计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800157

基于文本三区域分割的场景文本检测方法

李 煌 王晓莉 项欣光
南京理工大学计算机科学与工程学院 南京 210094
(lihuangleo@njust.edu.cn)

摘 要 随着卷积神经网络的发展,场景文本检测也得到了快速发展。然而,场景文本检测仍然存在很多问题:一方面,许多检测方法都采用矩形框作为检测框,这对于图像中不规则的文本是不友好的;另一方面,部分方法获取的检测框无法分离相邻的文本实例,从而导致图像中相邻文本的误检测。为了解决这两个问题,文中提出了一种基于文本三区域分割的场景文本检测方法,将图像的文本实例分别映射到整体区域、核心区域和边框区域空间中,以获取图像的文本实例在上述3个区域的分割图,然后利用整体区域分割图和边框区域分割图来指导核心区域分割图的生成。文本的核心区域虽包含了图像中的文本位置、大小等信息,但是缺少边界信息。为了获取更加精确的检测结果,所提方法利用文本的边框区域来对核心区域进行监督学习。最后将基于文本的核心区域分割图像,产生契合文本核心的外接多边形,并进行一定比例的扩张,获取检测结果。实验结果表明,所提方法在ICDAR2015数据集上的准确率可达到83%,与现有的检测算法相比,其F值获得了1%以上的提升,而且该算法在弯曲文本的检测上亦有着优异的表现。

关键词:场景文本检测;神经网络;实例分割;深度学习;计算机视觉
中图法分类号 TP391

Scene Text Detection Based on Triple Segmentation

LI Huang, WANG Xiao-li and XIANG Xin-guang
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Abstract Scene text detection has been developed rapidly with the development of convolutional neural network. However, there still exists some challenges. On the one hand, many detection algorithms use rectangular box as the detection box, which is inaccurate to locate the irregular texts. On the other hand, some methods may get the bounding boxes but fail to separate text instances that lie very close to each other, causing error detection. To solve these two problems, this paper proposes a novel triple segmentation (TS), text instances in image are mapped to score area, kernel area and threshold area, which generate three segmentation maps, the score map and threshold map are used to guide the generation of kernel map. Although kernel map has the information of texts in image, such as location, size and so on, it lacks the threshold information. In order to get a better result, this method uses threshold map to restrict the generation of kernel map. The detection result is based on instance segmentation to get the bounding polygon of text kernel instance, and then make an expansion. This algorithm achieves a precision of 83% on ICDAR2015 dataset, which outperforms the existing methods by more than 1% on F-measure, which proves this method is also effective to detect curve texts.

Keywords Scene text detection, Neural networks, Instance segmentation, Deep learning, Computer vision

1 引言

近年来,图像的文本检测工作是计算机视觉领域备受关注的问题,原因是其在图像文字翻译、自动驾驶、文本分类等应用场景均发挥着重要的作用。作为图像文本识别的重要前置工作,文本检测同样也是图像语义理解问题中的重要一环。因此,如何准确、快速地实现图像文本检测成为了当前图像领域的一个重要课题。

近年来,随着卷积神经网络(Convolutional Neural Networks,CNN)的飞速发展,各类基于CNN的方法在目标检测、图像分类、图像增强等领域均取得了十分优异的成果[1-4]。在图像文本检测的研究工作中,CNN也推动了研究的进展。

基于CNN的文本检测方法主要包括两类:一类基于候选框的文本检测方式,这类方法主要受到目标检测如Faster R-CNN[5],SSD[6],YOLO[7]等架构的影响。首先产生部分候选框,然后对候选框进行筛选并回归生成文本框坐标。这类多阶段生成回归坐标的检测方法在准确性上表现较好,但较为耗时。另一类则受到实例分割的启发,将图像划分为文本

到稿日期:2020-08-24 返修日期:2020-09-16 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。
通信作者:项欣光(xgxiang@njust.edu.cn)

所属区域与背景两大类，通过分割网络对图像进行像素级别的语义分割，再基于分割结果进行文本行的坐标构建。这种基于实例分割的文本检测方式简化了网络架构，更为高效，但在处理长文本检测问题上精度会有所下降。

在现实场景中，文本的形状是未知的，而传统检测方法往往无法检测出变形的文本，因此基于实例分割的文本检测方法成为极具潜力的一个研究方向。场景中通常会出现文本间紧密相连的情况，如图 1 所示。

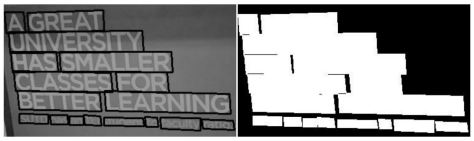


图 1　文本紧密相连

Fig. 1　Texts close to each other

之前的研究工作虽然提升了文本检测的效果，但是在处理紧密文本时，产生的标注框会使不同文本的检测框相交并产生边框粘连的问题。本文将用传统方法获取的标注框区域称为文本整体区域，同时将包含文本位置、形状等重要信息的区域称为文本核心区域。由于核心区域中不包含文本的边界信息，因此将核心区域与整体区域之间的区域称为文本边框区域。这 3 类区域包含了文本的所有信息。

针对上述问题，本文提出了一种基于三区域实例分割的文本检测方法(Triple Segmentation，TS)。如图 2 所示，本方法利用 3 个区域分割图像来确定更加精准的核心区域表示，最后将核心区域进行一定比例的扩张，得到最终的文本检测框。首先，利用 ResNet[8]＋FPN[9]作为模型的主干网络来提取图像的特征表示，并将 ResNet 中的卷积层替换成可变形卷积[10]，以解决因文本形状的未知性而带来的特征表示不全的问题。接下来将获取的图像特征分别映射到 3 个区域的表示空间中，得到每个区域下的文本信息特征表示。然后利用整体特征和核心特征来获取精确的核心区域表示，再根据核心区域的连通区域得到初步的检测框。最后按照指定比例扩张，从而获取最终的文本检测框。本方法通过三区域的表征，可以获取到更加精准的文本检测信息。在公开数据集上的实验表明，本文提出的检测方法优于现有的方法。

(a)原图　(b)整体区域　(c)核心区域　(d)边框区域

图 2　文本三区域

Fig. 2　Text triple segmentation

本文的主要贡献在于：

(1)本文设计了三区域特征表示方法，通过整体、核心和边框 3 个方面来更加全面地表征图像中的文本信息，同时三区域的表征方式可以有效解决文本粘连的问题。

(2)本文设计了 Minus 模块，利用边框区域信息来监督核心区域的预测，以去除核心区域中的边缘部分与文本边框信息产生的交叉噪声，使文本核心区域的预测更加精准。

(3)为了适用于多方向文本及弯曲文本的检测，本文利用实例分割对图像中的文本进行定位，以产生更贴合文本实例的多边形标注框，实验结果证明了 Triple Segmentation-Combine(TS-Combine)在检测弯曲文本时的有效性及优越性。

2　相关工作

2.1　传统方法

一般来说，文字的边缘密度较大，因此与背景的对比度也较高，依据图像的边缘特征可以将文本与背景区域进行区分。这类方法[11-13]一般采用人工设置的滤波算子对图像边缘信息进行提取，然后根据形态学特征剔除伪文本区域。

基于边缘特征的文本检测在文档图像的检测工作中取得了较好的效果，是过去常用的检测方案。但在自然场景的图像中，由于干扰因素增多，致使边缘特征的提取无法有效区分文本与非文本区域，因此上述检测算法在处理场景文本检测问题上存在一定的缺陷。

2.2　基于卷积神经网络的方法

基于卷积神经网络的文本检测方法大致可分为两类：基于坐标回归的文本检测算法与基于分割的文本检测算法。

基于坐标回归的文本检测算法即在网络的输出层直接回归图像文本实例的框选坐标。TextBoxes[14]就是一种基于 SSD 的回归文本坐标的检测算法，它调整了网络中候选框的比例，使其适用于文本行的检测工作。EAST[15]借鉴了全卷积网络[16]的思想，直接回归产生目标图像及其文本的旋转角度和坐标。类似的工作[17-19]能有效地作用于直线文本的检测上，但当文本的形状发生弯折时，这些方法的回归坐标将会产生偏差。

基于分割的文本检测算法主要源于全卷积网络，Zhang 等[20]提出的算法首次将全卷积网络引入文本检测的任务中。随后有检测方法[21]延用了其中的思想，但在分割产生的对象上做了各种改进工作。PixelLink[22]在进行文本实例分割的基础上，同时进行文本内字符关联的预测，以此进行文本的归类工作。最近提出的 PSENet[23]对文本进行了多尺度分割，分离出多个文本核心，并对核心进行阶段性的相融操作。

3　文本三区域分割的场景文本检测方法(TS)

真实场景中的很多文本距离很近，容易产生部分重叠的检测框，这限制了文本检测算法在实际场景中的应用。本文设计了一种基于三区域分割的文本检测方法，模型的具体流程如图 3 所示。首先利用主干网络提取图像的特征图，通过 FPN 及特征融合处理，获取融合特征图 G，随后将特征图 G 分别映射到 3 个区域对应的空间中，来初步获取文本在不同区域空间下的特征表示。3 个区域分别是文本的整体区域、核心区域和边框区域，通过整体区域特征图 S 与核心区域特征图 K 的融合来确定文本的核心区域预测图 CB。最后根据预测出的核心区域来获取文本定位坐标，从而得到最终的检测结果。为了使模型能够更好地控制核心区域的生成范围，引入了边框区域特征图 T 来指导模型获取到更加精准的核心区域信息。

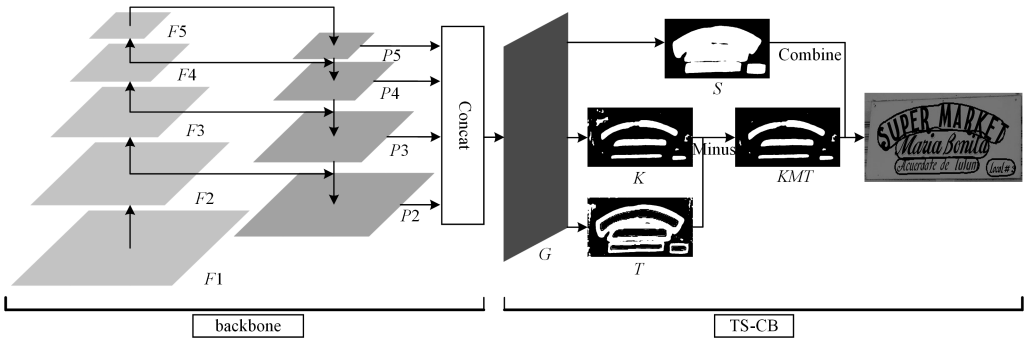


图 3 网络结构图

Fig. 3 Diagram of network structure

3.1 三类特征图像的标注生成

由于本文提出了基于三区域分割的文本检测算法，因此模型需要获取这 3 个区域的标注信息。公开的数据集中并不直接包含这些信息，因此需要由最初的文本定位坐标生成三区域下的图像标注信息来指导模型的学习。

如图 4 所示，图中最外缘的标注框即为初始的文本框，由此可生成文本的整体区域分布图。图像的文本核心区域并没有原始的坐标标注，为获取这个标注，本文采用了 Vatti clipping 算法[24]，使得图像中的标注多边形的各边均向内部收缩 d_i 个像素，以获取该多边形的核心区域框。收缩距离 d_i 的计算式如下：

$$d_i = A_i(1-s^2)/L_i \tag{1}$$

其中，A_i 为多边形面积；L_i 为周长；s 为收缩比例，这里 s 设为 0.5。

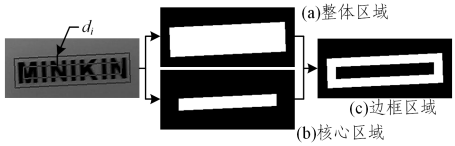


图 4 文本标注流程图

Fig. 4 Flow chart of text label

在获取上述两类图像后，将图像中文本框与核心区域框之间形成的连通区域定义为文本的边框区域。

3.2 特征提取模块

CNN 作为常用的特征提取方法，在不同的任务中均取得了很大的成就，但是对未知形状变换的建模存在固有的缺陷。在真实场景中，文本的形状存在极大的未知性，因此本文模型中引入了可变形卷积的方法来获取图像的特征。具体地，模型选用 ResNet 作为主干网络，同时将 ResNet 中第二到第五阶段中的卷积操作替换成可变形卷积，具体过程如图 3 所示。传统的卷积操作是在输入特征映射时使用规则网格进行采样，而可变形卷积在正常的采样坐标上增加了一个位移量，从而可以根据当前需要识别的图像内容进行动态调整。例如，在第一阶段得到的特征图 $F1$ 输入到第二阶段的过程中，设定一个基础的规则网格 R，这个网格体现出卷积块的感受野。定义一个 3×3 的卷积基础网格 R 为：

$$R=\{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\} \tag{2}$$

在此基础上，引入一个偏移量集合 Δr_i，利用可变形卷积来提取图像 X 中 k_0 点的特征 $y(k_0)$，计算操作如下：

$$y(k_0)=\sum_{r_i\in R} w(r_i)\cdot X(k_0+r_i+\Delta r_i) \tag{3}$$

其中，$w(r_i)$ 为卷积网格中 r_i 处对应的权重值，Δr_i 即卷积网格中相应点的偏移量。二者都将在网络的优化过程中不断地迭代更新，最终产生符合文本形态的变形卷积，从而获取更加贴合任务的特征。

在真实场景中文本的大小是不确定的，而在 ResNet 网络中，随着层数的增加，图像中的部分细节信息会逐渐丢失。为了避免图像中小目标文本信息的丢失，同时为了更加精确地捕获图像中所有文本的信息，模型引入了 FPN 结构，其过程就是将 ResNet 中每个阶段产生的特征进行融合。$P2$，$P3$，$P4$，$P5$ 分别是主干网络在不同阶段产生的特征图，其分别由 ResNet 的各阶段特征结果 $F2$，$F3$，$F4$，$F5$ 经过 1×1 卷积层获得，特别地，所有 P 类图像通道数将被统一为 256。

由于 $P2$ 是第二阶段产生的特征图，其分辨率是最大的，因此将 $P3$，$P4$，$P5$ 分别上采样到与 $P2$ 特征图相同的尺寸，然后将 4 个特征进行拼接，最后利用一个常规的卷积层将这几个特征融合起来，从而得到最终的特征图表示。其具体过程可以用下式来表示：

$$G=Conv(Concat(P2,upsample_{\times 2}(P3),upsample_{\times 4}(P4),upsample_{\times 8}(P4))) \tag{4}$$

其中，$Concat$ 表示通道方向上的拼接操作，$upsample$ 为上采样操作，下标的数字表示采样倍数。最终获取到包含丰富语义信息的特征图 G。

3.3 TS-Combine 架构(TS-CB)

TS 三区域分割将会产生与文本实例相关的 3 类分割图像，这 3 类分割图像将从特征图 G 中获取。结合了各阶段提取信息的特征图 G 包含了图像中的全部信息。在文本检测任务中，不需要关注图像的全部信息。为了将关注点放在任务目标上，模型设计了 3 个区域空间从不同的方面来获取图像中的文本信息。利用 3 个 1×1 卷积层分别获取图像中文本的整体区域特征图 S、核心区域特征图 K 和边框区域特征图 T。

TS-CB 模块仅利用 3 个独立卷积层将图像特征 G 映射到 3 个区域中，因此，从 3 个区域 S，K，T 及差异区域 KMT 中获取的特征将含有一些噪声信息，不利于获取精确的文本特征信息。因此在 TS-CB 结构中，首先提出了一种 Minus 方法，利用边框区域特征图 T 对核心区域特征图 S 进行约束，产生差异区域图 KMT，使文本的核心区域 K 在边缘分布预

测上的准确性得到提升。具体可用下式表示：

$$KMT_{(i,j)}=softmax(K_{(i,j)}-T_{(i,j)}) \tag{5}$$

其中，(i,j)为图像中各个像素点的坐标。

最后通过将整体区域特征 S 与差异区域 KMT 相融合，来降低噪声信息对文本识别的影响，同时获取到包含文本定位信息的图像 CB。图像 CB 中各点的取值如下：

$$CB_{(i,j)}=\begin{cases}1, & \text{if } (MS_{(i,j)}=1 \text{ and } MK_{(i,j)}=1\\ 0, & \text{if } (MS_{(i,j)}=0 \text{ or } MK_{(i,j)}=0\end{cases} \tag{6}$$

$$MS_{(i,j)}=\begin{cases}1, & \text{if } S_{(i,j)}>thre\\ 0, & \text{otherwise}\end{cases} \tag{7}$$

其中，(i,j)为图像中各个像素点的坐标，$thre$ 设为 0.5，$\boldsymbol{MS}$ 与 $\boldsymbol{MK}$ 分别为图 S 与 KMT 产生的二值掩码矩阵，以 $\boldsymbol{MS}$ 为例，其产生过程如式(7)所示。

CB 为最终的核心分布预测图像，在该图像中每一个连通域即为一个单独的文本实例核心，因此每个连通域的外接多边形将定位一个单独的文本核心。每个文本核心的预测定位多边形 O_i 并不包含文本边框的边界信息，为了获取精确的文本实例框，需要对 O_i 进行扩张操作，扩张操作同样采用 Vatti clipping 算法，对 O_i 的各边向外扩张 d_i' 个像素。扩张距离 d_i' 的计算式如下：

$$d_i'=A_i'(1+s')/L_i' \tag{8}$$

其中，A_i' 为多边形面积；L_i' 为周长；s' 为扩张比例，这里 s' 设为 0.5。

O_i 扩张 d_i' 像素产生的边界框即为文本的定位信息框，由此 TS-CB 预测产生了图像中所有文本的定位坐标信息。

在最终的结果中，由于核心区域特征图 K 的准确回归，差异区域图 KMT 的特征分布将与其基本一致。在预测中，为提升效率，可以利用图 K 来代替图 KMT。因此在进行预测时，仅将生成的三区域分割图像作为后续 TS-CB 架构的输入，来为得到最终的文本坐标定位信息提供依据。

3.4　损失函数

为了训练模型，将损失函数设置如下：

$$L=\alpha L_s+\alpha L_{kmt}+\beta L_k+\beta L_t \tag{9}$$

其中，L_s 为文本整体区域的损失函数，L_{kmt} 为差异区域的损失函数，L_k 为文本核心区域的损失函数，L_t 为文本边框区域的损失函数。α 与 β 为各类损失的权重，分别设置为 0.4 和 0.1。

因最终的检测以文本的整体区域 S 及文本的差异区域 KMT 为主要基准，并且以二者的交集作为检测核心，因此二者的重要性是一致的，故选取相同的权重；考虑到文本核心区域 K 与边框区域 T 将共同指导产生 KMT，这里对二者的损失选取相同的权值。

一般的分割任务常采用交叉熵作为损失函数，但其在处理正负例样本失衡的图像数据时，会导致模型偏向于图像的背景区域，造成正例样本的漏检。对于场景文本检测任务而言，文本区域与非文本区域(背景)是不均衡的关系，所以本文并没有使用交叉熵而是采用 Dice Loss[25] 损失函数。Dice Loss 中的 dice coefficeient 作为评价两个图像的相似度指标，在处理类别不均衡问题上效果更为优秀，更加适用于场景文本任务。

以文本区域 S 的分割为例，其真实的分割图像为 GS，二者的相似度由 dice coefficeient 计算获取：

$$D(S,GS)=\frac{2\sum_{(i,j)}(S_{(i,j)}\times GS_{(i,j)})}{\sum_{(i,j)}S_{(i,j)}^2+\sum_{(i,j)}GS_{(i,j)}^2} \tag{10}$$

其中，(i,j)为图像中各个像素点的坐标。

由于场景图像中的文本区域所占的比例往往低于非文本区域，且存在围栏、篱笆等容易被识别为文本实例的负例样本。因此在文本整体区域特征图 S 的预测中，本文引入了 OHEM[26] 来挖掘样本以产生掩码 M。文本区域的损失 L_s 可通过下式来表示：

$$L_s=1-D(S\cdot M,GS\cdot M) \tag{11}$$

特别地，文本的核心区域 GK 与文本的边框区域 GT 均被文本整体区域 GS 完全涵盖，且在进行文本核心区域及文本边框区域的预测时，应更关注图像的文本实例区域。在计算文本核心区域与文本边框区域的损失函数时，本文忽略了文本整体区域特征图 S 中的非文本区域，以消除冗余。因此在计算 L_k 与 L_t 时，使用 $\boldsymbol{MS}$ 作为前置掩码，$\boldsymbol{MS}$ 的产生过程如式(7)所示，即 L_k 与 L_t 有：

$$L_k=1-D(K\cdot \boldsymbol{MS},GK\cdot \boldsymbol{MS}) \tag{12}$$

$$L_t=1-D(T\cdot \boldsymbol{MS},GT\cdot \boldsymbol{MS}) \tag{13}$$

由于差异区域的最终目标是与文本核心区域保持基本一致，所以其损失函数 L_{kmt} 与 L_k 类似：

$$L_{kmt}=1-D(KMT\cdot \boldsymbol{MS},GK\cdot \boldsymbol{MS}) \tag{14}$$

4　实验结果

4.1　数据集

为了验证所提算法的有效性，本文在两个极具代表性的公开数据集(CTW1500 数据集和 ICDAR2015 数据集)上进行验证实验。本文的所有实验仅使用原始数据集，并未引入外部数据集进行预训练。

CTW1500 数据集[27] 是一个针对弯曲文本检测的数据集，由华南理工大学金连文团队构建，其中包含 1 000 张训练集和 500 张测试集，每张图片至少包含一个弯曲文本行。文本的标注是以文本行来对每个文本实例进行标注，因此该数据集的长文本检测也是一项富有挑战性的检测任务。

ICDAR2015 数据集[28] 是当前文本检测算法中常用的数据集，共包含 1 500 张图片，其中 1 000 张用于模型训练，其余图片用于测试，所有的图像尺寸均为 720×1 280。这些图像均是在各类场景中随机拍摄的图片，文本的标注均以单词进行标注区分，即对每个单词进行框选标注。

4.2　实验细节

本文实验使用了 pytorch 进行模型的框架构建，主干网络为 ResNet。在网络模型参数的优化上，使用了 Adaptive Moment Estimation(Adam)进行参数更新。初始学习率设置为 1×10^{-4}，在所有的数据训练完成 200 次迭代后，学习率将下降至 1/10，最终学习率停留在 1×10^{-6}。

为使模型具有更强的泛化能力，在训练中对数据进行数据增强处理，具体采用如下策略：

(1)对图片进行随机旋转，旋转角度范围为(−10°，10°)。

(2)对图片进行随机缩放，缩放比例限定于[0.5，1.0，2.0，3.0]。

(3)对图片进行随机裁剪，仅选取图片中的部分区域进行训练，裁剪尺寸为 640×640。

4.3 消融实验

本文在 ICDAR2015 数据集上对 TS 模型进行了多个消融实验，来验证本文提出的三区域实例分割及其 Minus 方法的有效性。如表 1 所列，K2 表示去除文本边框区域，仅利用二区域实例分割进行效果预测。类似地，K3 表示三区域实例分割，但未引入 Minus 方法的测试结果，相比 K2，其效果的提升并不明显。不难发现，相比两种方法，本文提出的 TS 方法的 F-值分别取得了 2.5%（相比 K2）与 1.7%（相比 K3）的提升。

相比上述 K2 与 K3，TS 使用了额外的标注信息及卷积操作，因此计算开销有一定的增加，但在网络的整体结构中，主要的计算开销存在于主干网络 ResNet50，其参数量为 45 M，而后续的单区域处理参数量为 100 K，则三区域参数量为 300 K，计算开销仅有 0.44%（200 K/45 M）的提升，但效果却取得了明显的提升。

表 1 ICDAR2015 上的消融实验

Table 1 Ablation experiment on ICDAR2015

Method	*Precision*	*Recall*	*F*
K2-ResNet50	79.8	79.4	79.6
K3-ResNet50	80.1	80.6	80.3
Ours	**83.3**	**81.0**	**82.1**

为分析三区域的各类损失参数对结果的影响，本文同时分析了不同损失权重下的检测效果。具体操作为：α 的取值由 0.25 增至 0.45，相应地，β 的取值为 $(1-2\alpha)/2$。对应的结果如表 2 所列，可以看出，当 α 的取值为 0.40，β 的取值为 0.10 时，F 值将达到最高，为 82.1%。

表 2 ICDAR2015 上的参数分析

Table 2 Parameter analysis on ICDAR2015

α	β	*Precision*	*Recall*	*F*
0.25	0.25	82.0	**81.9**	81.9
0.30	0.20	81.7	81.5	81.6
0.35	0.15	82.5	81.3	81.9
0.40	0.10	**83.3**	81.0	**82.1**
0.45	0.05	82.2	81.5	81.8

4.4 对比实验

为验证本文算法在多方向文本检测方面的优越性，将其与其他算法进行了对比，对比方法包括 CTPN，SegLink[29]，EAST，Corner[30]，DeepReg[31]，PSENet 和 PAN[32]。本文算法在 3 个指标上均取得了不错的表现，尤其是与 DeepReg 相比，召回率提升了 1%，最终 F 值实现了 1.1% 的增长。本文算法在 ICDAR2015 数据集上的测试结果如表 3 所列。

表 3 ICDAR2015 检测结果

Table 3 Detection results on ICDAR2015

Algorithm	*Precision*	*Recall*	*F*
CTPN[17]	74.2	51.6	60.9
SegLink[29]	73.1	76.8	75.0
EAST[15]	83.57	73.47	78.2
Corner[30]	**94.1**	70.7	80.7
DeepReg[31]	82.0	80.0	81.0
PSENet[23]	81.5	79.7	80.6
PAN[32]	82.9	77.8	80.3
Ours	83.3	**81.0**	**82.1**

在弯曲文本的检测实验中，本文主要在 CTW1500 数据集上进行了测试，结果如表 4 所列，该结果验证了本文算法在弯曲文本检测上的有效性。与专注于弯曲文本检测的方法 CTD+TLOC[27]，Textsnake[33] 以及 LOMO[34] 相比，本文算法的各项指标都有所提升。相比 Textsnake，其准确率提升了 10.3%，最终 F 值也实现了 2.4% 的提升。

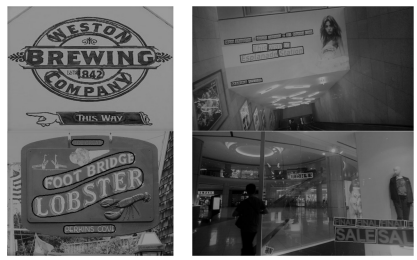
表 4 CTW1500 检测结果

Table 4 Detection results on CTW1500

Algorithm	*Precision*	*Recall*	*F*
CTPN*	60.4	53.8	56.9
SegLink*	42.3	40.0	40.8
EAST*	78.7	49.1	60.4
CTD+TLOC[27]	77.4	69.8	73.4
TextSnake[33]	67.9	85.3	75.6
PSENet[23]	**80.6**	75.6	78.0
LOMO[34]	69.6	**89.2**	**78.4**
Ours	78.2	77.8	**78.0**

注：* 表示数据来自文献[27]；LOMO 使用了外部数据进行预训练

本文算法对自然场景下文本的检测结果如图 5 所示。结果表明，TS 算法对 ICDAR2015 中的相邻文本具有很好的分离效果，准确地定位了每个单词级的文本实例，没有出现文本粘连的现象。在 CTW2015 上的检测图像中，TS 算法对弯曲文本的定位也很准确，没有出现冗余框选定位。



(a)CTW1500 (b)ICDAR2015

图 5 检测结果

Fig. 5 Detection results

结束语 本文提出了一种基于文本三区域分割的场景文本检测方法 TS。TS 将图像中的文本划分为 3 个区域进行实例分割，结合文本的整体区域与核心区域来分离场景图像中的粘连文本；另一方面，在本文的 Minus 模块中，文本的边框区域对核心区域的预测起到了更好的监督效果，产生了更为准确的核心分割图像。在生成的三区域分割图像上进行文本实例的定位，可以产生更贴合文本形态的文本框。实验结果表明，TS 方法适用于场景图像中各类形状的文本检测。本文方法在文本检测方面取得了不错的成果，但在文本实例的识别对接上，尚未有更好的端到端识别方案，这将是我们下一步的研究方向。

参 考 文 献

[1] LI Z C, TANG J H, ZHANG L Y, et al. Weakly-supervised Semantic Guided Hashing for Social Image Retrieval[J]. International Journal of Computer Vision, 2020, 128(8): 2265-2278.

[2] PENG Z, LI Z, ZHANG J, et al. Few-Shot Image Recognition With Knowledge Transfer[C]// International Conference on Computer Vision. 2019: 441-449.

[3] LI Z, TANG J, MEI T, et al. Deep Collaborative Embedding for

Social Image Understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,41(9):2070-2083.
[4] ZHOU H,LI Z,NING C,et al. CAD:Scale Invariant Framework for Real-Time Object Detection[C]// International Conference on Computer Vision. 2017:760-768.
[5] REN S,HE K,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,2017,39(6):1137-1149.
[6] LIU W,ANGUELOV D,ERHAN D,et al. SSD:Single Shot MultiBox Detector[C]// European Conference on Computer Vision. 2016:21-37.
[7] REDMON J,DIVVALA S K,GIRSHICK R,et al. You Only Look Once:Unified,Real-Time Object Detection[C]// Computer Vision and Pattern Recognition. 2016:779-788.
[8] HE K,ZHANG X,REN S,et al. Deep Residual Learning for Image Recognition[C]// Computer Vision and Pattern Recognition. 2016:770-778.
[9] LIN T,DOLLAR P,GIRSHICK R,et al. Feature Pyramid Networks for Object Detection[C]// Computer Vision and Pattern Recognition. 2017:936-944.
[10] DAI J,QI H,XIONG Y,et al. Deformable Convolutional Networks[C]// International Conference on Computer Vision. 2017:764-773.
[11] JAMIL A,SIDDIQI I,ARIF F,et al. Edge-Based Features for Localization of Artificial Urdu Text in Video Images[C]// 2011 International Conference on Document Analysis and Recognition (ICDAR). IEEE,2011.
[12] SHI C,WANG C,XIAO B,et al. Scene text detection using graph model built upon maximally stable extremal regions[J]. Pattern Recognition Letters,2013,34(2):107-116.
[13] EPSHTEIN B,OFEK E,WEXLER Y,et al. Detecting text in natural scenes with stroke width transform[C]// Computer Vision and Pattern Recognition. 2010:2963-2970.
[14] LIAO M,SHI B,BAI X,et al. TextBoxes:a fast text detector with a single deep neural network[C]// National Conference on Artificial Intelligence. 2017:4161-4167.
[15] ZHOU X,YAO C,WEN H,et al. EAST:An Efficient and Accurate Scene Text Detector[C]// Computer Vision and Pattern Recognition. 2017:2642-2651.
[16] LONG J,SHELHAMER E,DARRELL T,et al. Fully convolutional networks for semantic segmentation[C]// Computer Vision and Pattern Recognition. 2015:3431-3440.
[17] TIAN Z,HUANG W,HE T,et al. Detecting Text in Natural Image with Connectionist Text Proposal Network[C]// European Conference on Computer Vision. 2016:56-72.
[18] JIANG Y,ZHU X,WANG X,et al. R2CNN:Rotational Region CNN for Orientation Robust Scene Text Detection[J]. arXiv:1706.09579,2017.
[19] HE P,HUANG W,HE T,et al. Single Shot Text Detector with Regional Attention[C]// International Conference on Computer Vision. 2017:3066-3074.
[20] ZHANG Z,ZHANG C,SHEN W,et al. Multi-oriented Text Detection with Fully Convolutional Networks[C]// Computer Vision and Pattern Recognition. 2016:4159-4167.
[21] YAO C,BAI X,SANG N,et al. Scene Text Detection via Holistic,Multi-Channel Prediction[J]. arXiv:1606.09002,2016.
[22] DENG D,LIU H,CAI D,et al. PixelLink:Detecting Scene Text via Instance Segmentation[C]// National Conference on Artificial Intelligence. 2018:6773-6780.
[23] WANG W,XIE E,LI X,et al. Shape Robust Text Detection With Progressive Scale Expansion Network[C]// Computer Vision and Pattern Recognition. 2019:9336-9345.
[24] VATTI B R. A generic solution to polygon clipping[J]. Communications of The ACM,1992,35(7):56-63.
[25] MILLETARI F,NAVAB N,AHMADI S,et al. V-Net:Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation[C]// International Conference on 3D Vision. 2016:565-571.
[26] SHRIVASTAVA A,GUPTA A,Girshick R,et al. Training Region-Based Object Detectors with Online Hard Example Mining [C]// Computer Vision and Pattern Recognition. 2016:761-769.
[27] LIU Y,JIN L,ZHANG S,et al. Detecting Curve Text in the Wild:New Dataset and New Solution[J]. arXiv:1712.02170,2017.
[28] KARATZAS D,GOMEZBIGORDA L,NICOLAOU A,et al. ICDAR 2015 competition on Robust Reading[C]// International Conference on Document Analysis and Recognition. 2015:1156-1160.
[29] SHI B,BAI X,BELONGIE S,et al. Detecting Oriented Text in Natural Images by Linking Segments[C]// Computer Vision and Pattern Recognition. 2017:3482-3490.
[30] LYU P,YAO C,WU W,et al. Multi-oriented Scene Text Detection via Corner Localization and Region Segmentation[C]// Computer Vision and Pattern Recognition. 2018:7553-7563.
[31] HE W,ZHANG X,YIN F,et al. Deep Direct Regression for Multi-oriented Scene Text Detection[C]// International Conference on Computer Vision. 2017:745-753.
[32] WANG W,XIE E,SONG X,et al. Efficient and Accurate Arbitrary-Shaped Text Detection With Pixel Aggregation Network [C]// International Conference on Computer Vision. 2019:8440-8449.
[33] LONG S,RUAN J,ZHANG W,et al. TextSnake:A Flexible Representation for Detecting Text of Arbitrary Shapes[C]// European Conference on Computer Vision. 2018:19-35.
[34] ZHANG C,LIANG B,HUANG Z,et al. Look More Than Once:An Accurate Detector for Text of Arbitrary Shapes[C]// Computer Vision and Pattern Recognition. 2019:10552-10561.

LI Huang, born in 1992, master. His main research interests include text detection and so on.

XIANG Xin-guang, born in 1982, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include multimedia analysis, computer vision and so on.