

OCR技术发展

一、OCR 数据集分类

按输入方式分类

- 1. 印刷体文字
- 2. 手写体文字（由扫描仪输入 / 由手写板输入）

The Initiative Application of Tool Like SCIgen

Spiciest Robot , Bye Another Earth and Hello World

Abstract

SCIgen is a program that generates random computer science research papers, including papers, figures, and citations. Since almost all academic paper of these days are meaningless, papers generated by SCIgen are sufficient to fill up all journals. Researchers and universities are no longer needed, large amount of money can be saved for that.

2 Methodology

Do not repeat yourself. At first, human mind is too big, humans is the only source of all improvement. In addition, academic corruption is inevitable, evil will take over the whole academic measure. However, if they do not currently, all journal and conference papers will be completely meaningless. To achieve the sacred will of humanity, we should tell to truth, no dream is good dream. Or even better, we can still trust education system, which bring significant benefit to our society.

1.1 Introduction

SCIgen was a hand written content free program to form all elements of the papers. The aim is to maximize amusement rather than coherence. One useful purpose for such a program is to auto generate sub missions to conferences that can, suspect might have very low submission standards. In fact, one of our papers was accepted to SCIRUG. Our plan is:

Fill up all journals with a pre-generated system.

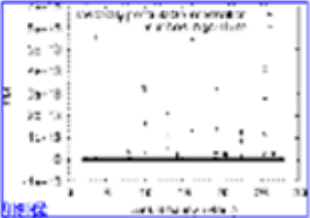
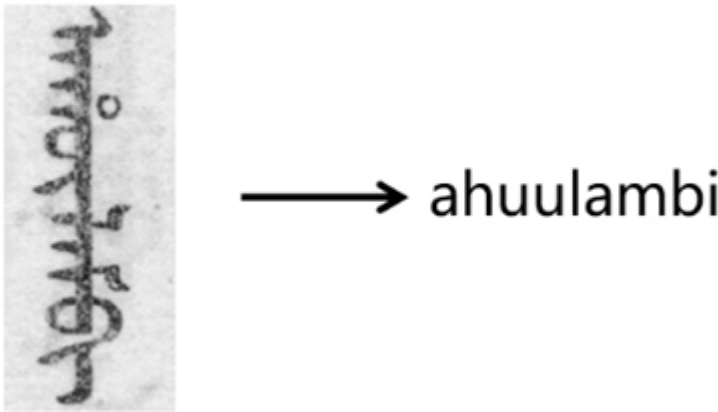


Figure 1: A figure generated by SCIgen

按识别字符集分类

- 1. 英文，中文，日文，韩文等
- 2. 中文及少数民族语言，常用各种字体



常规的 OCR 文字识别处理的过程包括：1、图像输入、预处理：二值化图片、噪声去除、倾斜校正；2、版面分析：把页面分为横排文本、竖排文本、表格、图片等不同区域，帮助字符切割、识别 OCR；3、

设置语种：选择需要什么 OCR 语种的引擎程序；4、输出结果：输出 OCR 识别结果为原版原样的优质文件；

传统印刷体文字识别

倾斜校正

印刷体文本资料大多是由平行于页面边缘的水平 (或者垂直) 的文本行 (或者列) 组成的，即倾斜角度为零度。然而在文本页面扫描过程中，不论是手工扫描还是机器扫描，都不可避免地会出现图像倾斜现象。而倾斜的文档图像对后期的字符分割、识别和图像压缩等工作将产生很大影响。为了保证后续处理的正确性，文本图像进行倾斜检测和校正是十分必要的。

文本图像的倾斜校正分为手动校正和自动校正两种。手动校正，是指识别系统提供某种人机交互手段，实现文本图像的倾斜校正。自动校正，是指由计算机自动分析文本图像的版面特征，估计图像的倾斜角度，并根据倾斜角度对文本图像进行校正。

文本图像的倾斜检测方法有许多种，主要可以划分为以下五类：基于投影图的方法，基于 Hough 变换的方法，基于交叉相关性的方法，基于 Fourier 变换的方法和基于最近邻聚类方法。

1. 最简单的基于投影图的方法是将文本图像沿不同方向进行投影。当投影方向和文字行方向一致时，文字行在投影图上的峰值最大，并且投影图存在明显的峰谷，此时的投影方向就是倾斜角度。

2.Hough 变换也是一种最常用的倾斜检测方法，它是利用 Hough 变换的特性，将图像中的前景像素映射到极坐标空间，通过统计极坐标空间各点的累加值得到文档图像的倾斜角度。

3.Fourier 变换的方法是利用页面倾角对应于使 Fourier 空间密度最大的方向角的特性，将文档图像的所有像素点进行 Fourier 变换。这种方法的计算量非常大，目前很少采用。

4. 基于最近邻聚类方法，取文本图像的某个子区域中字符连通域的中心点作为特征点，利用基线上的点的连续性，计算出对应的文本行的方向角，从而得到整个页面的倾斜角。

OCR 识别效果影响因素

1· 图片：质量通常建议 150dpi 以上 建议扫描仪分辨率设置为 300DPI 规格的参数；手机拍照的话建议摄像头像素为 500 万像素以上的摄像头；

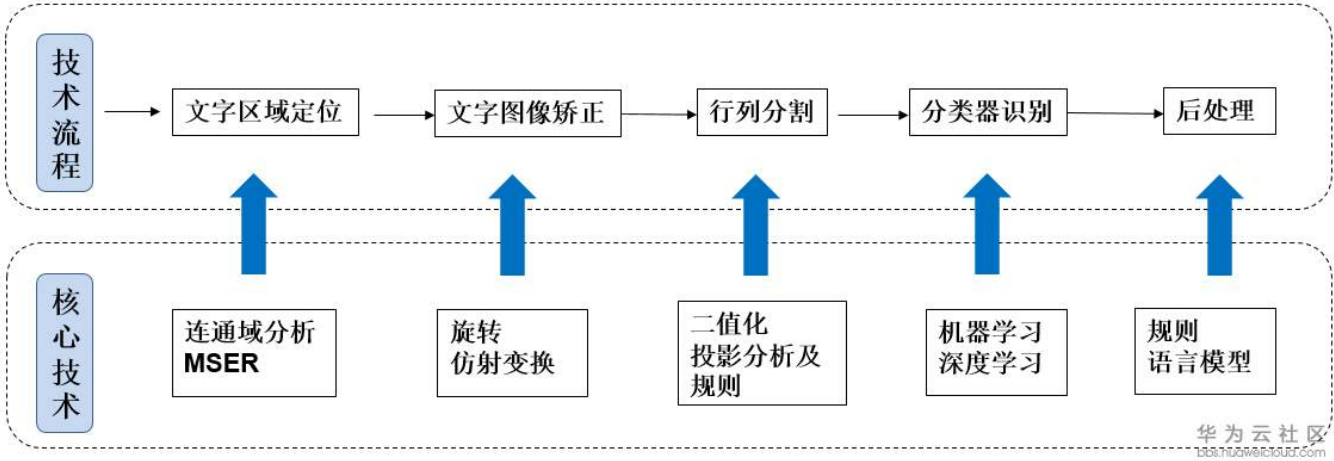
2· 颜色：通常对彩色识别不好，黑白的图片较高 常规的传统的 OCR 识别要求为白底黑字或者浅底黑字；如果是彩色背景图片的文字则需要特殊的 OCR 识别程序，比如文通视频文字识别系统；

3· 字体：目前全世界范围内的 OCR 技术都是针对“宋体印刷字”的字形做识别库的，若是是手写识别率低。

4· 程序：好的 OCR 程序遇到好的图片，识别效果才会优秀；在这里推荐北京文通科技推出的 OCR 程序，包括 OCR-SDK 开发包产品、OCR 技术识别文档 APP 产品等等。

传统 OCR 与深度学习 OCR 比较

传统 OCR 文字识别方法：基于图像处理（二值化、投影分析）和统计机器学习 SVM 等提取图片文本内容，使用 OpenCV 等。



OCR 文字检测

一、文字检测介绍

OCR 技术是一种可以从扫描或拍摄的图片中提取文本信息的技术。它的主要任务是找到图片中的文字区域，并将其转换成可以编辑的文本。如下图所示，在医疗领域中，高效准确地检测出医疗文本区域对 OCR 识别系统的整体性能至关重要。

主要编码	附加编码	疾病名称	英文名称
	M80000/0	良性肿瘤	Neoplasm, benign
	M80000/2	动态未定肿瘤	Neoplasm, uncertain whether benign or malignant
	M80000/3	恶性肿瘤	Neoplasm, malignant
	M80000/6	转移性肿瘤	Neoplasm, metastatic
	M80001/3	溃疡恶变	Ulcer malignant change
	M80002/3	息肉恶变	Polyp malignant change
	M80010/0	良性瘤细胞	tumour cells, benign
	M80010/2	良性或恶性未肯定瘤细胞	tumour cells, uncertain whether benign or malignant
	M80010/3	恶性瘤细胞	tumour cells, malignant
	M80020/3	小细胞型恶性肿瘤	Malignant tumour, small cell type
	M80030/1	良性或恶性未肯定巨细胞瘤	giant cell type, uncertain whether benign or malignant
	M80030/3	巨细胞型恶性肿瘤	Malignant tumour, giant cell type
	M80040/3	梭形细胞型恶性肿瘤	Malignant tumour, fusiform cell type
	M80100/0	良性上皮肿瘤	Epithelial tumour, benign
	M80100/2	原位癌	Carcinoma in situ NOS
	M80100/3	癌	Carcinoma NOS
	M80100/6	转移性癌	Carcinoma, metastatic NOS
	M80101/2	上皮内癌	Intraepithelial carcinoma
	M80110/0	良性上皮癌	Epithelioma, benign
	M80110/3	恶性上皮癌	Epithelioma, malignant
	M80120/3	大细胞癌	Large cell carcinoma NOS
	M80120/6	转移性大细胞癌	Large cell carcinoma, metastatic
	M80130/3	大细胞神经内分泌癌	Large cell neuroendocrine carcinoma
	M80140/3	具有柱状型大细胞癌	Large cell carcinoma with rhabdoid phenotype
	M80150/3	玻璃状细胞癌	Glassy cell carcinoma
	M80200/3	未分化型癌	Carcinoma, undifferentiated NOS
	M80200/6	转移性未分化癌	Carcinoma, undifferentiated metastatic NOS
	M80210/3	癌，间变	Carcinoma, anaplastic NOS

二、传统的文字检测方法

传统的文字检测方法主要是基于图像处理和机器学习的算法，大致可分为两类：基于连通域的方法和基于滑动窗口的方法。

2.1 基于连通域的自然场景文本检测方法：

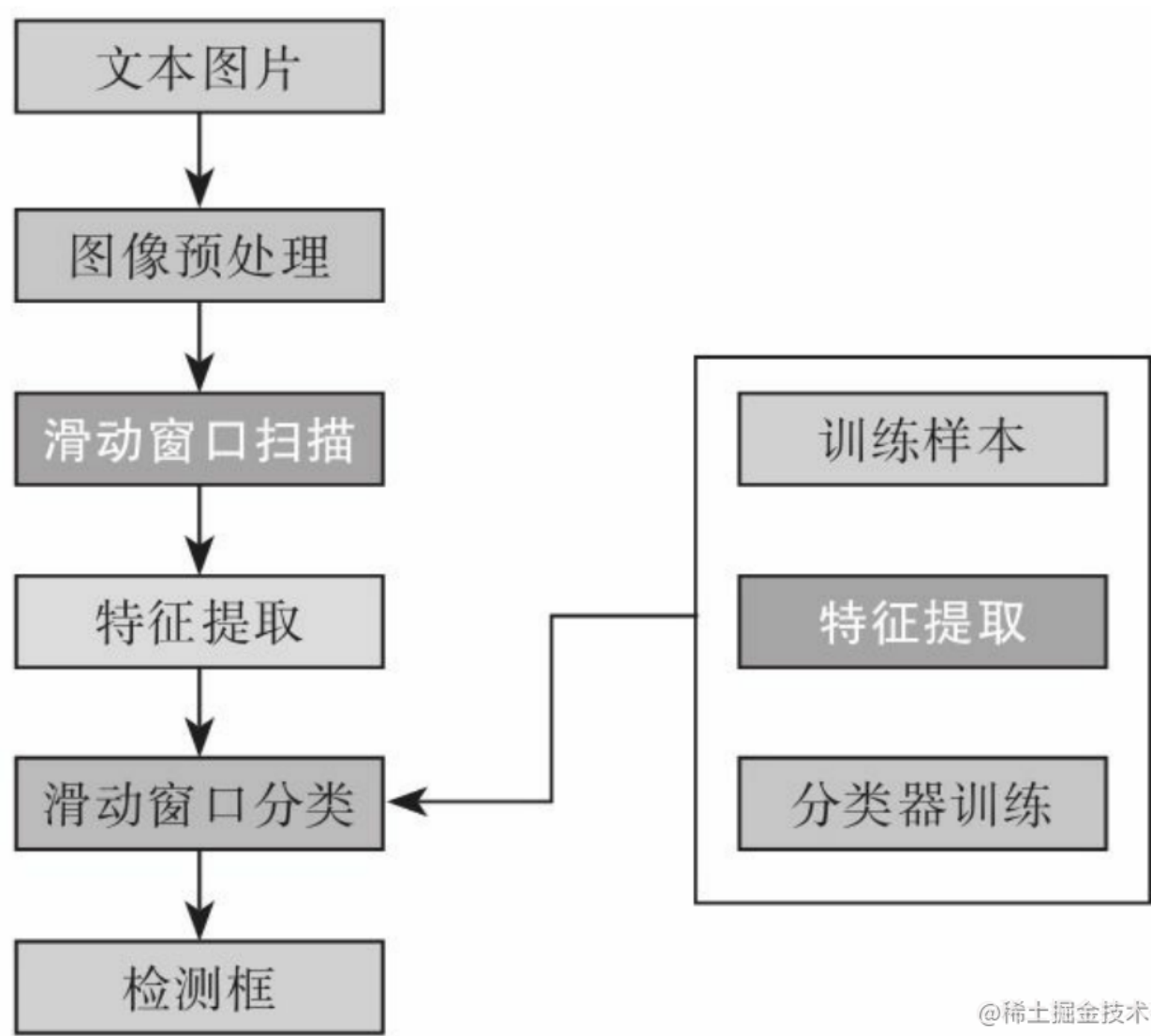
传统文本检测方法最常见的是基于连通性检测算法。基于连通区域分析通过检测像素值相等、且前景像素点位置邻近区域的形状、大小、位置等特征，判断是否为文本区域。通过获取候选文本区域，可以将搜索范围缩

小。但是这种方法很大程度上依赖于检测出来的文本连通区域，对于文本检测的召回率和文本轮廓的准确性有很大影响。

2.1.1 基于边缘检测的文本检测方法：基于各类算子的边缘检测的算法，如 Canny 边缘检测、Sobel 边缘检测等，通过检测图像中的边缘来定位文本区域。

2.2 基于滑动窗口的文本检测方法：

这种方法通常是基于单个字符的分类器的，它使用滑动窗口来处理候选框。方法是在图像上设置一个滑动子窗口，并在每个位置应用检测算法。然而，在复杂的场景中（如光照、阴影等），这种方法可能导致字符分类效果不好。同时，如何选择适当的检测窗口滑动步长也很复杂。



@稀土掘金技术社区

2.3 传统文字检测技术总结

传统文本检测方法	年份	优势	劣势
Zhong et al.	1995	支持自然场景文本图像检测	局限于简单场景
Jain et al.	1998	支持自然场景文本图像检测	局限于简单场景，依赖于人工设计的规则
Li et al.	2000	支持检测和追踪视频中的文本	仅支持水平文本检测
Kim et al.	2003	支持自然场景文本图像检测和视频文本检测	局限于简单场景，仅支持水平文本检测
Chen et al.	2004	支持复杂场景下的文本图像检测，速度快	仅支持水平文本检测
Lyu et al.	2005	支持视频文本检测，多语言	仅支持水平文本检测
Liu et al.	2006	支持自然场景文本图像检测	仅支持水平文本检测
Wang et al.	2010	支持复杂场景下的文本图像检测	仅支持水平文本检测，需要词典
Epshtein et al.	2010	支持复杂场景下的文本图像检测，多语言，速度较快	仅支持水平（和近似水平）文本检测，依赖于人工设计的规则
Neumann et al.	2010	支持复杂场景下的文本图像检测，速度较快	仅支持水平（和近似水平）文本检测
Yi et al.	2011	支持多方向文本图像检测，多语言	局限于简单场景，依赖于人工设计的规则

(续)

传统文本检测方法	年份	优势	劣势
Shivakumara et al.	2011	支持多方向文本图像检测	生成文本块（区别于词条或文本行，依赖于人工设计的规则）
Yao et al.	2012	多语言，速度较快	依赖于人工设计的规则
Huang et al.	2013	支持复杂场景下的文本图像检测，非常健壮	仅支持水平文本检测，依赖于人工设计的规则
Huang et al.	2014	支持复杂场景下的文本图像检测，性能优越	仅支持水平文本检测

@稀土掘金技术社区

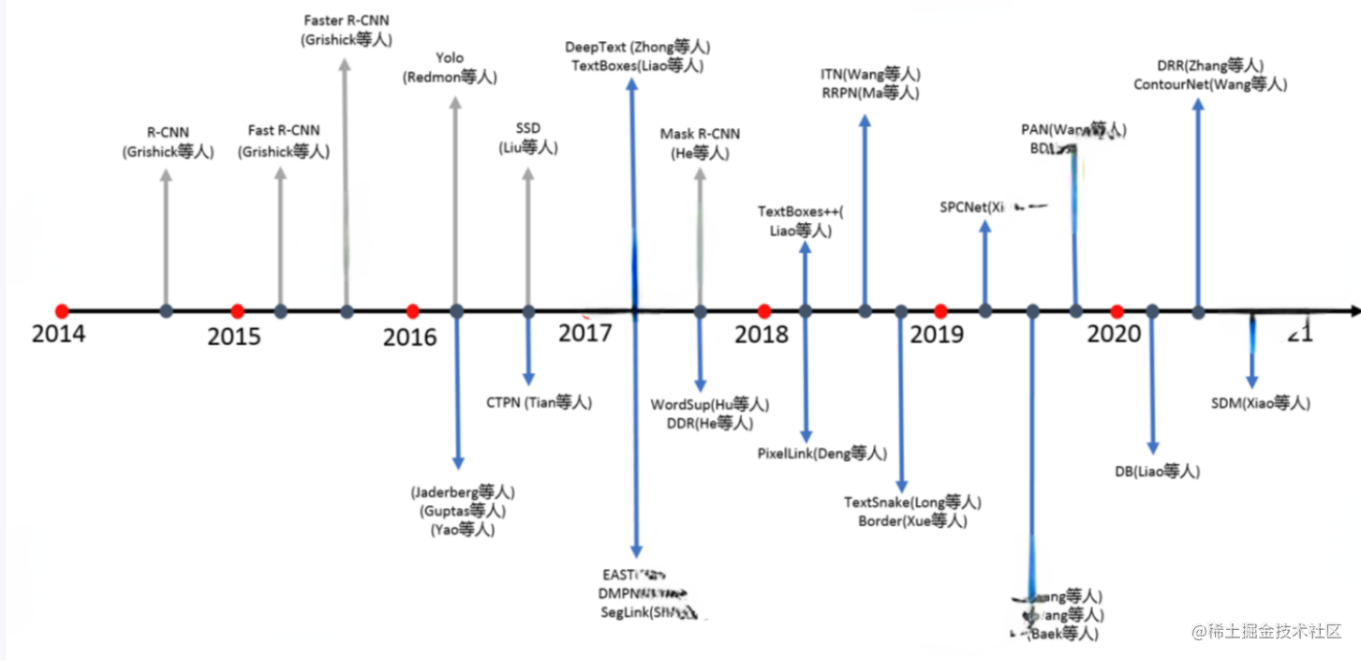
传统的文字检测方法通常结合了图像处理技术和机器学习算法，比如边缘检测、连通域和形态学等。虽然这些方法在理论上和稳定性上表现良好，但由于光照变化、褪色、噪声等因素的干扰，以及文本区域的不规则性、多方向性和多尺度性等复杂场景，使得检测的效果不太理想。然而近年来，基于深度学习的文字检测方法逐渐成为主流。

三、基于深度学习的文字检测方法

3.1 深度学习文字检测技术发展

深度学习的文本检测方法是从小目标检测算法（比如 faster rcnn 和 ssd）启发而来的。研究人员开始尝试修改这些算法中的锚框尺寸来适应文本目标的特点。例如，在 2016 年出现的 CTPN 算法是在 faster rcnn 的基础上做的改进，通过修改锚框的尺寸来更好地检测文本。而在 2017 年，*Textbox 算法则是在 ssd 上做了改进，同样是为了适应文本的特点。随后，研究人员在这些基础上提出了各种改进方法，不断优化文本检测算法的性

能。目前主流的文字检测算法包括 DB 算法及其改进算法等。



3.2 场景文字检测方法分类

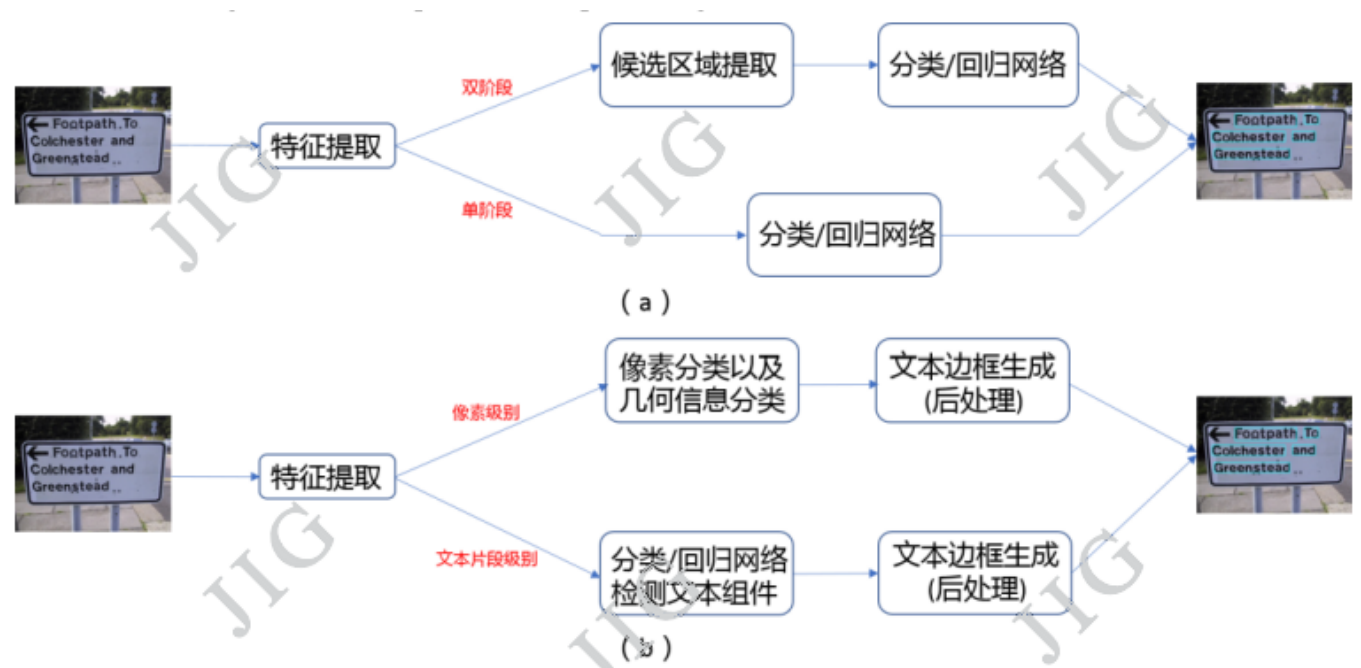


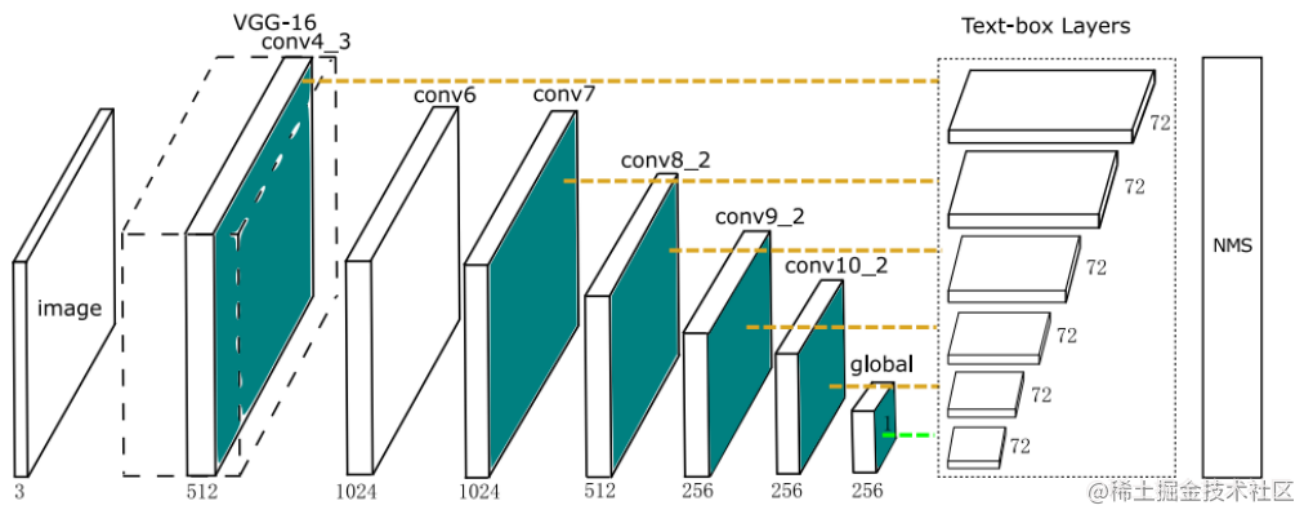
图 2 自然场景文本检测方法流程图。(a) 自顶向下；(b) 自底向上

Textboxes TextBoxes: A Fast Text Detector with a Single Deep Neural Network

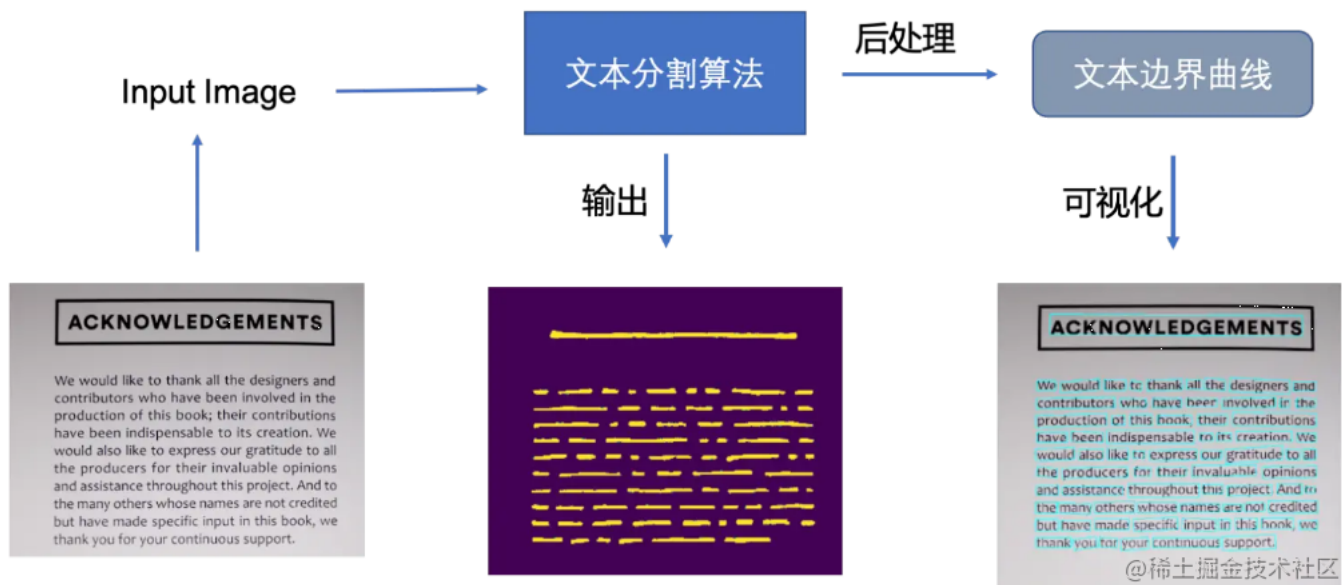
TextBoxes 是一种针对多方向文本检测的算法，使用多个检测框对文本区域进行精细化检测，并可以对文本区域进行旋转。主要对 SSD 算法进行优化，采用更大长宽比的预选框，采用 1x5 的卷积核；但只支持横向文本检测。

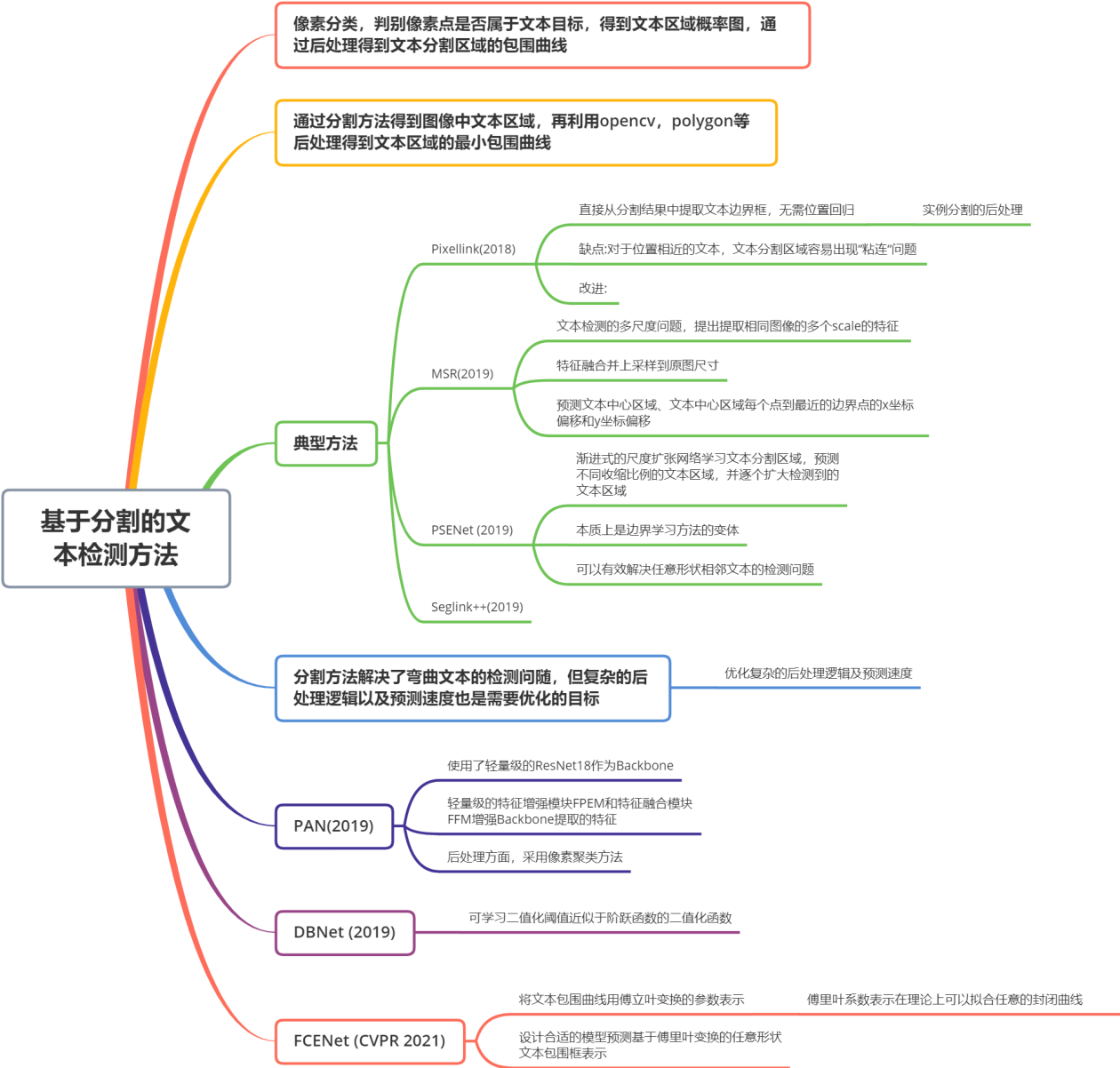
模型简介：Textboxes 为基于 SSD 框架检测模型，在其基础上进行改进，采用 28 层的全连接卷积网络，采用端到端训练，其主要改进如下：为了适应文字行细长型的特点，候选框的长宽比增加不同初始值。为适应文本行细长型特点，特征层也用长条形卷积核代替了其他模型中常见的正方形卷积核。为防止漏检文本行，还在垂直方向增加了候选框数量。为检测大小不同的字符块，在多个尺度的特征图上并行预测文本框，然后对预测结

果做 NMS 过滤。



3.3基于分割算法

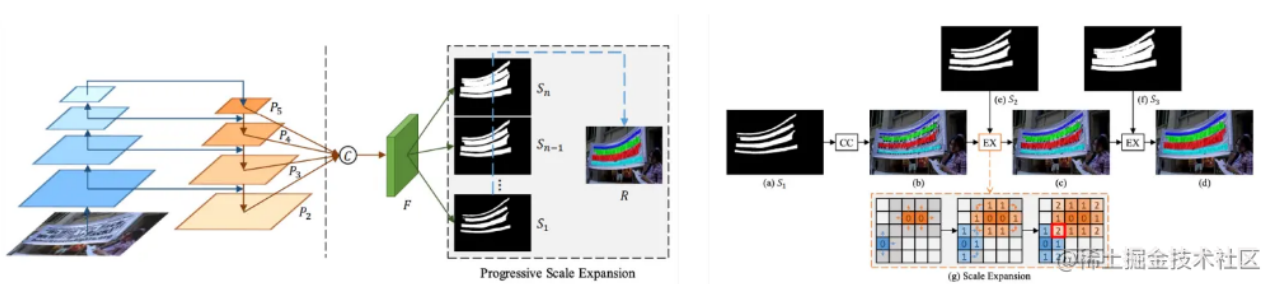




@稀土掘金技术社区

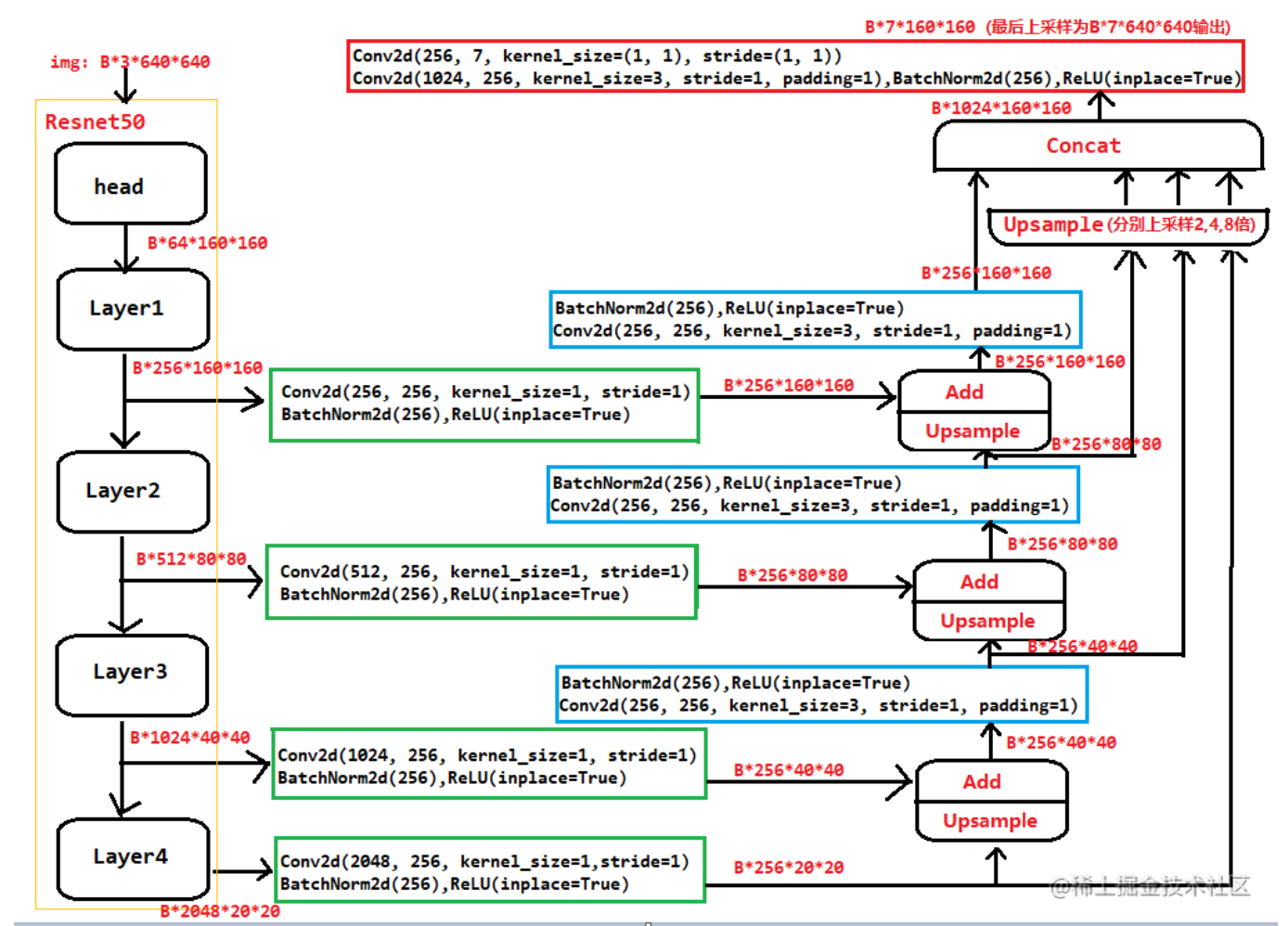
PSENet 2019 PSENET

PSENet 算法提出了一种解决基于分割的文本算法难以区分相邻文本问题的方法。这个算法使用了渐进式的尺度扩张网络来学习文本的分割区域。其主要思想是，通过预测不同收缩比例的文本区域，然后逐个扩大检测到的文本区域。这一过程可以有效地解决任意形状相邻文本的检测问题。



模型结构：

Psenet 网络采用了 resnet+fpn 的架构，通过 resnet 提取特征，取不同层的特征送入 fpn 进行特征融合，其结构如下图所示：



四、文字检测常见指标

IOU

IOU（IntersectionOverUnion，交并比）是目标检测中常见的评价标准，主要是衡量模型生成预测检测框（PredictedBoundingBox）和标注（GroundTruthBox）间的重叠程度，计算公式如下：



OCR 文字检测评估指标结果

通过对比上述方法在不同数据集上的文字检测评估指标，参考其他资料如下结果。评价指标为准确率 召回率

及 F1 值。

Method	ICDAR2015			MSRA-TD500			Total-Text			CTW1500		
	P	R	F	P	R	F	P	R	F	P	R	F
TextSnake (Long et al. 2018)	84.9	80.4	82.6	83.2	73.9	78.3	82.7	74.5	78.4	67.9	85.3	75.6
TextField (Xu et al. 2019)	84.3	83.9	84.1	87.4	75.9	81.3	81.2	79.9	80.6	83.0	79.8	81.4
PSE-Net (Wang et al. 2019a)	86.9	84.5	85.7	-	-	-	84.0	78.0	80.9	84.8	79.7	82.2
LOMO (Yu et al. 2018)	91.3	83.5	87.2	-	-	-	88.6	75.7	81.6	89.2	69.6	78.4
CRAFT (Baek et al. 2019)	89.8	84.3	86.9	88.2	78.2	82.9	87.6	79.9	83.6	86.0	81.1	83.5
PAN (Wang et al. 2019b)	84.0	81.9	82.9	84.4	83.8	84.1	89.3	81.0	85.0	86.4	81.2	83.7
DB (Liao et al. 2019)	91.8	83.2	87.3	91.5	79.2	84.9	87.1	82.5	84.7	86.9	80.2	83.4
ContourNet (Wang et al. 2020)	87.6	86.1	86.9	-	-	-	86.9	83.9	85.4	84.1	83.7	83.9
DRRG (Zhang et al. 2020)	88.5	84.7	86.6	88.1	82.3	85.1	86.5	84.9	85.7	85.9	83.0	84.5
MOST (He et al. 2021)	89.1	87.3	88.2	90.4	82.7	86.4	-	-	-	-	-	-
Raisi <i>et al.</i> (Raisi et al. 2021b)	89.8	78.3	83.7	90.9	83.8	87.2	-	-	-	-	-	-
TextBPN (Zhang et al. 2021)	-	-	-	86.6	84.5	85.6	90.7	85.2	87.9	86.5	83.6	85.0
DBNet++ (Liao et al. 2022)	90.9	83.9	87.3	91.5	83.3	87.2	88.9	83.2	86.0	87.9	82.8	85.3
DPTNet-Tiny(Ours)	90.3	77.4	83.3	88.4	82.9	85.6	88.9	83.9	86.3	86.4	81.5	84.1
DPTNet-Normal(Ours)	92.0	85.0	88.4	92.5	84.3	88.2	91.2	86.2	88.6	88.1	83.4	85.5

@稀土掘金技术社区