CVPR 2019 中的弯曲文本检测技术介绍

陈毅
3150102334
计算机科学与技术学院

摘要

目前的自然场景文本检测技术逐渐成熟，逐渐解决了不同语言、不同方向以及不同形状的文本检测问题，在弯曲文本的检测上仍然有许多新的方案提出。文本先对已有的文本检测技术作了综述，然后介绍了在 2019 年CVPR会议上提出的 4 篇支持弯曲文本检测的文章CRAFT、LOMO、PSENet以及CSE，并重点分析了CRAFT这一方法，然后通过对各个方法的介绍以及实验分析来进行直观的对比，对未来的文本检测方案改进与创新有更好的指导作用。

1. 自然场景文本检测技术综述

文本检测与识别是光学字符识别（OCR）技术的基础，目前许多互联网公司都推出了自己的OCR系统，支持文档、票据和证件等图像文本内容的识别，说明对于一般纸质文档，文本检测和识别的技术已经比较成熟。而自然场景下的文本往往具有特殊字体、形状和排版，而且通常会受到复杂背景的干扰，因此要对这些图像中的文本进行检测和识别仍然存在着一定的困难。近年来，针对自然场景下的文本检测，许多研究者提出了新的思路和方法，突破不同的难点，取得了越来越好的效果。

已有的文本检测方法通常分为基于回归的方法以及基于分割的方法。基于回归的方法与一般目标检测中的回归过程类似，并在其他环节上进行改进，比如CTPN[1]额外加入了BLSTM结构，利用文本行的序列特征来帮助检测文本；TextBoxes[2]调整了参考边界框和卷积核的形状来适应文本行特点；EAST[3]及TextBoxes++[4]等方法则支持了倾斜文本框或者任意四边形文本框的标注形式；SegLink[5]将文本行分割为多个小文本块后分别进行回归，再连接相邻小块得到检测结果。PixelLink[6]属于基于分割的文本检测方法，它对每个像素进行文本/非文本预测以及链接预测后，通过正链接将属于文本的像素连接在一起得到文本实例分割的结果，进而直接得到文本行的位置。

然而，即便上述提到的一些方法支持使用倾斜矩形或者任意四边形的标注进行训练和预测，解决了倾斜和变形文本行的检测问题，也难以完美实现自然场景中存在的弯曲文本的检测。因此近来出现了一些检测弯曲文本的方法，从弯曲文本行的特点出发进行针对性的创新。比如MSR[7]预测文本的中心区域以及在水平和垂直方向上距离边界的距离；TextField[8]类似地为每个像素预测一个二维向量，大小非 0 表示属于文本像素，方向指向最近的文本实例边界，经过后处理便可得到实例分割结果；TextMountain[9]将文本实例中心和边界形容为山峰的山顶和山脚，进行文本/非文本预测、山峰预测以及方向预测；TextSnake[10]则将文本区域表示为由一条中心线贯穿的多个圆形区域，每个圆形区域有各自的半径和关联的方向角度。

在今年举办的CVPR2019 会议中收录的一些自然场景文本检测方法都专注于解决弯曲变形文本的检测问题，包括CRAFT[11]、LOMO[12]、PSENet[13]以及CSE[14]等，本文主要对CRAFT方法进行介绍。

2. CRAFT方法分析

CRAFT文章的创新点主要在于：（1）预测字符位置和字符间亲和度（文中称为Affinity，可以理解为相邻两个字符间的关联程度），与PixelLink等方法类似，采用自底向上的检测思路，不过它关注的不是文本行片段或者小块文本区域，也不是单独的像素，而是文本行中的单个字符，从字符和字符之间的亲和度出发来得到整个文本行；（2）采用了弱监督学习的方式来训练，由于缺少字符级别位置的标注数据，作者以此方式通过单词级别标注的数据来生成单个字符级别标注的训练数据，结合额外的合成训练数据，一起对网络进行训练。下面介绍CRAFT的总体结构和部分细节的实现方法。

2.1. 网络架构

首先是CRAFT网络的主要架构。与大多数检测方法一样，也采用了VGG-16 作为基本的特征提取结构，同时使用了批量归一化的操作，此外，在上采样的过程中还加入了跳跃连接，因为它融合了特征提取过程中相同特征图尺寸大小的特征，因而能够更好地利用浅层特征。整个网络结构与全卷积网络类似，通过最后的卷积

层来得到两个以热图形式表示的得分图：区域得分图和亲和度得分图。

区域得分图（图 1 上）表示出了每个字符所在的位置，通过设定阈值可以得到每个字符的边框位置。而亲和度得分图（图 1 下）则表示了每两个字符间的关联程度，亲和度高表示这两个字符更有可能具有相关性，应该是同一个文本行中的前后字符。

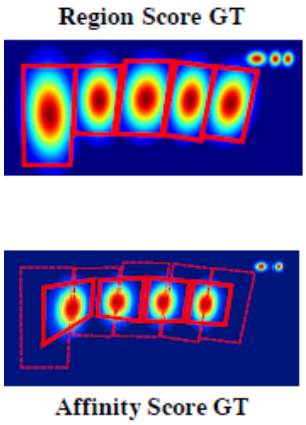

图 1 区域得分图（上）和亲和度得分图（下）

2.2. 得分图标注生成

那么如何为每一张文本图像生成对应的区域得分图以及亲和度得分图标注作为网络的训练数据呢？假设我们的训练图像数据已经具有单个字符的标注信息（四个顶点的坐标位置），那么首先利用字符的边界框位置来生成表示亲和度的边界框位置。如图 2 左所示，绘制两个字符边界框的对角线，在每个边界框中都会得到上下两个三角形，连接这四个三角形的中心点，便得到一个新的四边形，作者以此作为亲和度的边界框位置信息。对于这两类边界框，作者通过对一个 2D的高斯分布热图进行透视变换变形到对应的边界框形状，来得到适合这个形状的高斯热图以进行填充。

2.3. 字符级别的标注生成

人工合成的数据集在生成过程中便能获得字符级别的标注信息，但是已有的真实图像的数据集往往不具有字符级别的标注，最多也只是提供了单词级别的标注，如何将单词级别的标注信息转化为字符级别的标注信息以便该网络能够进行训练是一个问题，因此作者使用了弱监督学习的方法，从每一个单词级别的标注来生成字符级别的标注。

整个网络训练的过程如图 3 所示。真实图像通过已有的单词级别的标注裁减得到各个单词的图片，输入部分学习的中间模型来预测裁剪单词图像的字符区域得分图，并根据此使用分水岭分割算法来生成字符级别的边界框结果，结合 2.2 节中的方法便可得到对应的亲和度得分图标注。

可以看到，这样得到的边界框个数可能与实际的字符数不相符，可能将两个字符视为同一个字符或者将一个字符拆分成了两部分，因而如果直接将这个预测结果作为标注结果，会影响学习效果。因此，作者为每个裁剪的图像设置了一个置信度，其值与该文本框中分割得到的字符框数和真实标签字符的数量的比值有关（如果真实图像中有 6 个字符，而实际只区分得到了 5 个边界框，那么其置信度只有 5/6），作为训练期间的学习权重。用公式表示为：

$$s_{conf}(w) = \frac{l(w) - \min(l(w), |l(w) - l^c(w)|)}{l(w)}$$

其中$l(w)$表示单词标注w的长度，$l^c(w)$则是检测到的字符框数。作者给出的置信度计算公式表示同时考虑到了得到的字符框数远多于实际标注长度（大于两倍长度）的情况。这个置信度会被乘到根据两个得分图结果计算的L2 损失和上，当然，如果置信度较低，那么该数据会被直接舍弃。作者定义的损失函数给予了文本区域位置得分图和亲和度得分图这两部分损失相同的权重大小，表明字符分割和字符连接两部分有着同等的重要性。

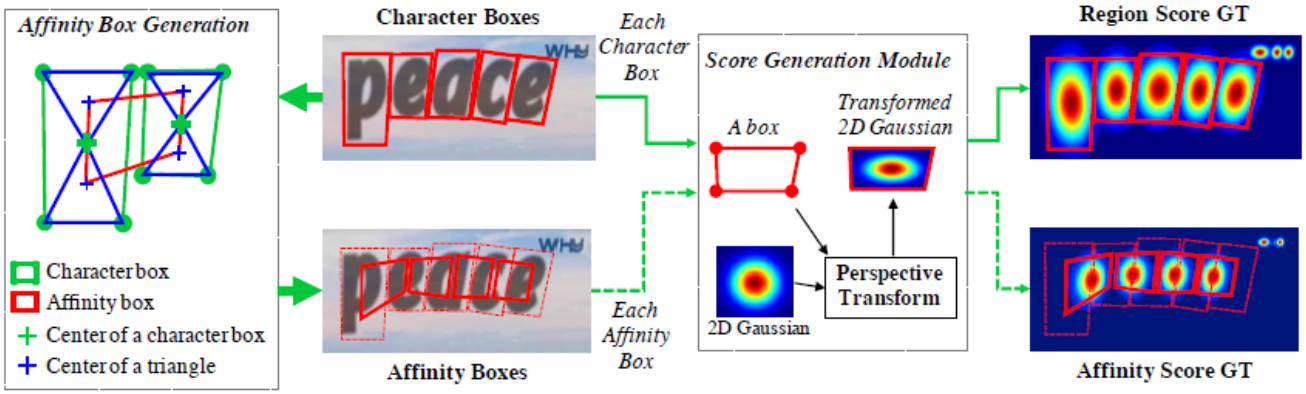

图 2 由字符级标注生成得分图标注的流程

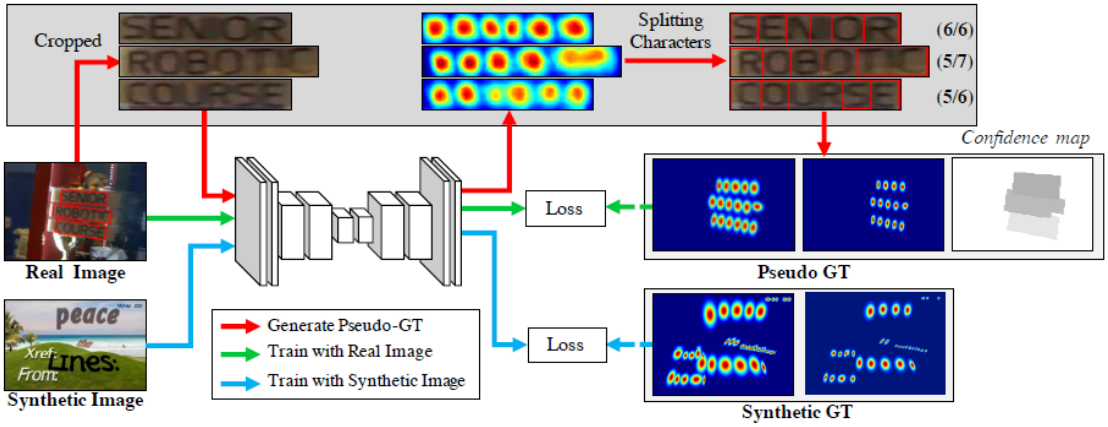

图 3 CRAFT网络的训练流程

网络使用合成图像和真实图像同时训练，随着网络训练过程的不断进行，网络给出的字符级别的标注结果也将越来越准确，进而促进网络的效果提升，实现了弱监督学习的目的。

2.4. 得到边界框结果

为了得到弯曲文本行的边界框结果，只需将区域得分图和置信度得分图中大于一定阈值的像素设为文本像素，然后计算连通区域的结果并使用最小矩形框包围这个连通区域。然后对这个矩形框从左往右扫描，找到每个字符区域中局部最大值点所在的竖直线段（图 4 中的蓝色线），并将所有的极值点连在一起称为中心线（黄色线），将每条蓝色线段旋转到与黄色线段垂直后得到的线段的端点即为最后边界框的顶点（两侧的线段需要往外平移一定距离以包围整个字符）。

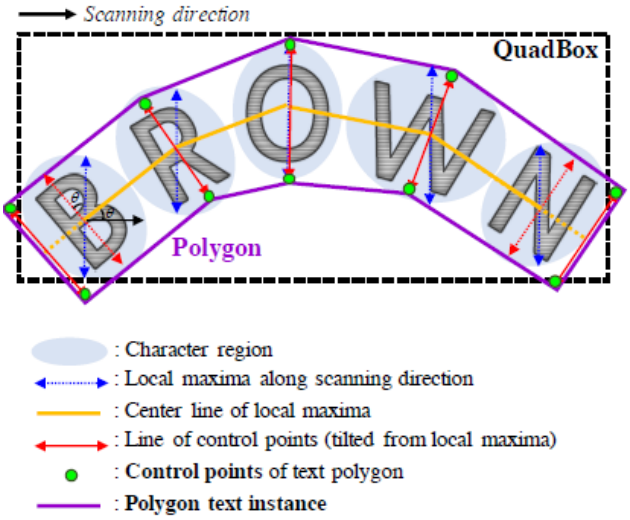

图 4 文本边界框生成过程

3. CVPR2019 中提出的其他文本检测方法

LOMO将整个检测过程分为三个部分，初步的位置回归（DR），矩形框位置迭代细化（IRM）以及生成特殊形状标注（SEM）。整个处理流程如图 5 所示。

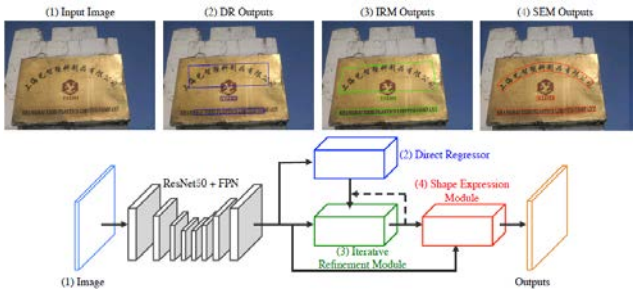

图 5 LOMO网络架构

三个处理阶段共用同一个特征提取结果，DR阶段给出文本行的粗略位置，可以看到由于CNN感受野的限制，对于长文本行的检测并不完全，该部分的损失结合文本/非文本分类损失及位置回归损失得到；IRM阶段通过预测每个矩形框四个角落的attention map，经过缩减得到矩形框四个顶点坐标的偏移向量，得到了位置偏移后可以修正文本框位置后继续进行迭代操作，因此也可以称为迭代的细化过程，该过程的损失通过计算对应的坐标偏移距离得到；最后的SEM则在前面步骤得到的矩形框预测文本区域、文本中心线以及对应的到文本行上下两侧的距离，以得到贴近文本形状的检测结果，这里和DR阶段均采用了FCN的网络结构，该部分的损失为三个预测部分的损失加权。整个网络可以同时进行训练，三个部分的损失使用了相同的权重。

PSENet也是基于分割的，思路很简单，如图 6 所示，第一步也是通过ResNet和FPN提取特征，然后进行多尺度融合得到图中的F，用来生成共n个不同尺寸的分割结果，即每个分割结果中的区域大小都不一样。显然，

分隔区域最小的S_1可以用来得到不同的文本实例，得到不同的连通区域，同时结合S_2的分割结果，使用广度优先搜索的方式同时对得到的每个连通区域周围进行扩展，得到与该分割结果类似大小的连通区域，后续进行类似的操作，最后得到整个文本块的检测结果。

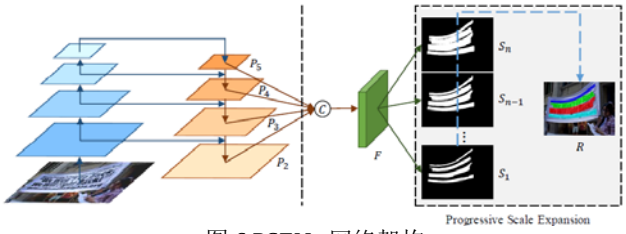

图 6 PSENet网络架构

该方法通过计算不同的d_i来表示不同的标注区域距离原始标注边界的距离，进而生成不同大小的分割标注。训练的损失包括了原始实例分割的损失以及每个缩小后的分割结果的损失。

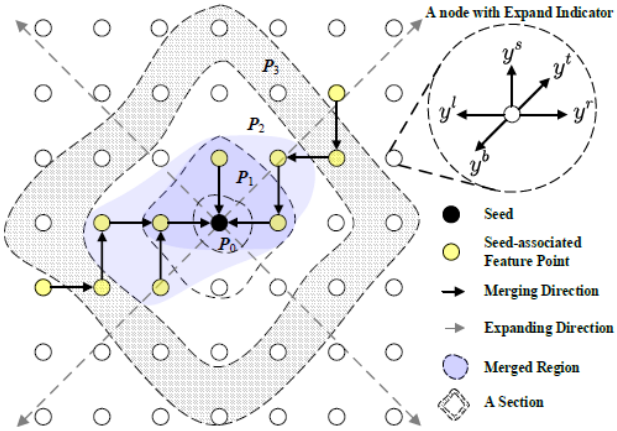

图 7 CSE检测结果生成过程

CSE检测文本的过程如图 7 所示，通过使用Fast RCNN等一般目标检测方法得到的目标检测框内的区域进行双线性插值得到一个大小为$S \times S$的特征图，并将中心作为种子（seed），初始的合并区域只包括seed一个点。接下来要做的就是从该中心seed往外扩展，首先对与该seed距离为 1 的那些点（section）计算一个扩展指示向量，这个向量表示了这个点往上下左右方向合并以及无合并方向的概率，如果其中的主合并方向指向已经在合并区域内的点，那么这个点也会被加入到合并区域内。接着继续往外扩展，每个section内的点到seed点的距离都相同。该方法由于选取的seed位置的不同，对应的扩展指示向量也应该有所不同，但同一section内的点是相互独立的，所以当前section内每一个点的扩展指示向量的计算都依赖于seed和前面section内的点的条件概率，后续便是对应的计算过程。

4. 实验结果

Total-Text[15]数据集和SCUT-CTW1500[16]数据集是最近提出的用于自然场景弯曲文本检测的数据集，前者包含了英文弯曲文本，后者包含了中英文的弯曲文本，表 1 和表 2 分别给出了上述四篇文章提出的方法在这两个数据集上对应的测试结果，具体数据由原始论文给出。

表 1 Total-Text数据集上的测试结果

方法	Precision/%	Recall/%	F-score/%
CRAFT	**87.6**	**79.9**	**83.6**
PSENet-1s	84.0	78.0	80.9
LOMO MS	**87.6**	79.3	83.3
CSE	81.4	79.7	80.2

表 2 SCUT-CTW1500 数据集上的测试结果

方法	Precision/%	Recall/%	F-score/%
CRAFT	**86.0**	**81.1**	**83.5**
PSENet-1s	84.8	79.7	82.2
LOMO MS	85.7	76.5	80.8
CSE	78.7	76.1	77.4

ICDAR 2015 数据集不包含弯曲的文本，但是包含了不同形状和方向的文本，上述方法在该数据集上测试的结果如表 3 所示，具体数据由原论文给出。

表 3 ICDAR 2015 数据集上的测试结果

方法	Precision/%	Recall/%	F-score/%
CRAFT	89.8	84.3	86.9
PSENet-1s	88.7	85.5	87.1
LOMO MS	87.6	**87.8**	**87.7**
CSE	**92.3**	79.9	85.7

可以看到，在弯曲文本行的检测上，CRAFT有着较好的表现，在这些方法之中有着最高的准确率以及召回率。同时CRAFT由于网络输出和后处理都比较简单，具有较高的FPS，在处理效率上比较高。在一般的场景文本检测数据集ICDAR 2015 上的测试结果表明，CRAFT仍然有着不错的效果，但是LOMO Multi-Scale方法表现最好，可能是由于其对边框的修正过程和形状预测过程起到了较好的帮助作用。

5. 分析与讨论

之前的文本检测方法通常通过回归的方式来得到边界框结果，在变形文本的检测上，借助分割结果来获得检测结果似乎也是一个较好的方式，本文提到的CRAFT和PSENet都使用到了分割的思想，而LOMO和CSE则在一般检测回归结果的基础之上继续进行后续的处理来得到更为精确的检测结果。

在文本中主要推荐了CRAFT方法的原因是其原理

简单易懂，字符级别的标注更具有实际意义，对于文本检测任务有着更好的效果，同时后处理获得检测框也非常容易，在语言的适应性方面需要单个字符便于分割，一般的英文和中文都是可以支持的，如果字符是等宽的，那对于弱监督学习过程也很有帮助。由于CRAFT从字符级别的检测结果开始与周围字符进行连接，因此不需要大的感受野信息，即不需要进行目标检测和大多数文本检测需要的多尺度训练。但是该方法和PixelLink、CSE等基于扩展和链接的方法都有一个很明显的缺点，那就是当文本行中的字间距较大时，很难在扩展过程中将相隔较远的文本连接起来，对于PSENet，由于基于文本行整体分割的结果，也具有类似的问题。总体来说，这些文本检测方法都考虑到了文本行的特点，提出了具有针对性的检测方案，具有创新意义。

References

[1] 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.

[2] Liao M, Shi B, Bai X, et al. Textboxes: A fast text detector with a single deep neural network[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[3] 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 Recognition. 2017: 5551-5560.

[4] Liao M, Shi B, Bai X. Textboxes++: A single-shot oriented scene text detector[J]. IEEE transactions on image processing, 2018, 27(8): 3676-3690.

[5] Shi B, 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. 2017: 2550-2558.

[6] Deng D, Liu H, Li X, et al. Pixellink: Detecting scene text via instance segmentation[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[7] Xue C, Lu S, Zhang W. MSR: Multi-Scale Shape Regression for Scene Text Detection[J]. arXiv preprint arXiv:1901.02596, 2019.

[8] Xu Y, Wang Y, Zhou W, et al. TextField: Learning A Deep Direction Field for Irregular Scene Text Detection[J]. IEEE Transactions on Image Processing, 2019.

[9] Zhu Y, Du J. TextMountain: Accurate Scene Text Detection via Instance Segmentation[J]. arXiv preprint arXiv:1811.12786, 2018.

[10] Long S, Ruan J, Zhang W, et al. Textsnake: A flexible representation for detecting text of arbitrary shapes[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 20-36.

[11] Baek Y, Lee B, Han D, et al. Character Region Awareness for Text Detection[J]. arXiv preprint arXiv:1904.01941, 2019.

[12] Zhang C, Liang B, Huang Z, et al. Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes[J]. arXiv preprint arXiv:1904.06535, 2019.

[13] Li X, Wang W, Hou W, et al. Shape robust text detection with progressive scale expansion network[J]. arXiv preprint arXiv:1806.02559, 2018.

[14] Liu Z, Lin G, Yang S, et al. Towards Robust Curve Text Detection with Conditional Spatial Expansion[J]. arXiv preprint arXiv:1903.08836, 2019.

[15] Ch'ng C K, Chan C S. Total-text: A comprehensive dataset for scene text detection and recognition[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017, 1: 935-942.

[16] Yuliang L, Lianwen J, Shuaitao Z, et al. Detecting curve text in the wild: New dataset and new solution[J]. arXiv preprint arXiv:1712.02170, 2017.