首先采用类似于 VGG16 的卷积特征提取网络来提取图像特征，然后将特征输入全卷积网络以预测每个像素点的置信度，最后通过一个加入注意力机制的字符来识别网络并输出文本的预测区域。Wang 等[10]提出的 PSNet 模型利用分割的方式对提取出的特征进行像素分类，然后采用渐进尺度扩展算法将分类的 4 个结果扩增到原始文本行大小，得到任意形状文本的检测结果。Baek 等[11]提出了一种检测字符级的弱监督学习模型，首先利用高斯热度图对训练集图片中的每个文本注释生成字符级边框，以供模型进行预训练，然后通过连通区域标记法，根据预训练模型的输出结果计算出多边框文本区域。

基于深度学习的自然场景文本检测算法虽然在检测结果的准确度和效率上都有较好的表现，但是模型通常包含多个处理步骤，每个步骤对检测结果都起着至关重要的作用，不合理的模型设计将严重影响检测结果和效率。同时，数据集的大小也会直接影响模型的性能，较小的数据样本会导致训练过程中出现过拟合进而生成不理想的检测模型。此外，目前深度神经网络的可解释性也是一个研究热点和难题。

2.3　多粒度认知计算

人类对世界的认知过程是一种从不同层次逐级认识的机制，在认知计算中被称为粒计算。粒计算中具体化的对象被称为粒，用来表示或解释对象的结构被称为粒结构。中国科学院生物物理研究所陈霖院士等通过实验研究发现，人类认知具有“大范围优先”的规律，视觉系统对全局拓扑特性尤为敏感[12]，即人类的认知规律是一个从粗粒度到细粒度的过程。

在图像识别的过程中，模型通过提取局部的像素特征和全局的图像特征来对原始图像进行识别分类。粒计算中，局部的像素特征属于最细的粒度特征，而图像的整体特征则属于最粗的粒度特征。算法或者模型对图像特征进行学习的过程，其实就是一个从细粒度到粗粒度的识别过程。而人类对一件事物的认知过程则恰好相反，通过对整体问题或者事物的分解来得到每一部分的认知结果，是一个从粗粒度到细粒度的识别过程。不管是粗粒度的信息还是细粒度的信息，都对事物的认知起到了至关重要的作用，因此结合不同粒度的特征将更有利于捕获图像的文本区域信息。

3　结合多粒度特征融合的自然场景文本检测方法

本节将介绍本文提出的一种结合多粒度特征融合的自然场景文本检测方法（TDMF），模型的详细架构如图 1 所示。该模型采用已经训练好的通用特征提取网络对输入的图片进行特征提取，结合多粒度认知思想，利用上采样的方式逐层进行特征提取，得到不同的粒度信息。然后将提取到的特征进行融合，输出每个像素点的预测得分图。预测得分图在附近像素的几何结构趋于高度相关的情况下加权合并几何体，再通过筛选过滤掉非文字边界框，得到最后的文字检测边界框。自然场景文本检测系统通常包含多个处理环节，神经网络模型由于功能的分散性，往往由多个不同功能的神经网络组成，因此结构相对复杂。TDMF 模型仅由两个步骤组成，避免了多个处理环节和多个组成部分可能局部最优但整体未必最优以及耗时的问题，使模型在保证检测准确率的同时尽可能地减少耗时。

图 1　多粒度特征融合的自然场景文本检测模型

Fig. 1　Natural scene text detection model combining multi-granular feature fusion

3.1　基础特征提取网络

TDMF 主干特征提取网络采用了 VGG-16[13]网络模型，对于输入的图片，在给定感受野的情况下，VGG-16 采用小卷积核取代大卷积核来减少模型的训练参数，同时通过增加多层非线性层来扩展网络深度以保证学习更复杂的模式。相比 ResNet[14]网络的 50 层复杂网络结构，VGG-16 需要训练的参数更少，减少了计算量。其网络结构如图 2 所示，VGG-16 包含 5 个 Block，13 个卷积层，卷积核大小均为 3×3，使每个卷积层与前一层保持相同的宽和高，5 个 Block 卷积提取出的特征最后通过 3 个全连接层输出。结构中采用的池化层均采用上一层 1/2 的输出，使特征的宽和高在每个 Block 中依次减半。

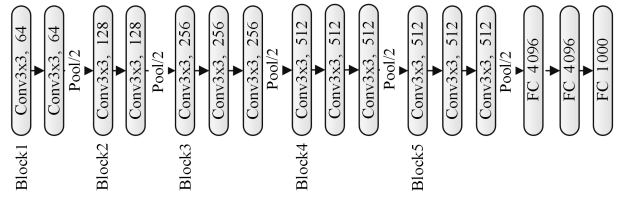

图 2　VGG-16 特征提取网络

Fig. 2　VGG-16 feature extraction network

3.2　多粒度特征融合

基于基础特征网络提取出的特征，从认知计算的角度出发，提取不同卷积层的粒度特征，其尺寸大小分别为输入图像的 1/4，1/8，1/16，1/32，这样可以得到由细粒度到粗粒度的图像特征。

多粒度特征融合过程如图 3 所示，首先将提取的 4 个粒度中最细的粒度层进行上采样处理，将其扩大为图像的1/16，与次细粒度特征图大小相等；然后采用两个大小分别为 1×1 和 3×3 的卷积核进行卷积后，将两个粒度进行特征融合，并以此方式逐层往上将特征两两融合；最终得到 3 张由不同粒度融合而成的特征图。依层次进行粒度融合的方式使 VGG-16 网络提取出来的特征图联系更紧密，让神经网络在训练模型的过程中能够更好地学习到不同粒度信息之间的关系。在得到 3 张融合后的特征图之后，将充分提取到图片特征的特征图进行连接，最后经过大小为 1×1 的卷积核卷积操作后输出得到文字与非文字的预测得分图。

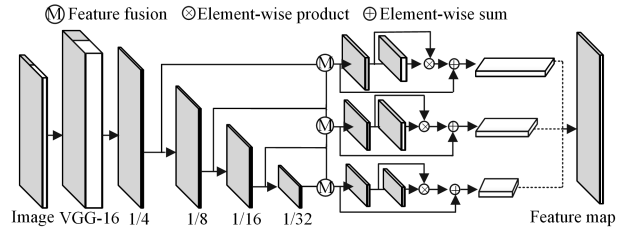

图 3　多粒度特征融合网络

Fig. 3　Multi-granularity feature fusion network

3.3 加入残差通道注意力的多粒度信息提取

在利用 VGG-16 网络进行特征提取后，分别提取其全连接层的前 1/4，1/8，1/16，1/32 的特征图，并经过 ReLU 函数激活，其表达式如下：

$$\boldsymbol{f}_i=\sigma(Conv(\boldsymbol{f}_i)) \tag{1}$$

其中，$i\in\{1,2,3,4\}$，$\boldsymbol{f}_i$ 为下采样过程中的特征。在多粒度信息提取过程中，先利用 3.2 节中的多粒度特征融合方式得到采样过程中上下连接的 3 个多粒度特征 $\boldsymbol{h}_i$，其表达式如下：

$$\boldsymbol{h}_j=concat(Unpool(\boldsymbol{f}_{j-1}),\boldsymbol{f}_j) \tag{2}$$

其中，$j\in\{2,3,4\}$，$concat(\cdot)$表示向量的拼接操作。另外，经过特征融合后的信息存在冗余问题，特别是在粒度信息提取时，必然存在重要度不同的像素信息。此时，重要度低的特征相对于文本信息来说变成了噪声，同时也会造成网络的冗余。因此，在进行粒度特征融合时加入基于通道注意力改进的残差通道注意力网络 RCAN[15]，不仅可以提升网络对文本信息的聚焦，也可以通过残差思想结合上下有联系的特征，防止有用特征的丢失。其表达式如下：

$$CA^i=\sigma(W^i\boldsymbol{h}^i+\boldsymbol{b}^i) \tag{3}$$

$$RCA^j=\boldsymbol{h}^j+CA_h^j \tag{4}$$

其中，$i,j\in\{1,2,3\}$，$\boldsymbol{h}^*$ 代表进行粒度融合后的特征向量，W^i，W^j 代表可学习的参数，$\sigma(\cdot)$代表 sigmoid 激活函数，CA^i 和 RCA^j 分别代表生成的通道注意力权重值和残差通道注意力权重值。最后，将所得的 3 个多粒度特征向量进行融合得到最终的预测特征，表达式如下：

$$Z=\sigma(Pool(concat(RCA^1,RCA^2,RCA^3))) \tag{5}$$

其中，$concat(\cdot)$表示向量的拼接操作，$Pool(\cdot)$表示最大池化，$\sigma(\cdot)$表示 ReLU 激活函数。

3.4 损失函数设计

整个模型的损失函数可表示为：

$$L=L_s+\lambda_g L_g \tag{6}$$

其中，本文将 λ_g 设为 1，L_s 和 L_g 分别表示分数图和几何图形的损失。文本区域的预测可以看作是一个像素级的二分类问题，因此 TDMF 采用与评价指标 IoU 度量相似的 Dice 损失函数来优化得分图。Dice 系数是一种集合相似度度量函数，通常用于计算两个样本的相似度，其表达式如下：

$$s=\frac{2|X\cap Y|}{|X|+|Y|} \tag{7}$$

因此，最终的分数图损失函数表达式如下：

$$L_s=1-s=1-\frac{2\sum y_{\text{true}}y_{\text{pred}}}{\sum y_{\text{true}}+\sum y_{\text{pred}}} \tag{8}$$

其中，y_{true} 和 y_{pred} 分别表示分数图的真实值和预测值。为了使大文本区域和小文本区域生成精确的文本几何预测，保持回归损失尺度不变，旋转矩形框 RBox，回归部分采用 IoU 损失函数，因为它对不同尺度的对象是固定的，其表达式为：

$$L_R=-\log(IoU(\hat{R},R^*))=-\log\frac{|\hat{R}\cap R^*|}{|\hat{R}\cup R^*|} \tag{9}$$

其中，$\hat{R}$ 和 R^* 分别表示预测的几何形状和真实的几何形状，相交矩形 $|\hat{R}\cap R^*|$ 的宽度和高度分别为：

$$w_i=\min(\hat{d}_2,d_2^*)+\min(\hat{d}_4,d_4^*) \tag{10}$$

$$\boldsymbol{h}_i=\min(\hat{d}_1,d_1^*)+\min(\hat{d}_3,d_3^*) \tag{11}$$

其中，d_1，d_2，d_3，d_4 分别表示特征图中像素到对应矩形的上、右、下和左边界的距离。联合区可由式(12)求出：

$$|\hat{R}\cup R^*|=|\hat{R}|+|R^*|-|\hat{R}\cap R^*| \tag{12}$$

旋转角度损失的计算式如下：

$$L_\theta(\hat{\theta},\theta^*)=1-\cos(\hat{\theta}-\theta^*) \tag{13}$$

其中，$\hat{\theta}$ 表示对旋转角度的预测，θ^* 表示实际值。最后可以计算出总的几何损失为：

$$L_g=L_R+\lambda_\theta L_\theta \tag{14}$$

在实验过程中，将 λ_θ 设置为 10。

3.5 模型训练

TDMF 模型采用 Adam 优化器进行训练，为了加快学习速度，将训练样本中的原始图像修改为 512×512，并打包成 24 个批次进行训练。Adam 的初始学习率设置为 1×10^{-3}，当迭代次数过半时衰减为初始学习率的 1/10，直到指定训练次数结束。

4 实验结果与分析

4.1 实验数据

本文实验采用的数据集为 ICDAR 2013 和 ICDAR 2015 公用数据集，这两个数据集也是文本检测算法中比较常用的数据集。ICDAR 2013 数据集共有 462 张图片，其中训练集包含 229 张图片，测试集包含 233 张图片。ICDAR 2015 数据集共有 1 500 张图片，其中训练集包含 1 000 张图片，测试集包含 500 张图片。训练集的标签是图片中文字检测框的 4 个顶点，对应图片中的 4 个像素点。由于这些图片是在自然场景下随机拍摄的，因此图中的文本信息可能受到各种自然环境或者人为环境的影响，这些负面影响有利于评估文本检测算法的鲁棒性和准确率。本文在 ICDAR 两个数据集上的测试效果如图 4 所示，可以看出 TDMF 能有效应对复杂场景下图像中的文字检测问题。

图 4 TDMF 在 ICDAR 数据集上的检测效果

Fig. 4 TDMF detects the effect on ICDAR dataset

4.2 评价指标

本文使用准确率(Precision)、召回率(Recall)、F 值(F-measure)来评估算法的有效性，具体表达式如下：

$$Precision=\frac{TP}{TP+FP} \tag{15}$$

$$Recall=\frac{TP}{TP+FN} \tag{16}$$

$$F\text{-}measure=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{17}$$

其中，TP，FP，FN 分别表示正确预测的文本区域数、错误预

测的文本区域数和没有检测到的文本区域数。

4.3　对比实验

为了进一步验证 TDMF 的有效性，本节将其与其他自然场景文本检测方法在 ICDAR 2013 和 ICDAR 2015 两个数据集上进行了对比。ICDAR 2015 数据集下的对比实验结果如表 1 所列，TDMF 在准确率和 F 值指标方面相比其他方法取得了较好的效果，其中准确率为 0.8530，F 值为 0.8253，均高于表 1 中的其他方法，但是在召回率指标上比 RRPN 低了 4%。ICDAR 2013 数据集下的对比实验结果如表 2 所列，TDMF 相比其他几种方法，在 F 值的综合评价指标上具有较好的效果。

表 1　ICDAR 2015 数据集下的实验结果

Table 1　Results of testing on ICDAR 2015 dataset

Method	Recall	Precision	F-measure
Yao et al.[16]	0.5869	0.7226	0.6477
Tian et al.[5]	0.5156	0.7422	0.6085
Zhang et al.[17]	0.4309	0.7081	0.5358
Zheng et al.[18]	0.6200	0.4000	0.4800
SegLink[6]	0.7680	0.7310	0.7100
SSTD[19]	0.7300	0.8000	0.7700
RRPN[20]	**0.8400**	0.7700	0.8000
EAST[21]	0.7831	0.8224	0.8022
TDMF	0.7992	**0.8530**	**0.8253**

表 2　ICDAR 2013 数据集下的实验结果

Table 2　Results of testing on ICDAR 2013 dataset

Method	Recall	Precision	F-measure
Faster-RCNN[22]	0.7500	0.7100	0.7300
Text[23]	**0.8500**	0.6300	0.7200
Wang et al.[24]	0.6500	0.8200	0.7400
I2R NUS[25]	0.7000	0.6600	0.6900
Yin et al.[26]	0.6500	**0.8400**	0.7300
TDMF	0.6835	0.8320	**0.7505**

表 3 列出了本文提出的 TDMF 算法在去掉必要的特征融合和残差通道注意力网络处理模块后，在 ICDAR 2015 数据集上的表现效果，其中 TDMF-RCAN 表示原有模型在移除残差注意力网络后的模型，TDMF-MF 表示移除多粒度特征融合网络后，将提取出来的特征进行拼接并加入残差通道注意力的模型。从实验结果可以看出，残差通道注意力单元在模型训练中对特征的提取发挥了重要的作用，有效增强了模型对不同粒度特征的关注。多粒度特征融合网络则对提取出来的不同粒度特征进行了有效利用，在一定程度上提高了文本的检测精度和 F 值。

表 3　TDMF 消融实验

Table 3　TDMF ablation experiment

Method	Recall	Precision	F-measure
TDMF-RCAN	0.4786	0.3505	0.4047
TDMF-MF	0.5017	0.3539	0.4150
TDMF	**0.7992**	**0.8530**	**0.8253**

结束语　本文提出了一种结合多粒度特征融合的自然场景文本检测方法(TDMF)，该方法利用深度卷积神经网络提取特征，并将提取出的多个粒度级别的信息加入残差通道注意力模型进行融合，然后生成得分图以及文字的检测边界框。在 ICDAR 2015 数据集和 ICDAR 2013 数据集上与其他一些方法的对比实验中，本文方法在准确率、召回率、F 值 3 个指标上取得了更具有竞争力的效果。尽管 TDMF 在实验中表现出了不错的性能，但仍有不足之处，在下一步的工作中，将针对不规则文本和多语种同时存在的文本检测进行更深入的研究。

参 考 文 献

[1] CHO H, SUNG M, JUN B. Canny Text Detector: Fast and Robust Scene Text Localization Algorithm[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Press, 2016: 3566-3573.

[2] NEUMANN L, MATAS J. A method for text localization and recognition in real-world images[C]// Asian Conference on Computer Vision. Berlin: Springer Press, 2010: 770-783.

[3] TIAN S X, PAN Y F, HUANG C, et al. Text flow: A unified text detection system in natural scene images[C]// Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE Press, 2015: 4651-4659.

[4] WANG K, BELONGIE S. Word spotting in the wild[C]// European Conference on Computer Vision. Berlin: Springer Press, 2010: 591-604.

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

[6] SHI B G, 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. Hawaii: IEEE Press, 2017: 2550-2558.

[7] XU H L, SU F. A robust hierarchical detection method for scene text based on convolutional neural networks[C]// Proceedings of the 2015 IEEE International Conference on Multi-media and Expo. Turin: IEEE Press, 2015: 1-6.

[8] WANG Y X, XIE H T, ZHA Z J. ContourNet: Taking a Further Step Toward Accurate Arbitrary-shaped Scene Text Detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE Press, 2020: 11753-11762.

[9] YANG X, HE D F, ZHOU Z H, et al. Learning to Read Irregular Text with Attention Mechanisms[C]// International Joint Conference on Artificial Intelligence Pacific Rim International Conference on Artificial Intelligence. Melbourne: Morgan Kaufmann Press, 2017: 3.

[10] WANG W H, XIE E Z, 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. California: IEEE Press, 2019: 9336-9345.

[11] 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. California: IEEE Press, 2019: 9365-9374.

[12] CHEN L. Topological structure in visual perception [J]. Science, 1982, 218(4573): 699-700.
[13] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015) [2020-10-23]. https://arxiv.org/pdf/1409.1556.pdf.
[14] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Press, 2016: 770-778.
[15] ZHANG Y L, LI K P, LI K, et al. Image superresolution using very deep residual channel attention networks[C]// Proceedings of the European Conference on Computer Vision. Munich: Springer Press, 2018: 286-301.
[16] YAO C, BAI X, SANG N, et al. Scene text detection via holistic, multi-channel prediction [EB/OL]. (2016) [2020-10-23]. https://arxiv.org/pdf/1606.09002.pdf.
[17] ZHANG Z, ZHANG C Q, SHEN W, et al. Multi-oriented text detection with fully convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Press, 2016: 4159-4167.
[18] ZHENG Y, LI Q, LIU J, et al. A cascaded method for text detection in natural scene images[J]. Neurocomputing, 2017, 238: 307-315.
[19] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]// European Conference on Computer Vision. Cham: Springer Press, 2016: 21-37.
[20] MA J Q, SHAO W Y, YE H, et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia, 2018, 20(11): 3111-3122.
[21] ZHOU X Y, 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. Hawaii: IEEE Press, 2017: 5551-5560.
[22] 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 and Machine Intelligence Press, 2016, 39(6): 1137-1149.
[23] SHI C Z, WANG C H, XIAO B H, et al. Scene text detection using graph model built upon maxially stable extremal regions [J]. Pattern Recognition Letters, 2013, 34(2): 107-116.
[24] WANG X B, SONG Y H, ZHANG Y L, et al. Natural scene text detection with multi-layer segmentation and higher order conditional random field based analysis[J]. Pattern Recognition Letters, 2015, 60: 41-47.
[25] JADERBERG M, VEDALDI A, ZISSERMAN A. Deep features or text spotting[C]// European Conference on Computer Vision. Zurich: Springer Press, 2014: 512-528.
[26] YIN X C, PEI W Y, ZHANG J, et al. Multi-orientation scene text detection with adaptive clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence Press, 2015, 37(9): 1930-1937.

CHEN Zhuo, born in 1993, master. His main research interests include computer vision and so on.

LIU Qun, born in 1969, Ph. D, professor, is a member of China Computer Federation. Her main research interests include data mining, complex network and so on.