文章编号:1003-0530(2017)04-0558-14

文本检测算法的发展与挑战

李翌昕 马尽文
(北京大学数学科学学院, 北京 100871)

摘 要: 对自然场景中的文字进行识别和理解是大量计算机视觉应用的基础。自然场景下的文本检测算法旨在检测出自然图像中的文字区域, 现已经成为计算机视觉和智能信息处理领域研究的一个热点。本文首先对文本检测算法的目标、技术路线及其所面对的挑战进行了分析与介绍。然后回顾了几种经典的文本检测算法, 并介绍了两种代表最新研究趋势的深度学习型文本检测算法。进一步, 本文阐述了几个主流的文本检测数据集并总结了一些代表性文本检测算法在这些数据集上的检测结果。最后, 本文讨论了文本检测的研究现状、面临的挑战和发展的趋势。

关键词: 文本检测; 文本定位; 文字识别; 计算机视觉; 机器学习
中图分类号: TP181 **文献标识码**: A **DOI**: 10.16798/j.issn.1003-0530.2017.04.016

The Developments and Challenges of Text Detection Algorithms

LI Yi-xin MA Jin-wen
(School of Mathematical Sciences, Peking University, Beijing, 100871, China)

Abstract: Recognition and understanding of the text from a natural scene is fundamental to a variety of practical applications in the fields of computer vision and intelligent information processing. The objective of text detection algorithms is to detect and localize text regions precisely in natural images. Therefore, text detection is a major part of recognizing and understanding the text from a natural scene, and has become a very popular topic in recent years. In this paper, we first introduce the objective, methods, and challenges of text detection. We then review some classic algorithms on text detection, and introduce two deep learning based algorithms that represent the trends of text detection research. Moreover, we summarize typical text detection datasets available, as well as the results of representative and leading algorithms on these datasets. Finally, we conclude the current researches on text detection and the challenges we face, and point out some prospective directions of text detection.

Key words: text detection; text localization; text recognition; computer vision; machine learning

1 引言

文字作为人类发明的抽象的交流符号,具有丰富的表达性,并在自然场景中作为信息表达大量出现。由于文字含有丰富的语义信息,识别自然场景中的文字成为大量视觉方面应用的基础,如目标定位[1]、人机交互[2]、图像搜索[3]、机器导航[4]和工业自动化[5]等等。因此,自然场景下对文字的识别和理解是近年的研究与应用的热点之一。

OCR[6-11](Optical Character Recognition,光学字符识别)被认为是文本识别的一个重要的突破且得到了广泛的应用。虽然市面上存在大量开源或付费的 OCR 软件系统可以在扫描文本上达到很高的识别准确率,但将其应用到自然图像上往往效果很差。这主要是因为 OCR 多为针对扫描文本设计,相对于扫描文本背景单一,文本整齐,而自然图像背景复杂,文字形式多样,因此 OCR 往往很难直接在自然图像上达到很好的识别效果。若将文本行从自然图像中切下,再送入识别系统,会有更好的识别效果。文本检测的任务便是在自然图像中对文字区域进行有效的定位,以提高后续识别系统的识别准确率,因此文本检测是在自然场景下对文字进行识别和理解的重要步骤。

文本检测最基本的任务则是判断出给定图像

收稿日期: 2016-12-20; 修回日期: 2017-02-07
基金项目: 国家自然科学基金(61171138)资助

中是否存在文字,若存在,则框出文字所在位置并返回。与文本检测十分相似的任务包括:文本定位[12],文字信息提取[13-15],自然环境下文本检测[16]。文本定位任务更加侧重返回文本框的位置,文字信息提取同时还包含定位与二值化任务,而自然环境下文本检测则将图像限定在户外拍摄的自然图像上。这些问题侧重点稍有不同,但在本质上所研究的问题是相同的。

对自然图像的文本检测研究从上个世纪 90 年代开始,而近年来随着计算机视觉领域的发展也逐步成为研究热点。许多国内外的人工智能巨头,例如微软[17-18],三星[19],百度[20],旷视[21]等等,在文本检测上都投入了大量的研究,涌现出许多新的思想与算法。在早年的研究中,图像中的文字被视为具有某种特定的纹理特征[22-39],或是某类特定的区域[17-18,40-43]。因此通常采用一些局部图像算子或者区域提取方法将图像中的文字的区域选出,这一大类的文本检测算法主要基于纹理特征与连通区域的提取。近些年来,由于卷积神经网络在计算机视觉领域的很多问题上有着突破性的成果[44-46],这类算法也被引入了文本检测中。文本检测问题可以看作一个特殊的物体检测问题[45,47-48],滑动窗口分类问题[49],或是语义分割问题[21,50],这一大类的文本检测算法主要基于深度学习框架实现。

本文的目的是通过对文本检测算法与通用数据集的分析与总结,以及对研究发展方向的分析,使读者了解文本检测的研究现状和所面对的挑战,能够对后续研究工作有所启发。本文的结构如下:第 2 部分介绍文本检测问题的背景;第 3 部分分别介绍基于连通区域和基于深度学习的两大类文本检测算法;第 4 部分介绍常用的文本检测数据集,以及近年一些文本检测算法在数据集上的效果对比;最后一部分对文本检测的研究现状与面临的挑战总结,并指出文本检测研究未来的发展方向。

2　背景

本节分为四部分介绍文本检测问题的背景:扫描文本与自然文本;文本检测相关的任务;文本检测的困难;以及文本检测算法。

2.1　扫描文本与自然文本

扫描图像一般是打印的文档通过扫描仪等设备获取。扫描图像的背景干净,扫描图像中的文本大多为打印文字,字体字号统一,排版规则。在对扫描文本进行简单的对准操作后,可以很容易的对其中的文字进行分割,获得很好的识别效果。绝大多数 OCR 系统,例如国内的汉王[6]和文通[7],国外的 ABBYY[8]和 IRIS[9],以及谷歌负责维护的开源 OCR 项目 Tesseract-OCR[10-11]等等,在扫描图像上都有非常高的识别准确率。

相比于扫描图像,自然图像由相机,手机或各种穿戴式设备在自然场景中获取。自然场景中的文本形式多样,在同一幅图像中,也有可能出现字体,大小,颜色,样式等完全不同的文字。自然场景的背景普遍较为复杂,使得其中的文字一般无法被 OCR 系统有限切割,也同时影响 OCR 系统的识别。另外由于拍照设备与环境的不同,自然图像可能会出现虚焦,模糊,变形,光照不均匀,文字被遮挡等情况,使得识别其中文字十分困难。

一个具体的例子如图 1 所示,左图为一幅自然图像。由于背景复杂干扰较多,OCR 系统无法将文字从图中切割出来,因此得到完全错误的识别结果。若先对图像中的文本进行初步的定位,如右图所示,将文本行框出逐条送入 OCR 系统,便去除了背景的干扰,可以得到较好的识别结果。

图 1　文本检测的作用和功能

Fig. 1　Function of text detection

2.2　文本检测相关任务

自然场景下的文字识别系统,从算法框架上可以分为分步系统与整合系统两类[51]。其中分步系统将自然场景下的文字识别问题分割为多个模块,包括文本检测[17,21,35,40,50],文字分割[12,52-54]与文字识别[5,10,41,55-57],通过一个前馈的过程完成对自然场景下文字的识别。相对应的,整合的文字识别系统也成为端到端[16,49,58-60](end-to-end)系统,以图像为输入,直接以识别为目标,通过检测与识别之间进行信息共享完成文字识别。从输入数据上可以分为图像与视频[14,27,29,34,37-38]。

ICDAR "Robust Reading Competition"[61-62]是自然场景下文字识别问题最为权威的比赛之一。该比赛将自然场景下的文字识别从任务上分为四个项目[20]:文本检测,文字分割,文字识别,以及端到端。以分别评价分步识别系统中各个模块以及端到端整合系统的性能。

在分步系统中,文本检测可以看作物体检测任

务的一个具体子任务,主要在自然场景中检测并定位文本的位置,并不对文本进行识别,本文也主要针对文本检测任务进行分析和探讨;文字分割任务针对文本检测得到的文本行,将文本行图像进行单字符分割,得到分割后的结果送给文字识别系统;文字识别使用机器学习算法对切割后的单字符进行识别,但近些年使用深度学习方法的文字识别系统,不再需要提前对文本进行切割,可以直接输入文本行图像得到识别结果。

在端到端的整合系统中,输入是原始自然图像,输出直接为图像中的文字内容。但一般的端到端系统并不是分步系统中三个任务的简单组合,在文本检测与文字识别两个步骤中往往共享信息或使用联合优化策略[51]。端到端的识别系统在通过检测结果得到识别结果的同时,往往还需要利用识别结果反馈指导检测结果,以得到更好的整体识别效果。或是用深度学习的方法训练一个端到端的神经网络,此时直接得到最后的识别结果,很难对中间的检测结果进行可视化的表示。

2.3 文本检测困难与挑战

自然场景下的文本检测是一项十分困难的任务,遇到的挑战主要可以分为以下三个方面[63]:文字的多样性,复杂的背景,诸多干扰因素。

相比起扫描文本中字体,字号,颜色,以及排版都较为统一的文字,自然场景下的文字更加复杂,甚至在同一场景下的文字都会出现不同的字体,颜色,方向等等。

另外,自然场景下的背景可以非常复杂,诸如围栏,草地,砖墙,标识等等物体都很难与文字区分开,文字也有可能被这些物体所遮挡,从而造成检测错误。

特别地,自然图像存在大量的干扰因素,主要由不同的拍摄条件所产生,例如图像分辨率较低,含有噪声,模糊,变形,光照不均匀等情况,这都是文本检测算法所要面对的问题。

2.4 文本检测算法

文本检测算法大致可以分为两类:基于纹理与连通区域的算法和基于深度学习的算法。

基于纹理与连通区域的算法,大多将文字看作具有某类特定的特征[22-39],或某类特定的区域[17,40-43]。因此这一大类算法一般利用文字的低阶局部特征,从自然图像中提取文字候选区域,并筛选融合为文本行候选,最终得到检测结果。

由于 CNN(Convolutional Neural Network,卷积神经网络)近些年来给计算机视觉领域带来了突破性的进展,近些年来也涌现了大量基于深度学习的文本检测算法。这类方法使用神经网络,直接得到检测结果[16],或是得到图像中文字显著图[21,49-50],再对显著图进行一系列后续处理得到文本检测结果。

在深层神经网络结构中,信息逐层前馈可以看作特征逐层提取的过程,高层神经元可以表示图像全局上更高阶的语义信息,因此神经网络具有很强的理解图像内容的能力。而基于纹理与连通区域的文本检测算法只使用了图像中低阶的局部特征,例如边缘纹理等信息。

3 两类基本的文本检测算法

在近 20 年中,人们针对不同类型的自然图像,甚至是视频,提出了大量的文本检测算法。这些算法大致可以分为两类:以基于筛选连通区域为代表的定位检测算法和以 CNN(Convolutional Neural Network,卷积神经网络)为代表的深度学习算法。以下分别介绍两类算法中较为经典的方法。

3.1 基于连通区域的文本检测算法

这类方法将图像中的文字看成某些特殊的区域或者具有某些特定的纹理特征。首先,我们可使用一些特征或方法在自然图像中提取候选区域作为文字的候选,这些特征包括颜色特征[22-25]、纹理特征[26-30]、边缘特征[31-34]、笔画宽度变换[35-39]、极值区域[17,40-43]等等。

经过筛选后滤去非字符的候选区域,将留下的区域视作字符并融合为文本行候选,再对文本行候选进行筛选得到最终的文本检测结果。其流程如图 2 所示。过滤筛选的方法可以通过人工设计特征,选取阈值进行筛选[35,58],或者使用统计模型或机器学习算法[36,41-43,59,64]对特征进行学习,自适应的对文字候选区域进行筛选。

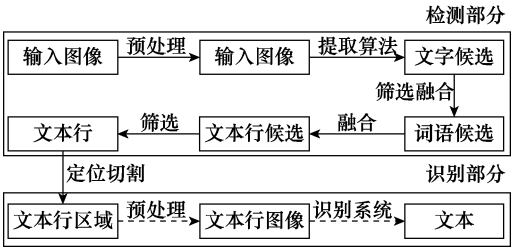


图 2 基于连通区域的文本检测算法流程

Fig. 2 Framework of component-based algorithm

基于连通区域的方法的关键在于提取文字候选的算法。本文选取介绍近几年较为常用或经典的两种提取方法:SWT[35](Stroke Width Transform,

笔画宽度变换)和 MSER[65](Maximally Stable Extremal Regions,最大稳定极值区域)。

SWT[35](Stroke Width Transform,笔画宽度变换)提取文字候选的方法是建立在一系列一般性的假设上:文字都是由笔画构成,而笔画有一定的宽度,同一行文本的笔画宽度应该较为接近,而非文字部分不是由笔画构成,因此没有笔画宽度。基于这种假设,我们可以对图像进行笔画宽度变换,计算输入图像中每一个像素点所在的笔画的宽度值,将连通区域作为文字候选。

笔画宽度具体的计算方法如图 3 所示:首先用 Canny 算子对图像进行边缘检测,结果如图 3 中第二幅图所示,白色像素为边缘点,黑色部分为非边缘点。对于每一个边缘点 P 有一个梯度方向 $\vec{d}_p$,沿 $-\vec{d}_p$ 方向进行搜索,可以找到边缘点 Q,则 $\boldsymbol{PQ}$ 之间的欧式距离 d 即为笔画宽度,此时可将 $\boldsymbol{PQ}$ 之间的所有像素点的笔画宽度值赋值为 d。

图 3　笔画宽度的计算

Fig. 3　Calculation of Stroke Width

一些特殊情况需要注意,例如经过某个像素点搜索线段不止一条的时候,该像素点的笔画宽度值取所有搜索线段的最小值。直角笔画拐角处的笔画宽度值并不为真实的笔画宽度值,此时可以取搜索线段上的平均值。当文字区域为亮底暗字(如白底黑字)时,沿 $-\vec{d}_p$ 方向进行搜索,但当文字区域为暗底亮字(如黑底白字)时,则需要沿 $\vec{d}_p$ 方向进行搜索。

以 SWT 为文字候选提取方法的一种文本检测算法流程如图 4 所示。四幅图从左至右分别为:原图,SWT 变换后结果(文字候选),筛选后结果(文本行候选),文本检测结果。可以看出原图中的文字并不显著,有许多与文字十分相似的线条,并且图像背景十分复杂,例如树叶部分边缘信息十分密集,对检测算法产生很大影响,但基于 SWT 的文本检测算法能成功地框出了文字区域,达到了很好的检测效果。

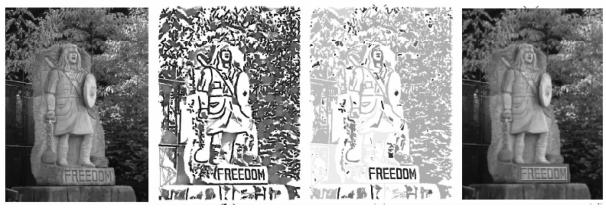


图 4　SWT 文本检测算法流程

Fig. 4　Framework of SWT text detection algorithm

基于 SWT 的文本检测算法[35]最早于 2010 年提出,并且在此基础上产生了许多类似的文本检测算法[36-39],由于其较好的检测结果流行一时。但此类算法建立在边缘检测上,当图像较为模糊时边缘信息较弱,或图像中高频区域密集,有较多复杂边缘的干扰,此时算法会受到较大影响。另外在筛选文字与文本行时,若使用手工特征会产生大量参数,基于不同类别的图像需要选取不同的参数与其相配,没有一组参数对所有类型的图像都适用,并且参数的调整没有理论性的指导,多为实验得到或经验选取,这也是基于连通区域的文本检测算法普遍存在的缺陷。

MSER[65](Maximally Stable Extremal Regions,最大稳定极值区域)方法所采用的 MSER 区域是那些在一系列灰度阈值范围内能保持形状和大小的区域。它们有着锐利的边缘,并且与背景有很强烈的灰度值对比。一般由于形态上的特性,文字都含有丰富的边缘信息,另外文字作为一种信息传递方式,为了让人能够看清,都与背景有较为强烈的颜色与灰度值对比,因此文字基本都为 MSER 区域,而此类方法便是通过提取 MSER 区域作为文字候选。

MSER 区域的提取首先需要对图像根据一系列阈值进行二值化,可以得到在不同阈值 t 下黑色(或白色)区域的面积,如图 5 所示。定义二值化图像某一区域 R 在二值化阈值 t 下的面积为 R_t,以黑色区域为例,随着二值化阈值 t 的增大,区域 R 的面积 R_t 也随之增大,此时可以定义变化率:

$$\nu(R_t) \triangleq \frac{|R_{t+\Delta} - R_{t-\Delta}|}{|R_t|} \tag{1}$$

给定阈值 t 的一个范围,当变化率 ν 在这个范围内取到极小值时,区域 R_t 称为最大稳定极值区域 MSER。同理可对图像进行反色处理得到白色区域的 MSER 区域。将 MSER 区域作为文字区域的候选,其后的流程与 SWT 算法类似,将文字候选融合为文本行候选进行筛选,得到最终的检测结果。

图 6 展示了一种基于 MSER 的文本检测算法流程。首先对原图提取 MSER 区域作为文字候选,得到

图 6 左数第一幅图。对文字候选区域进行筛选并融合,成为文本行候选,得到图 6 左数第二幅图。其后对文本行候选进行筛选,去除非文本行形状的区域,得到文本行区域为图 6 第三幅图。最后根据连通区域将三个文本行框出,得到最终的文本检测结果。

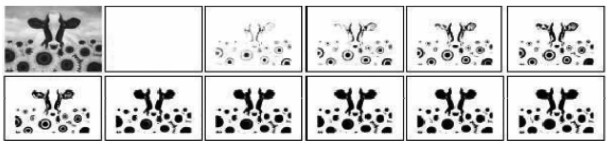


图 5 MSER 区域二值化

Fig. 5 Binarization of MSERs

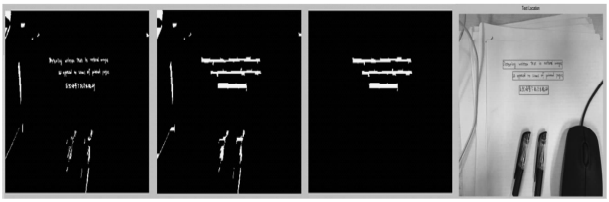


图 6 基于 MSER 的文本检测算法流程

Fig. 6 Framework of MSER text detection algorithm

MSER 方法最早于 2002 年作为一种仿射不变区域提取方法被提出[65],直到 2010 年才被引入到文本检测领域[58]。由于 MSER 具有良好的仿射不变性,可以检测出不同角度拍摄的文字,基于 MSER 的文本检测算法也成为主流,大量以此为基础发展而来的文本检测算法应运而生,例如直接使用树状结构的极值区域 ER[66],以及颜色增强极值区域 CER[17-18]等等。

直到 2014 年之前,基于 MSER 的文本检测算法一直占领着这个领域的领先地位,其中 USTB_STAR[43]算法甚至赢得了 ICDAR2013[61](International Conference on Document Analysis and Recognition)比赛的冠军。但此类方法仍然存在手工特征造成的参数过多情况,另外在最新的数据集中加入的新图像中文字并不是 MSER 区域,例如反光玻璃上面的文字,这也让基于 MSER 的文本检测算法效果大幅下降,最终被学习能力与泛化能力更强的深度学习打败。

3.2 基于深度学习的文本检测算法

近些年来,随着硬件成本的降低,人们可以训练结构更加复杂的多层神经网络,并在诸多领域(例如图像[44-46]、语音[67-68]、自然语言[69-70]等等)很轻易的击败传统的模型和算法,其中在一些任务上可以接近人类平均水平(例如图像分类[44],人脸识别[71],语音识别[67-68]等等),甚至可以击败人类最高水平(例如 AlphaGo[72]击败李世石九段),因此深度学习迅速成为了人工智能领域最热门的一大类方法。

CNN(Convolutional Neural Network)作为深度学习中的一种模型,在计算机视觉领域得到广泛的应用[44-46]。相比于传统方法直接处理像素点或提取图像局部的低层特征,CNN 的卷积层使用卷积窗口在图像上滑动,得到下一层特征,网络的浅层代表了图像低层次的一些边缘,轮廓等信息,而高层则提取出了图像的高阶语义信息,从而更好的完成许多视觉任务。近年来神经科学的研究发现,人类大脑对图像的处理是从具体到抽象的多层次综合处理方式(视网膜到视皮层,到额叶,再到海马体),这与 CNN 的结构十分相似。CNN 正是很好的模拟了人类大脑理解图像的方式,因此在计算机视觉的许多任务上有着接近人类水平的效果。图 7 是一个 CNN 网络的简单示意图,输入猫的图像,通过层次化的特征提取,最终得到猫的分类。

近年来,人们尝试采用 CNN 网络进行文本检测。在检测问题上,最为经典的网络结构是 R-CNN[45](Region-Convolutional Neural Network,区域卷积神经网络)以及在其基础上发展而来的一系列算法[47-48]。但由于文字形式特殊,文本检测问题不同于一般意义上的问题。若在文本检测任务上直接使用这类方法则检测效果一般,因此本文不介绍此类方法。

由于许多大型网络在图像分类任务上有很好的结果,人们首先开始尝试将这类分类网络直接用于对文字候选区域的筛选上[17-18,66,73-74]。这类方法一般基于 MSER 区域,在提取出的 MSER 区域上使用神经网络进行分类,或生成概率图对 MSER 区域筛选,因此这类算法并不基于深度学习。

另外的研究试图将分类的网络修改为用作检测的网络[16,21,49-50],两种具有代表性的方法分别是:text-spotting[49], FCN[46](Fully Convolutional Network,全卷积网络)。相比于 R-CNN 可以直接得到检测到的文本框,这两个方法由于是从分类网络修改得到,只能得到一幅代表文字区域概率的显著图(Saliency Map),对其进行二值化后可以使用类似基于连通区域的方法,通过筛选融合得到最终的文本检测结果。

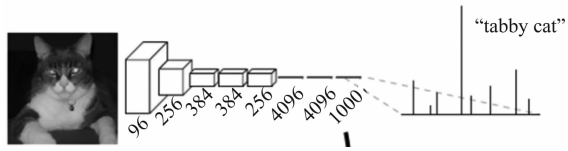


图 7 CNN 分类网络

Fig. 7 CNN classification network

Text-Spotting[49]根据文本检测的不同任务,首选训练一个小型的神经网络作为分类器。若只是进行文本检测,训练一个二元分类器(区分文本/非文本)即可。若想得到一个端到端的文本识别系统,则需要训练一个多类分类器(包含数字英文等字符)。按照滑动窗口的思想,得到的分类器在输入图像上进行滑动,每个窗口得到一个分类结果,可以理解为图像在该窗口中的内容为文字的概率。这样便可以得到一幅概率显著图,其中每个像素点的值可以代表该像素周围区域是文字的概率值。最后只需要用一个合适的阈值对概率图进行二值化,使用类似基于连通区域的筛选融合方法,即可得到最终的文本检测结果。

在具体的实现细节上,训练分类器的样本可以如文章中从网络上爬取[49],或者使用公开的字符数据集(例如 chars74k[75])。由于分类器训练好后,窗口的大小也固定,此时我们可以对图像进行多尺度的缩放,再经过分类器的滑动,这样相当于改变了窗口的大小,可以检测不同大小尺度的文字,最后对多尺度的概率显著图做一个统一的归一化即可。

图 8 展示了该算法的流程。从左至右分别对应输入的原图,生成的概率显著图(红色区域概率值高,蓝色区域概率值低),文字融合为文本行的结果,文本行区域二值化结果,以及文本检测结果。可以看到最右图文本检测的结果中,由于输入图像中复杂背景的干扰,返回了许多非文字区域,可以用手工特征进行一定筛选,或机器学习方法进行分类,滤去非文字的框,得到最终的检测结果。

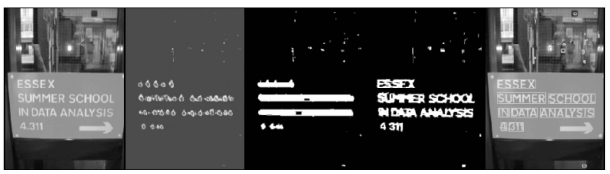


图 8　Text-Spotting 算法流程

Fig. 8　Framework of Text-Spotting

此方法于 2014 年提出[49],是在文本检测任务上使用 CNN 一个初步尝试,并且其文本检测结果达到了当时较为先进的水平。但该方法最为主要的问题是,滑动的窗口为了检测文字,尺寸很小(文献[49]中为 24 * 24),因此只利用了像素周围非常小的一个局部的特征。这造成,很多小局部有很像文字的线条,但实际并不是文字的干扰区域,也会被该方法当作文字返回。由于自然图像中的复杂背景,这类干扰区域一般较多,这使得该算法的检测效果存在一定的局限性。

FCN[46](Fully Convolutional Network)于 2015 年针对图像语义分割任务提出,在像素级别的图像分割问题上取得了突破性的成果。FCN 区别于传统的分类 CNN 网络,其输入与输出都是图像,而传统的分类 CNN 网络输出为一个类别或者一个一维特征。FCN 的结构如图 9 所示,CNN 网络通过卷积层提取下一层特征,通过池化层(pooling)后缩小特征尺寸,FCN 可以使用 CNN 网络中间层的特征,通过反卷积(deconvolution)与上采样(upsample)将特征恢复为原图大小。在这个结构上进行微调(fine-tuning),可以得到图像到图像的像素级别预测,输出图像的像素值可以看作输入图像中对应像素点属于某一类别的概率值。

文本检测任务可以看作一个特殊的像素级图像分割问题。如果可以通过 FCN 网络对图像中文字区域与非文字区域进行准确的分割,只需要将文字区域框出即可完全文本检测任务。目前使用 FCN 进行文本检测任务主要有两项工作[13-14],其想法均是将文本检测数据集制作为图像/像素级标签的图片对,输入 FCN 网络进行训练,得到的网络对图像进行像素级别分类,最后融合为文本行得到最终的文本检测结果。

使用 FCN 网络做文本检测任务主要的困难在于像素级别的图像标注。几乎所有文本检测数据集的标注都是文本框类型,即只标注文字在某一矩形区域内,但没有标注矩形区域内哪些像素点是文字。如果将矩形区域内所有的像素点都作为文字像素,训练得到的网络很难将多行文本分开,会识别为一整块文字区域。一种比较有效的解决方法[21],如图 10 所示,(a)为原图,(b)为原图中文字的文本框标签,将文本框中的文字进行切割得到如(c)图的切割结果,此时可以制作如(d-f)图的像素级标签,其中(d)是文本框标签,用来检测文本区域,(e)是文字标签,将(c)图的文字框缩小一定比例得到,用来检测文字位置,(f)图是文本行角度标签,用来检测文本行的倾斜角度。三个标签可以分别得到三个损失函数,对三个损失函数用合适的比例系数进行加权可以得到一个混合的损失函数,对这个新的损失函数进行训练得到的网络,可以同时检测文本行区域,文字位置,以及文本行倾斜的角度,可以得到目前领先的文本检测效果。

FCN 网络使用了不同层次的特征,同时考虑了输入图像的局部与全局,因此有很好的文本检测效果。另外由于神经网络高度并行的计算特点,FCN 结构虽然

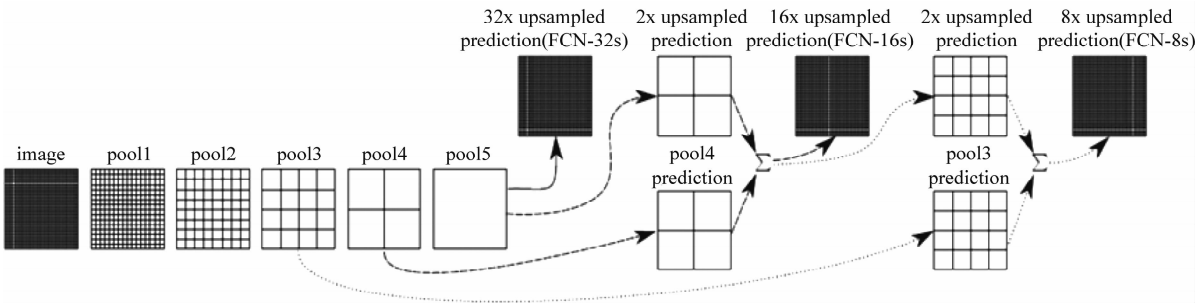


图 9 FCN 结构示意图

Fig. 9 Main structure of FCN

相比传统方法复杂,但在 GPU 上可以做到实时的检测(size384 * 384, 18fps on GTX1080, 30fps on GTX1080 * 2),传统方法大多只能在 CPU 上运行,普遍需要秒级的处理时间,无法做到完全实时。另外深度学习方法大多依赖于训练数据的质量,随着文本检测数据集的完善,此类方法效果还有不小的提高空间。

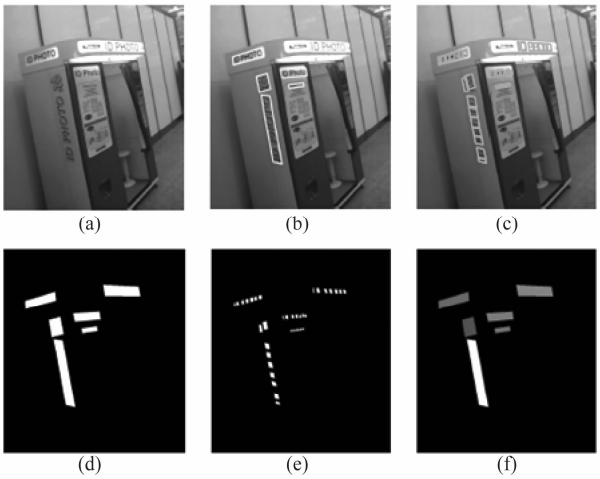


图 10 多目标像素级标注

Fig. 10 Multi-target pixel level label

4 数据集

一般的文本检测传统方法,若人工设计特征与参数选取,可以做到完全无监督的形式,即不需要额外的数据去训练模型。若参数由机器学习方法得到,则需要在训练数据上进行学习。对于基于深度学习的文本检测算法,一般是一个有监督的形式,神经网络需要在大量标注好的训练数据上进行学习,而一旦网络结构确定,模型所能达到的效果完全由训练数据的数量和质量决定。另外,各种文本检测算法的性能需要进行量化,才能进行比较,因此需要有一个统一的评价指标,这也正是公开数据集的作用。

本文首先介绍一个文本检测上常用的数据集 ICDAR,简单总结了这个数据集的内容与特点。随后会对文本检测模型的评价标准进行简单介绍,最后总结几种经典算法和现阶段领先的算法在这 ICDAR 数据集上的效果。

4.1 公开数据集

公开数据集中一般将数据分为两部分,训练集和测试集。其中训练集如上文所述,是为模型提供训练数据,因此训练集中的样本都进行了完全的标注信息,提供了完整的正确答案。测试集是为评价各种算法性能提供了一种方法,训练时不允许使用测试集中的样本,训练好的模型在测试集上运行的效果作为评价算法性能的统一指标。测试集可以不提供正确答案,此时需要将算法在测试集上返回的结果上传至数据集的网站,以在线的方式返回评价指标,这种方式更为公平。若测试集提供正确答案,便可自己根据算法在测试集上的结果离线的算出评价指标,但若存在不诚信的行为,例如将测试集中的数据加入训练集中,则最终的评价指标将缺乏可信度。有时在训练集中会分割出一部分数据作为验证集,验证集的作用是实时监测模型的泛化能力,防止出现过拟合现象。

ICDAR[20](International Conference on Document Analysis and Recognition)会议从 2003 开始每两年举办一次,其中的“Robust Reading”比赛是文本检测上最为权威的交流平台。ICDAR 从 2003 年开始为文本检测比赛提供公开数据集,并在 2005 与 2013 年对数据集进行了两次更新,在 2015 年推出了全新的一批数据集,本文主要介绍 ICDAR2013 与 ICDAR2015 数据集。

ICDAR2015 包含四个部分[20]。第一部分称为“Born-Digital”,是一批从网页上抓取数据,如图 11 左图所示,这一部分的图像更偏向于扫描文本而不是自然图像。第二部分称为“Focused Scene Text”,这批图片沿用了 ICDAR2013 中的所有图片,完全在自然场景下拍摄,专门针对文字拍摄,文字都在图像中较

为显眼的位置,如图 11 中图所示。第三部分称为“Text in Videos”,顾名思义是提供了一些视频以及其中文字的标注。第四部分称为“Incidental Scene Text”,是 ICDAR2015 全新加入的一批数据,这部分图片由 Google Glass 在自然场景下无先验知识的随意拍摄,并没有专门针对文字,如图 11 右图所示。

由于第一部分图片不属于自然图像范畴,第三部分为视频,因此一般文本检测任务选取第二与第四部分的数据。其中第二部分完全继承自 ICDAR2013,有 229 张训练图与 233 张测试图片,其中文本行均为水平,采用文本框形式进行了标注。第四部分为 ICDAR2015 新加入内容,包括 1000 张训练图与 500 张测试图,其中包括多角度的文本行,采用一个平行四边形的框对文本进行了标注。由于第四部分由穿戴设备随意采集,并没有对准文字,文字不一定在图像中心位置,并且可能会出现模糊遮挡等情况,因此这部分图像最为接近真实的自然图像,其检测难度也最大。

图 11　ICDAR2015 数据集

Fig. 11　ICDAR2015 datasets

由于 ICDAR 对图像中所有的文本都进行了完整的标注,数据质量很高,因此 ICDAR 也成为了文本检测上最为常用的数据集。

MSCOCO[76](Microsoft Common Object in Context)是一个大型的图像数据集,包含 82783 张训练图,40504 张验证图,与 2014 年 40775 张测试图和 2015 年的 81432 张测试图。标注信息为像素级别的物体,其中包括共 11 大类 91 小类的生活中常见物体。因此 MSCOCO 为物体检测,图像分类,语义分割等等问题提供了一个通用的大型数据库,以及一个统一的评价平台。

COCO-Text[77]是康奈尔大学在 MSCOCO 数据集基础上提供了文本信息的标注,从而为文本检测以及文字识别等领域的研究者提供了一个大型的图像数据集。COCO-Text 的每一条标注信息对应 MSCOCO 图像中的一条文本,但不是像素级别,只提供矩形框形式的标注。另外在提供矩形框参数的同时,COCO-Text 中的每一条标注还包含一个多属性的标签,包含文本内容,是手写体还是打印体,是否清晰,方便研究者根据需求对数据进行清洗。COCO-Text 一共标注了 63686 张图片中的 173589 条文本,分为 43686 张训练图像和 20000 张验证图像。COCO-Text 的标注如图 12 所示[77]。

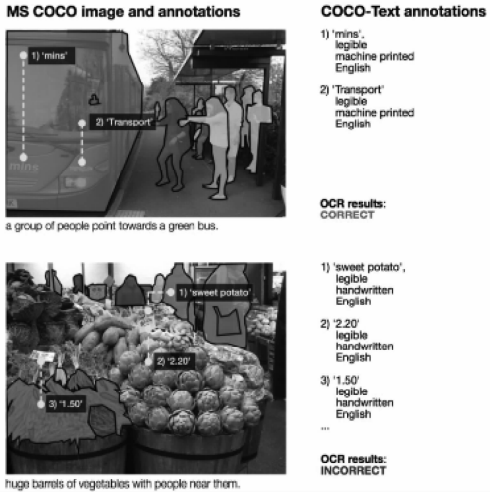


图 12　COCO-Text 数据集

Fig. 12　COCO-Text datasets

4.2　评价标准

文本检测作为一类检测问题,同样使用准确率,召回率,以及 F 值作为算法的评价指标。假设算法检测出的结果中,正类(文字)被正确检测出的个数为 *TP*,负类(非文字)被检测出正类的个数为 *FP*,正类被算法检测出为负类(非文字)个数为 *TN*,则准确率可以定义为 $TP/(TP+FP)$,召回率可以定义为 $TP/(TP+TN)$。准确率可以理解为算法返回的结果中有多少比例真正是文字,召回率可以理解为有多少比例的文字被算法检测出来。

由于对于图像中文本行的标注是文本框形式,算法返回的检测结果也是文本框形式,一般情况下两个文本框很难完全重合,因此 ICDAR 使用的评价标准中[72],需要根据返回的文本框与正确的文本框的重合部分比例来判断该文本框是否被算法返回,并且需要对一对多与多对一情况进行一定惩罚[72]。其准确率与召回率的定义如下:

$$\text{precision}(G,D,t_r,t_p)=\frac{\sum_j \text{Match}_D(D_j,G,t_r,t_p)}{|D|} \tag{2}$$

$$\text{recall}(G,D,t_r,t_p)=\frac{\sum_j \text{Match}_G(G_j,D,t_r,t_p)}{|G|} \tag{3}$$

其中 G 是正确的文本框,D 是算法检测出的文本框,t_r

与 t_p 是两个和召回率与准确率有关的参数,一般取 0.8 和 0.4,Match 函数当完全匹配时取值为 1,当完全不匹配是为 0,当出现一对多或多对一情况时惩罚分值,值在 0-1 之间。具体的评价细节可以参看 ICDAR2013 的官方文档[61],与 ICDAR 使用的评价方法[72]。

一般来说,对于一个算法准确率与召回率是一对此消彼长的评价指标。例如将算法的参数调整为只返回置信度较高的区域时,算法的准确率会提高,但同时召回率会降低,反之亦然。因此如果只用一个指标评价算法性能,一般使用 F 值,即准确率与召回率的调和平均值:

$$F = \frac{1}{\frac{\alpha}{precision} + \frac{1-\alpha}{reall}} \tag{4}$$

一般情况下 α 取 0.5,即为:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{5}$$

只有准确率与召回率都在较高水平时,F 值才会较高,我们才认为算法的性能较好。

4.3 算法结果

ICDAR 每两年举办一次的文本检测比赛[20]是文本检测问题最为权威的比赛,在比赛结束后依然可以提交结果,因此 ICDAR 也成为了文本检测与识别算法的主流的评价标准之一,其排行榜上的算法可以代表目前文本检测研究的最高水平。本文主要关注 ICDAR2013[61]和 ICDAR2015[62]比赛,总结了一些经典的算法,以及在 ICDAR 排行榜[20]上目前领先的算法(截止至 2016.12),在 ICDAR 上的评价指标如下表所示(表中 Algorithm 代表算法名称,Method 为算法核心部分所使用的方法,Y 代表算法发表的年份,R 代表算法召回率,P 代表算法准确率,以及 F 表示算法的 F 值)。

表 1 ICDAR2013 数据集上算法比较

Tab. 1 Comparison of algorithms on ICDAR2013

Algorithm	Method	Y	R	P	F
Baidu IDL v2 *	CNN	2016	0.85	**0.93**	**0.88**
MS_CER[17]	CER	2014	0.86	0.87	0.86
Megvii[21]	FCN	2016	**0.89**	0.80	0.84
HUST[50]	FCN	2016	0.88	0.78	0.83
MSER_RTree	MSER	2015	0.74	0.85	0.79
SWT_best	SWT	2015	0.73	0.81	0.77
USTB_STAR[43]	MSER	2013	0.66	0.88	0.76
Text-Spotter[49]	CNN	2014	0.65	0.88	0.74
ABBYY[8]	OCR	2012	0.35	0.61	0.45

(* 文章在准备中)

表 2 ICDAR2015 数据集上算法比较

Tab. 2 Comparison of algorithms on ICDAR2015

Algorithm	Method	Y	R	P	F
Baidu IDL v2 *	CNN	2016	0.73	**0.77**	**0.75**
HUST_OText *	CNN	2016	**0.74**	0.69	0.71
RRPN-2 *	RPN	2016	0.73	0.69	0.71
Megvii[21]	FCN	2016	0.72	0.59	0.65
HUST[50]	FCN	2016	0.71	0.43	0.54
Stradvision-2 *	ER	2016	0.37	0.77	0.50

(* 文章在准备中)

ICDAR2013 数据集的标准线由通用 OCR 软件 ABBYY[8]提供,F 值只有 0.45,这说明通用的 OCR 软件在自然图像上无法得到理想的文本检测结果。SWT 的基本算法[35]在 2010 年提出,在 ICDAR2013 与 ICDAR2015 数据集上没有结果,目前可以找到基于 SWT 最好的算法在 ICDAR2013 上 F 值为 0.77。目前不使用深度学习框架,效果最好的算法是微软亚洲研究院基于 MSER 提出的颜色增强极值区域 CER[17],但仍然使用一个小型的神经网络对极值区域进行筛选,在 ICDAR2013 上的 F 值是 0.86。值得一提的是,2013 年 ICDAR 比赛的冠军 USTB_STAR[43],依然使用的是基于 MSER 的文本检测算法,但经过三年的发展,USTB_STAR 目前只能排在 ICDAR2013 排行榜第 23 的位置。在 ICDAR2013 排行榜[20]上目前效果最好的算法是百度深度学习研究院的 Baidu IDL v2 *,但目前没有针对这个算法所发表学术论文。

由于 ICDAR2013 只提供了 233 张测试图,并且提供了测试图的正确标注信息,因此很容易通过训练或调整算法参数使得算法在这样小规模的数据集上达到很好的效果,但此时算法很可能已经在测试集上过拟合,并不具有和测试结果同样好的泛化能力。ICDAR2015 拥有更多的测试图像(500),并且由于其无先验随意拍摄得到,其图像更接近真实的自然图像,难度更大,另外没有提供测试图像的正确标注信息,因此我们认为 ICDAR2015 对于近两年的文本检测算法是更好的评价标准。

早于 2015 年的方法一般没有在 ICDAR2015 上进行实验,绝大多数算法未提供代码,因此我们无法测试这些算法在 ICDAR2015 数据集上的效果。从表 2 中可以看出,ICDAR2015 数据集明显难度更大。在 ICDAR2015 排行榜[20]上,几乎所有算法都使用了深度学习框架,这可能是由于 ICDAR2015 的

第 4 个挑战十分接近自然图像,并且使用了全新的数据,导致传统方法普遍效果较差。目前使用非深度学习框架下效果最好的算法是 stradvision-2 *,基于 ER 区域,F 值只有 0.50。ICDAR2015 排行榜上大量新算法的文章仍在准备中,还未发表。

从表中结果可以看出,未使用深度学习的传统文本检测算法,除了 MS_CER[17],在 ICDAR2013 上的 F 值很难超过 0.8,而近些年基于深度学习的算法明显有着更好的效果,F 值达到接近 0.9 的水平。这说明 ICDAR2013 数据集已经基本被解决,因此全新的并且难度更大的 ICDAR2015 第 4 个挑战被提出,目前算法在这个数据集上的 F 值最高为 Baidu IDL v2 *,只有 0.75,因此算法效果在这个数据集上还有很大的提升空间。

由于 ICDAR 比赛中只需要提交算法效果,绝大部分算法在文章中并没有对算法执行的效率进行分析,并且没有提供可重现实验的代码,因此本文无法对各算法的速度进行详细定量的对比。

基于纹理连通区域的传统文本检测算法,需要先按照一定规则提取文字候选区域,这一步往往是比较耗时的。以 MSER 为例,在 MatLab 环境下,使用 Vlfeat[79]对一幅 640 * 480 的图像提取 MSER 区域,耗时为 2 s 左右(3.0 Hz CPU * 2)。并且在后续对文字候选进行筛选融合的步骤中,在文字候选较杂的情况下,若进行连通性分析则较为耗时,处理时间仍为秒级。相较而言,基于深度学习的文本检测算法在训练时一般较为耗时,但使用训练好的模型进行检测时,只需对图片进行一次前馈过程,配合 GPU 硬件加速,即使是十分复杂的网络结构这个过程一般也是实时的。以 FCN[46]为例,一幅 640 * 480 的图像经过 FCN-VGG-16 得到概率显著图,耗时仅为 0.09 s(GTX1080 GPU * 2)。经典的检测算法 Faster R-CNN[48]在同样使用大型的 VGG-16 网络的条件下,对同样尺寸级别图像的检测效率为5 fps,即完成整个检测步骤仅需要 0.2 s,这相比基于 MS-ER 的算法秒级的处理速度提高了一个数量级。

从 ICDAR2015 排行榜[20]中前几名的算法无一例外使用了深度学习框架来看,目前基于深度学习的算法是文本检测任务上最为流行,同时也是效果最好的一类方法。另外由于深度学习算法的高度并行性,可以使用 GPU 进行加速的特性,这使得基于深度学习的文本检测算法,相比传统方法在速度上有更大的优势,这对于相当一部分的实际应用是至关重要的。因此可以预见在今后的几年中,文本检测上的研究必定是在更难的更真实的 ICDAR2015 数据集上,围绕着深度学习框架来进行。

5　总结与展望

自然场景下对文字的识别和理解是许多计算机视觉任务的基础,而一般 OCR 系统对自然图像中文字的识别效果很差。如果对自然图像中的文本先进行定位,再送入 OCR 系统往往会有更好的识别效果,因此文本检测是识别自然图像中文字很重要的步骤。自然场景下的文字区别于扫描文档中的文字,文字形式更加多样,背景更加复杂,另外由于拍照条件不佳会引起图像质量较差的情况,这也是文本检测任务所面对的主要困难[63]。

文本检测的算法大致可以分为两类:基于纹理与连通区域的文本检测算法和基于深度学习的文本检测算法。基于连通区域的算法,主要通过一些特征提取方法从图像中提取文字候选区域,将文字候选进行筛选融合为文本行,得到最终的文本检测结果。基于深度学习的算法,主要通过神经网络对输入图像中的像素点进行预测,得到一幅可以表示文本区域的概率显著图,框出其中概率较高部分即可完成文本检测任务。

基于纹理或连通区域的文本检测算法,往往由人为设定的规则提取文字候选区域,再由人工设计的特征对这些区域进行筛选。此类算法一般将文本看作具有某种纹理或某类区域,因此大多比较直观,当规则与特征设计合理时也会有较好的检测效果,例如 USTB_STAR[43],MS_CER[17]等等。但此类算法最大的问题在于人为设计的规则往往无法适应文字的多样性,例如基于 SWT 的算法对于边缘信息较少例如模糊的图像效果很差,基于 MSER 的算法无法检测出不是 MSER 区域的文字。另外人工设计的特征会给算法带来大量的参数,这些参数往往需要根据某类特定的图像进行调节,无法适应所有类型的图像,这导致算法的鲁棒性较差,可能无法区分与文字及其相似的一些背景区域,在一些场景更加复杂的自然图像上检测效果往往不理想[62]。

基于深度学习的文本检测算法使用一个神经网络对图像中的文字区域进行预测。而大型的神经网

络由于参数规模极大,具有十分强大的学习和泛化能力,因此模型有足够的能力适应文字的多样性。并且这些参数由算法在大量数据上学习得到,无需人为设定,使得深度学习模型有很好的鲁棒性。基于深度学习的文本检测算法近些年在公开数据集上取得了不错的效果[20,61-62],例如 Megvii[21], HUST[50], Baidu IDL v2 * 等等。但此类算法所面临的最大问题是数据。现有的文本检测数据集,由于需要人工标注,因此标注质量较高的数据集一般数据量较小,例如 ICDAR2013[61],ICDAR2015[62],在这类数据上训练神经网络,极易出现过拟合的情况;而数据量较大的数据集往往标注质量一般,例如 COCO-Text[77],在这类数据上训练,网络往往无法学到文字的本质,模型的泛化能力不强。

以 CNN 为代表的深度学习算法由于很好的模拟了人类大脑理解图像的方式,在计算机视觉领域的大多数问题上都能有很好的效果。深度学习算法引入文本检测问题后,也迅速的打败了发展多年的传统方法,取得了目前领先的检测结果[61-62]。因此可以预见,在文本检测任务上,基于深度学习的算法必将成为主流的方法,对训练数据与网络结构的研究也必将成为热点。而更加真实也更加困难的 ICDAR2015 的第 4 个挑战[62],也将成为研究者突破的重点。我们认为有如下的几个角度值得进行更深入的探索:

数据集。深度学习网络的训练依赖于大量的训练数据。规模更大,形式更加丰富的训练数据可以使得网络学习到文字更本质的特征,从而使得深度学习模型拥有更强的泛化能力。而现有的文本检测数据集的数据规模与标注质量有限,如何利用数据扩增的方法,例如旋转缩放等等,将标注质量较高但数据量较小的数据集扩容,或如何利用增强标注信息,例如多目标训练等等,将数据规模大但标注质量较差的数据集增强,将是提高基于深度学习的文本检测算法效果的关键。

深度学习网络结构。目前文本检测问题上使用的深度学习框架,大多是从其他问题上直接将网络移植而来。例如直接使用传统物体检测问题的网络结构[45],将文字直接当作一类特殊的物体进行训练;及将文本检测问题看作语义分割问题[46],使用 FCN 网络对文字进行预测[21,50]等等。这类网络一般被设计为解决通用问题,而文本相较于其他物体类别,其形式更加复杂丰富,这些网络的结构、特征表示、以及输出形式并不完全适合文本检测问题。因此,如何针对文本检测进行网络结构的设计,值得进行更加深入的研究。

后续处理。在各个公开的文本检测数据集上[20,77],计算文本检测算法的性能指标,需要得到文本框形式的返回结果。因此对于得到特征图[21,49-50]或文字候选区域[17,35,43]的文本检测算法,往往需要进行筛选融合等后续处理才能得到最终文本框级别的检测结果。使用人工特征进行筛选的方法[35]鲁棒性较差,因此完善后续处理过程对算法效果的提升有重要的意义。未来的研究中,可以加入例如文本框回归[47-48]等优化过程,提高输出文本框的准确率。而由深度学习实现的文本框回归网络,可接在文本检测网络结构之后,进行整体的微调训练,从而得到一个由输入图像经过文字概率图直接得到文本框位置的端到端网络,由此得到更好的文本检测结果。

参考文献

[1] Barber D B, Redding J D, Mclain T W, et al. Vision-based Target Geo-location using a Fixed-wing Miniature Air Vehicle[J]. Journal of Intelligent & Robotic Systems, 2006, 47(4):361-382.

[2] Kisačanin B, Pavlovic V, Huang T S. Real-Time Vision for Human-Computer Interaction[J]. Human Computer Interaction in the Work Place, 2010:224-229.

[3] Tsai S S, Chen H, Chen D, et al. Mobile visual search on printed documents using text and low bit-rate features [C]// IEEE International Conference on Image Processing,2011:2601-2604.

[4] Desouza G N, Kak A C. Vision for Mobile Robot Navigation: A Survey[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(2):237-267.

[5] Park R H. Recognition of raised characters for automatic classification of rubber tires[J]. Optical Engineering, 1995, 34(1):102-109.

[6] Website of Hanwang OCR products[EB/OL]. http://ka.hanwang.com.cn/sbjs/ocrjs/, 2016.12.

[7] Website of Hanwang OCR products[EB/OL]. http://www.wintone.com.cn/a/prods/ocr/, 2016.12.

[8] Heliński M. Report on the comparison of Tesseract and ABBYY FineReader OCR engines[J]. 2012.

[9] Jantz R. IRIS serves up snappy, accurate OCR[J]. Pc

World, 2006.
[10] Smith R. An Overview of the Tesseract OCR Engine[J]. 2007, 2:629-633.
[11] Address of open source Tesseract-OCR[EB/OL]. https://github.com/tesseract-ocr/tesseract, 2016.12.
[12] Lienhart R, Wernicke A. Localizing and segmenting text in images and videos[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2002, 12(4):256-268.
[13] Jung K, Kim K I, Jain A K. Text information extraction in images and video: a survey[J]. Pattern Recognition, 2004, 37(5):977-997.
[14] Zhang J, Kasturi R. Extraction of Text Objects in Video Documents: Recent Progress[C]//The Eighth Iapr International Workshop on Document Analysis Systems. IEEE Computer Society, 2008:5-17.
[15] Zhang J, Kasturi R. Extraction of Text Objects in Video Documents: Recent Progress[C]//The Eighth Iapr International Workshop on Document Analysis Systems. IEEE Computer Society, 2008:5-17.
[16] Jaderberg M, Simonyan K, Vedaldi A, et al. Reading Text in the Wild with Convolutional Neural Networks[J]. International Journal of Computer Vision, 2016, 116(1):1-20.
[17] Sun L, Huo Q, Jia W, et al. Robust Text Detection in Natural Scene Images by Generalized Color-Enhanced Contrasting Extremal Region and Neural Networks[C]// International Conference on Pattern Recognition. IEEE, 2014:2715-2720.
[18] Sun L, Huo Q, Jia W, et al. A robust approach for text detection from natural scene images[J]. Pattern Recognition, 2015, 48(9):2906-2920.
[19] Park J M, Chung H, Seong Y K. Scene text detection suitable for parallelizing on multi-core[C]//IEEE International Conference on Image Processing. IEEE Press, 2009:2425-2428.
[20] Leaderboard and benchmark of ICDAR robust reading competition[EB/OL]. http://rrc.cvc.uab.es/?com=introduction, 2016.12.
[21] Yao C, Bai X, Sang N, et al. Scene Text Detection via Holistic, Multi-Channel Prediction [J]. arXiv: 1606.09002v2, 2016.
[22] Jain A K, Yu B. Automatic text location in images and video frames[J]. Pattern Recognition, 1998, 31(12): 2055-2076.
[23] Wang K, Kangas J A. Character location in scene images from digital camera[J]. Pattern Recognition, 2003, 36(10):2287-2299.
[24] Mancas-Thillou C, Gosselin B. Spatial and Color Spaces Combination for Natural Scene Text Extraction[C]// IEEE International Conference on Image Processing. IEEE, 2006:985-988.
[25] Mancas-Thillou C, Gosselin B. Color text extraction with selective metric-based clustering[J]. Computer Vision & Image Understanding, 2007, 107(1-2):97-107.
[26] Yi C, Tian Y L. Localizing Text in Scene Images by Boundary Clustering, Stroke Segmentation, and String Fragment Classification[J]. IEEE Transactions on Image Processing, 2012, 21(9):4256-68.
[27] Shivakumara P, Phan T Q, Tan C L. New Fourier-Statistical Features in RGB Space for Video Text Detection [J]. IEEE Transactions on Circuits & Systems for Video Technology, 2010, 20(11):1520-1532.
[28] Zhong Y, Zhang H, Jain A K. Automatic caption localization in compressed video[J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 1999, 22(4): 385-392.
[29] Li H, Doermann D, Kia O. Automatic text detection and tracking in digital video[J]. IEEE Transactions on Image Processing, 2000, 9(1):147-156.
[30] Pan Y F, Liu C L, Hou X. Fast scene text localization by learning-based filtering and verification[C]//International Conference on Image Processing, ICIP 2010, September 26-29, Hong Kong, China. DBLP, 2010:2269-2272.
[31] Shivakumara P, Huang W, Tan C L. Efficient video text detection using edge features[C]//International Conference on Pattern Recognition, IEEE, 2008.
[32] Huang R, Shivakumara P, Uchida S. Scene Character Detection by an Edge-Ray Filter[C]//International Conference on Document Analysis and Recognition. IEEE Computer Society, 2013:462-466.
[33] Li M, Wang C. An adaptive text detection approach in images and video frames[C]//IEEE International Joint Conference on Neural Networks. IEEE, 2008:72-77.
[34] Shivakumara P, Phan T Q, Tan C L. A Gradient Difference Based Technique for Video Text Detection[C]//International Conference on Document Analysis and Recognition. IEEE Computer Society, 2009:156-160.
[35] Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform[C]//Computer Vision and Pattern Recognition. IEEE, 2010:2963-2970.
[36] Chowdhury A R, Bhattacharya U, Parui S K. Scene text detection using sparse stroke information and MLP[J]. 2012:294-297.
[37] Bai B, Yin F, Liu C L. A Fast Stroke-Based Method for

Text Detection in Video[C]//Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, IEEE, 2012:69-73.

[38] Liu X, Wang W. Robustly Extracting Captions in Videos Based on Stroke-Like Edges and Spatio-Temporal Analysis [J]. IEEE Transactions on Multimedia, 2012, 14(2): 482-489.

[39] Mosleh A, Bouguila N, Hamza A B. Image text detection using a bandlet-based edge detector and stroke width transform[C]//British Machine Vision Conference, 2012.

[40] Chen H, Tsai S S, Schroth G, et al. Robust text detection in natural images with edge-enhanced Maximally Stable Extremal Regions[C]//International Conference on Image Processing, 2011:2609-2612.

[41] Neumann L. Real-time scene text localization and recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2012:3538-3545.

[42] Koo H I, Kim D H. Scene Text Detection via Connected Component Clustering and Nontext Filtering[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2013, 22(6):2296-305.

[43] Yin X C, Yin X, Huang K, et al. Robust Text Detection in Natural Scene Images[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 36(5): 970-83.

[44] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems, 2012, 25(2):2012.

[45] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. Computer Science, 2014:580-587.

[46] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2015: 3431-3440.

[47] Girshick R. Fast R-CNN[C]//IEEE International Conference on Computer Vision. IEEE, 2015:1440-1448.

[48] 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 & Machine Intelligence, 2015:1-1.

[49] Jaderberg M, Vedaldi A, Zisserman A. Deep Features for Text Spotting[M]. Computer Vision-ECCV 2014. Springer International Publishing, 2014:512-528.

[50] Zhang Z, Zhang C, Shen W, et al. Multi-Oriented Text Detection with Fully Convolutional Networks[J]. arXiv: 1604.04018v2, 2016.

[51] Ye Q, Doermann D. Text Detection and Recognition in Imagery: A Survey[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(7):1480-1500.

[52] Song Y, Liu A, Pang L, et al. A Novel Image Text Extraction Method Based on K-Means Clustering[C]//Ieee/acis International Conference on Computer and Information Science. IEEE, 2008:185-190.

[53] Nomura S, Yamanaka K, Katai O, et al. A novel adaptive morphological approach for degraded character image segmentation[J]. Pattern Recognition, 2005, 38(11): 1961-1975.

[54] Roy P P, Pal U, Llados J, et al. Multi-Oriented and Multi-Sized Touching Character Segmentation Using Dynamic Programming[C]//International Conference on Document Analysis & Recognition. IEEE, 2009:11-15.

[55] Sawaki M, Murase H, Hagita N. Automatic acquisition of context-based images templates for degraded character recognition in scene images[J]. British Journal of General Practice the Journal of the Royal College of General Practitioners, 2000, 4(517):615-9.

[56] Campos T E D, Babu B R, Varma M. Character Recognition in Natural Images[C]//Visapp 2009-Proceedings of the Fourth International Conference on Computer Vision Theory and Applications, Lisboa, Portugal, February. 2009:273-280.

[57] Jawahar C V. Top-down and bottom-up cues for scene text recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2012:2687-2694.

[58] Neumann L, Matas J. A Method for Text Localization and Recognition in Real-World Images[C]//Computer Vision-ACCV 2010-, Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers. 2010:770-783.

[59] Wang K, Babenko B, Belongie S. End-to-end scene text recognition[J]. Proceedings, 2011, 24(4):1457-1464.

[60] Bissacco A, Cummins M, Netzer Y, et al. PhotoOCR: Reading Text in Uncontrolled Conditions[C]//IEEE International Conference on Computer Vision, 2013:785-792.

[61] Karatzas D, Shafait F, Uchida S, et al. ICDAR 2013 Robust Reading Competition[C]//International Conference on Document Analysis and Recognition. IEEE, 2013:1484-1493.

[62] Karatzas D, Lu S, Shafait F, et al. ICDAR 2015 competition on Robust Reading[C]//International Conference on Document Analysis and Recognition, 2015.

[63] Zhu Y, Yao C, Bai X. Scene text detection and recognition: recent advances and future trends[J]. Frontiers of Computer Science, 2016, 10(1):19-36.

[64] Pan Y F, Hou X, Liu C L. Text Localization in Natural Scene Images Based on Conditional Random Field[C]// International Conference on Document Analysis and Recognition. IEEE, 2009:6-10.

[65] Matas J, Chum O, Urban M, et al. Robust wide-baseline stereo from maximally stable extremal regions[J]. Image & Vision Computing, 2004, 22(10):761-767

[66] Huang W, Qiao Y, Tang X. Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees [C]// Computer Vision-eccv, 2014:497-511.

[67] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013:6645-6649.

[68] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.

[69] Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning[J]. Journal of Parallel and Distributed Computing, 2008:160-167.

[70] Manning C D, Surdeanu M, Bauer J, et al. The Stanford CoreNLP Natural Language Processing Toolkit [C] // Meeting of the Association for Computational Linguistics: System Demonstrations, 2014.

[71] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering[C]//Computer Vision and Pattern Recognition. IEEE, 2015:815-823.

[72] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search [J]. Nature, 2016, 529(7587):484-489.

[73] Ye J, Huang L L, Hao X. Neural Network Based Text Detection in Videos Using Local Binary Patterns[C] // Pattern Recognition, 2009. CCPR 2009. Chinese Conference on, 2009:1-5.

[74] Ren X, Chen K, Sun J. A Novel Scene Text Detection Algorithm Based on Convolutional Neural Network[J]. 2016.

[75] Chars74k Dataset Web Site[EB/OL]. http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k.

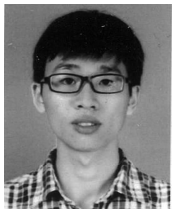
[76] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[M]. Computer Vision-ECCV 2014. Springer International Publishing, 2014:740-755.

[77] Veit A, Matera T, Neumann L, et al. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images[J]. arXiv:1601.07140, 2016.

[78] Wolf C, Jolion J M. Object count/area graphs for the evaluation of object detection and segmentation algorithms [J]. International Journal on Document Analysis and Recognition (IJDAR), 2006, 8(4):280-296.

[79] Vedaldi A, Fulkerson B. Vlfeat: an open and portable library of computer vision algorithms[C] // International Conference on Multimedea 2010, Firenze, Italy, October. DBLP, 2010:1469-1472.

作者简介



李翌昕　男,1992 年生,北京人。2014 年 7 月毕业于中国科学技术大学数学科学学院,获得理学学士学位。现为北京大学数学科学学院信息科学系博士研究生,主要研究方向为计算机视觉、图像处理和模式识别。

E-mail: liyixin@pku.edu.cn

马尽文　男,1962 年生,陕西人。1992 年毕业于南开大学数学系,获理学博士学位。现为北京大学数学科学学院信息科学系主任、教授、博士生导师。中国电子学会信号处理分会常务委员,中国工业与应用数学学会理事,主要从事智能信息处理、神经计算、模式识别、生物信息学等方面的研究。

E-mail:jwma@math.pku.edu.cn