# OCR就在身边

光学字符识别（英语：Optical Character Recognition，OCR）传统是指对文本资料的图像文件进行分析识别处理，获取文字及版面信息的过程。场景文字识别（Scene Text Recognition，STR） 指识别自然场景图片中的文字信息。自然场景图像中的文字识别，其难度远大于扫描文档图像中的文字识别，因为它的文字展现形式极其丰富。也有人用OCR技术泛指所有图像文字检测和识别技术， 包括传统OCR技术与场景文字识别技术。这是因为，场景文字识别技术可以被看成是传统OCR技术的自然演进与升级换代。

## 1、OCR的发展历史

OCR的概念是在1929年由德国科学家Tausheck最先提出来的，后来美国科学家Handel也提出了利用技术对文字进行识别的想法。在60、70年代，研究的初期，多以文字的识别方法研究为主，且识别的文字仅为0至9的阿拉伯数字。早期的OCR软件，由于识别率及产品化等多方面的因素，未能达到实际要求。同时，由于硬件设备成本高，运行速度慢，也没有达到实用的程度。

进入20世纪90年代以后，随着平台式扫描仪的广泛应用，大大推动了OCR技术的进一步发展，使OCR的识别正确率、识别速度满足了广大用户的要求。直到今天，针对格式化文档图像的OCR，尤其是扫描文档，在理论与应用方面，都已经日趋成熟。

值得一提的是，过去OCR的识别以人工筛选特征为主，随着近几年神经网络的再次繁荣，OCR领域也开始搭上神经网络这辆快车，现在学界以及工业界先进的OCR识别算法已经离不开使用神经网络了。

## 2、应用概述

图像文字检测和识别技术有着广泛的应用场景。已经被互联网公司落地的相关应用涉及了识别快递单、识别身份证、识别银行卡、识别车牌、识别路牌、识别答题卡、识别会议白板、识别广告主干词、识别试卷、识别单据等等。

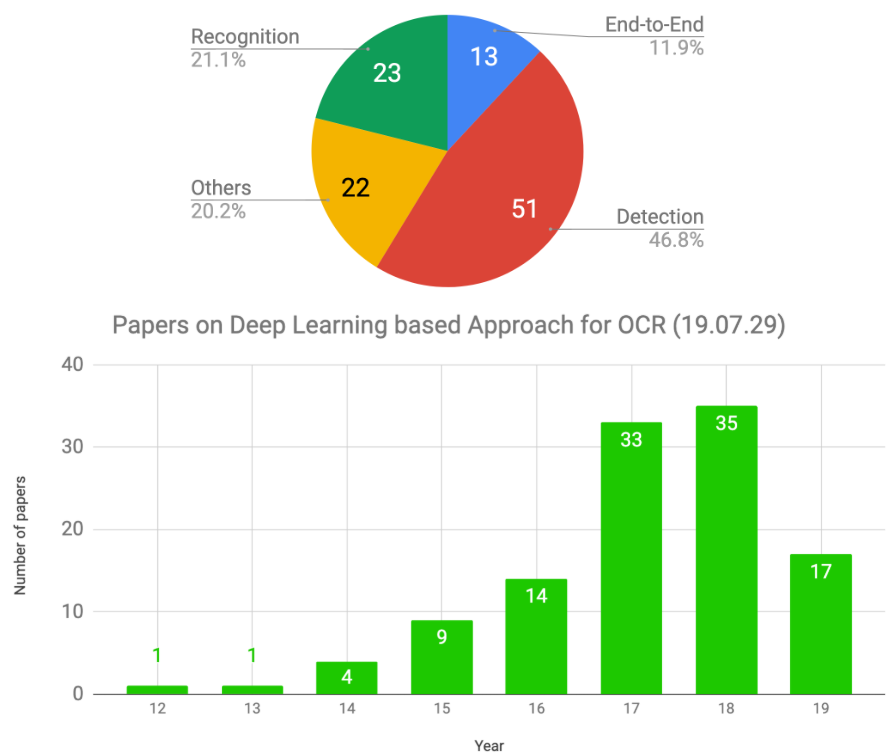
已经有不少服务商在提供图像文字检测和识别服务，这些服务商既包括了腾讯、百度、阿里、微软、亚马逊、谷歌等大型云服务企业，也包括了一些活跃在物流、教育、安防、视频直播、电子政务、电子商务、旅游导航等垂直细分行业的服务企业。这些企业既可以使用提前训练好的模型直接提供场景图文识别、卡证识别、扫描文档识别等云服务，也可以使用客户提供的数据集训练定制化模型（如票据识别模型），以及提供定制化AI服务系统集成等。

## 3、GitHub上OCR相关的开源社区

由于国际开源社区以及神经网络的快速发展，以及相关领域相关人员的分享知识共同促进AI发展的思想的普及，各大开源社区存在很多OCR相关的内容。

### awesome-deep-text-detection-recognition

这个GitHub仓库列出了从2012年到2019年发表的用深度学习的方法解决OCR问题的论文清单。



这个清单分为四个板块：1、文本检测类论文2、文本识别类论文3、端到端的文本识别论文4、其他。每个板块中的论文都是按照发表时间排序。这些收集工作是NAVER-LINE公司里面的OCR小组做的，这个仓库会跟新国际上主要的几个AI会议上新出的OCR相关的论文。

截止2019年12月15日，本项目在github平台上Star1.2k次，Fork 308次，共有Issues 4个，Pull request0个。

### Tesseract开源OCR引擎

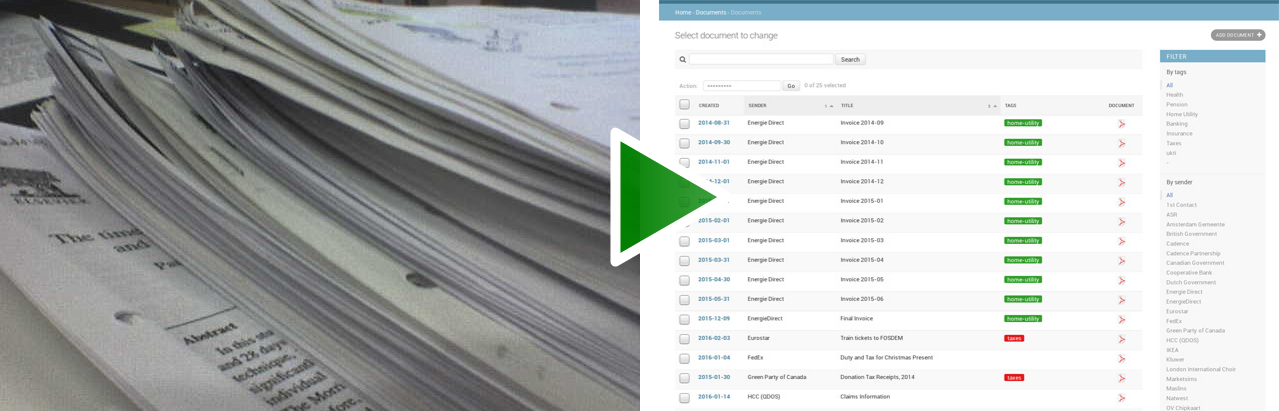
Tesseract，最初是由Hewlett Packard在1980年的惠普公司开发的，在2005年开源。2006年以后Google开始赞助这个项目。Tesseract可以处理很多自然语音，英语、葡萄牙语系、意大利语等。截止到2015年为止支持超过100种书面语言，并且可以通过训练学习轻松掌握其他语言。

最初Tesseract是用C语言写的，在1998年改用C++。Tesseract是无GUI交互的，可以通过命令后被执行。但是有一些其他软件提供GUI对Tesseract进行了封装。目前该项目的首席开发人员是Ray Smith。维护者是Zdenko Podobny。

截止2019年12月15日，本项目共Watch 1.6k次，Star 31.9k次，Fork 6k次，共有Issues 247个，Pull request 10个。

### Paperless

项目致力于通过OCR识别扫描文件的方法索引并存档所有的纸质文件。该项目创始于2015年，世界范围内使用的人很多。



上一次更新是在三个月前，到现在该项目不再更新，因为它已经完全可以满足我们目前的需求。但是开发团队并没有放弃这个项目，有疑问的用户还是可以提出issue并会得到解答，项目开发者也欢迎用户添加新的东西。

还有一个名为Mayan EDMS的项目，与Paperless的技术惊人地相似。它基于Django并使用Tesseract和Unpaper，Mayan EDMS有更多的其他功能，并配有一个漂亮的UI，但使用Python 2编程。因此Paperless占用资源相对少。

截止2019年12月15日，本项目共Watch 173次，Star 6.2k次，Fork 408次，共有commits1183个，contributor76个。

### ChineseOcr

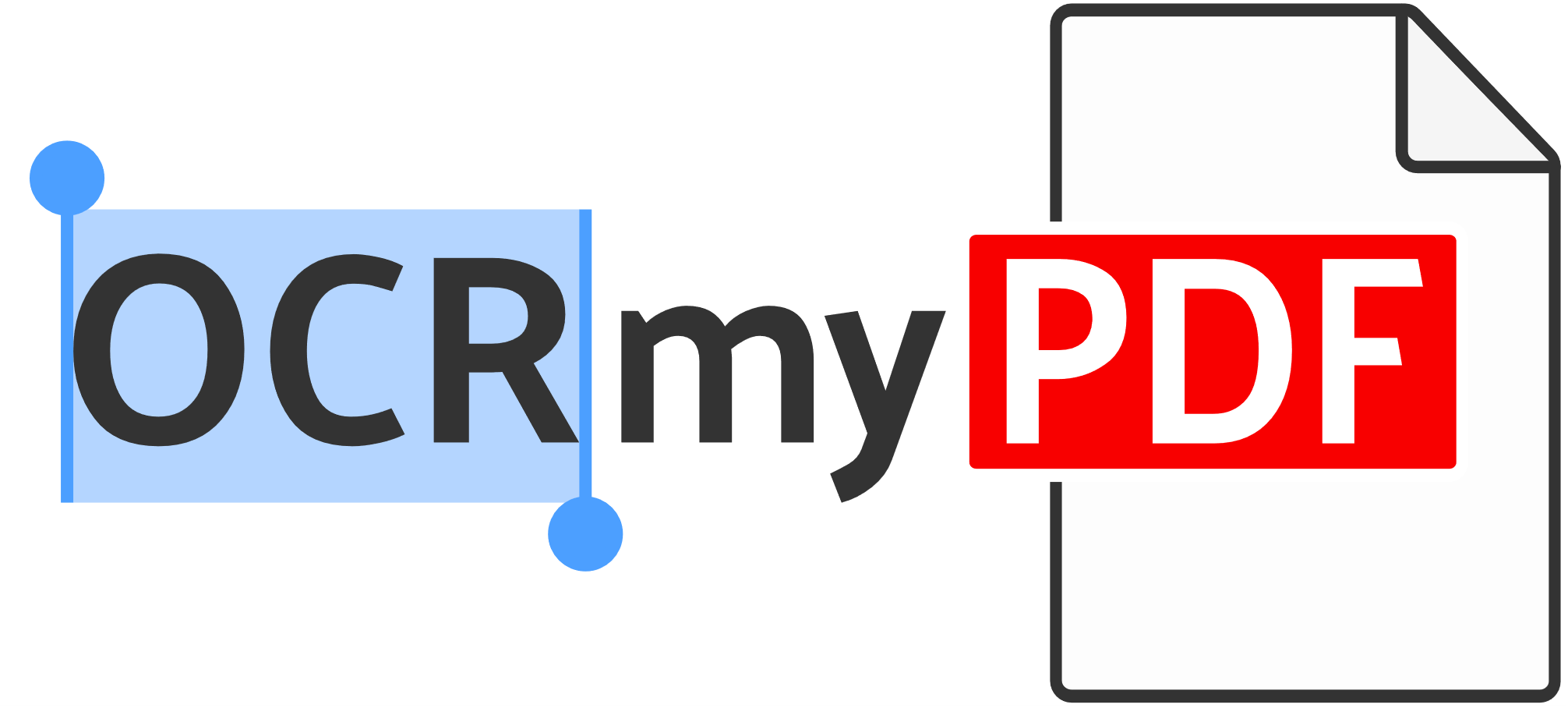
本项目基于[yolo3](https://github.com/pjreddie/darknet.git) 与[crnn](https://github.com/meijieru/crnn.pytorch.git) 实现中文自然场景文字检测及识别。这个项目实现了很多功能，例如：

* 文字方向检测 0、90、180、270度检测（支持dnn/tensorflow）
* 支持(darknet/opencv dnn /keras)文字检测,支持darknet/keras训练
* 不定长OCR训练(英文、中英文) crnn\dense ocr 识别及训练 ,新增pytorch转keras模型代码(tools/pytorch\_to\_keras.py)
* 支持darknet 转keras, keras转darknet, pytorch 转keras模型
* 身份证/火车票结构化数据识别
* 新增CNN+ctc模型，支持DNN模块调用OCR，单行图像平均时间为0.02秒以下
* 

由于这个项目融合多个框架，维护较困难，为了更好的部署应用，后续将只围绕darknet框架进行开发，不在支持keras、tensorflow、pytorch。

截止2019年12月15日，本项目共Watch 156次，Star 2.2k次，Fork 849次，共有issues299个，PR11个。

### OCRmyPDF



OCRmyPDF这个项目会分析PDF的每个页面，以确定捕获该页面上的所有信息而不丢失内容所需的色彩空间和分辨率（DPI）。OCRmyPDF使用多种策略，具体取决于输入选项和输入PDF本身，但通常来说，它会光栅化OCR的页面，然后将OCR嫁接到原始文档上。因此，它可以处理复杂的PDF，并且仍然尽可能保留其内容。OCRmyPDF使用开源OCR引擎Tesseract来执行OCR识别部分。

截止2019年12月15日，本项目共Watch 77次，Star 2.1k次，Fork 262次，共有issues48个，PR2个。