03 光学字符识别

场景描述

很多视频（如 Hulu 的海量娱乐视频）中都包含了大量的图像文字，比如新闻视频中的文本标题、体育视频中的比分牌、电影视频中内嵌的字幕等。这些图像文字携带了大量的文本信息，如果能将它们挖掘出来，就可以更好地理解视频及图像的内容，从而促进很多应用的发展，如图像 / 视频的搜索和推荐。挖掘图像中的文本信息，需要对图像中的文字进行检测和识别，这也称作光学字符识别（Optical Character Recognition，OCR）。光学字符识别的确切定义是，将包含键入、印刷或场景文本的电子图像转换成机器编码文本的过程。从二十世纪五十年代第一个识别英文字母的 OCR 产品面世以来，OCR 的领域逐步扩展到数字、符号和很多语言文字。如今，我们日常生活中见到的卡证识别、表单识别、增值税发票识别等都是基于 OCR 技术开发的。

OCR 算法通常分为两个基本模块，即文本检测（严格来说是物体检测的一个子类）和文本识别。

传统的文本检测主要依赖于一些浅层次的图像处理和分割方法，比如早期的基于二值化的连通区域提取，或者后期的基于最大极值稳定区域（Maximally Stable Extremal Regions，MSER）的字符区域提取等，以完成最终的文本定位。传统方法中使用的特征也主要以人工设计的特征为主，比较经典的有方向梯度直方图（Histogram of Oriented Gradient，HOG）。这些技术使传统 OCR 所处理的对象往往局限于成像清晰、背景干净、字体简单且排列规整的文档图像。

最初的文本识别算法是基于字符分割的。检测出文本行之后，需要将文本行分割成一个个独立的字符，之后对单独的字符进行识别，并利用隐马尔可夫模型（Hidden Markov Model，HMM）对最终的词语进行预测。这种方法思路相对简单，但识别准确率很大程度上受到字符分割效果的影响。

现在随着深度学习的兴起，越来越多的研究者采用神经网络来解决文本检测及文本识别的问题。研究问题的场景也从单纯的印刷文档图像、车牌等简单规则的文本形式拓展到了自然场景中千变万化的文本。本节介绍了一些利用深度神经网络实现自然场景下的文本检测及文本识别的方法。