摘要

文本作为人类最具影响力的发明之一，在日常生活中发挥了重要作用。文本中包含的丰富而准确的信息在计算机视觉的应用领域显得极为重要，自然场景的文本检测也成为了当前研究的热点话题。尽管这一领域存在着诸多的挑战，如噪声，模糊，扭曲等，但是在最近几年文本检测仍然取得了巨大的进展。本篇综述主要包含三大内容：（1）总结最近几年场景文本检测的主要方法和思路（2）介绍当前最新的算法和模型（3）预测将来潜在的研究方向。本文还介绍了文本检测领域对模型性能的评估标准，旨在向该领域的研究者提供大致的参考。

关键字：文本检测，自然场景，算法，数据库，评估标准

Introduce

作为一种人类抽象出来的信息载体，自然场景中的文本蕴含着高级语义，这一属性使图像和视频中的文本成为了一项重要的信息来源。嵌入在文本中的丰富信息在计算机视觉的应用领域发挥着重要作用，如图像检索[1]，目标定位[2]，人机交互[3]，机器人导航[4]和工业自动化[5]。因此，自动文本检测，提供了在图像和视频中访问和利用文本信息的方法，已经成为计算机视觉和文档分析中的积极研究课题。然而，对文本检测和定位仍然是一项艰难的任务，场景文本检测面临着的挑战大致分为以下三个方面：

1. 场景文本的分散性。相较于文档文本具有规则的字体，单一的颜色，一致的尺寸和排版，自然场景下的文本的字体，颜色和方向存在较大差别，即使在同一场景下这样的差别仍然存在。
2. 背景的复杂性。自然场景下的图片通常是非常复杂的，存在大量和文本相似的物体（窗户，砖块和栅栏等），因此容易造成混淆和错误。
3. 其他干扰因素。如噪声，扭曲，模糊，低分辨率和不均匀光照等都会增加文本检测和定位的难度和风险。

。。。

场景文本识别的相关工作

在过去几年里，研究人员已经提出了大量的算法和模型来检测场景图像和视频下的文本。这些方法主要分为以下几类：基于纹理的方法，基于连通体的方法和混合方法。

基于纹理的方法

这类方法把文本视为一种特殊的纹理并充分利用文本的纹理特性（如局部强度，滤波器响应，小波系数）来区分文本和非文本区域。由于要扫描图像所有的位置和尺寸，这类方法通常计算量很大，而且只能处理水平文本，对图像旋转和尺寸改变比较敏感。

在早期的工作中，Zhong et al. [35] 提出了一种在彩色图像中定位文本的方法，首先使用水平方向的空间不变性粗略的定位文本，然后在文本区域中使用颜色分割来寻找文本。Li et al. [38] 提出了一种在视频中检测和定位文本的文本检测系统，在这个系统中，使用小波系数的均值分解图像，把排序第一和第二的分量当作局部特征。Kim et al. [36] 直接使用原生像素值作为局部特征训练了一个SVM分类器来分类像素点，并通过自适应MeanShift算法寻找文本区域。这种方法对背景简单的图像或视频会表现出绝佳的性能，但是很难将其泛化到背景更加复杂的图像或视频中。

为了处理多种语言的文本（主要是中文和英文），Lyu et al. [40] 提出了一种由粗到细的多尺度搜索方案，这个方案需要利用文本强烈边沿和高对比度的特性来区分文本和非文本区域，而且该方法提供了一种局部自适应二值化的策略来分割检测到的文本区域。但是该方法同样引进了大量的经验规则和参数，因此很难处理不同分辨率或不同类型文本的图像和视频。Zhong et al. [41] 提出了一种完全不同的方法能够直接在离散余弦变换（DCT）域中检测文本，由于在检测前不需要对图像进行解码，所以该算法实时性很强同时精度受到限制。为了加速文本检测流程Chen *et al.* [8] 提出了一种快速文本检测器，该检测器是级联的Adaboost分类器，算法首先会提取平均强度，像素值方差，水平差分，垂直差分和梯度直方图等特征，然后根据特征集合训练弱分类器。该方法的检测效率明显高于其他算法，但是检测精度也受到限制。

最近，Wang *et al.* [46] 提出了在自然场景图像中对特定单词进行定位的方法。首先算法会使用滑动窗口得到单个字符，然后根据字符间的结构关系对字符所有的可能结合进行打分，最后在给定的单词列表中，选择出与其最相似的结合关系作为最终的输出结果。与其他检测结果不同的是，该算法只能检测到给定列表中的单词，不能检测到不在列表中的单词。在现实生活中，对每幅图像不可能给出所有单词的列表，这很大程度减小了该算法的应用范围。

基于连通域的方法

这类方法首先会通过一系列算法（颜色聚类，笔画宽度变换，最大稳定极值区域等）来提取连通区域，然后通过手工设计的规则或训练分类器的方式滤除非文本的连通区域，最后通过聚类等算法将连通区域链成单词或文本行区域，并训练分类器进一步过滤非文本区域。一般来说，这类方法只需要处理少量的连通区域而不用遍历图像的每个位置，从而更加高效。而且对旋转变化，颜色变化，尺度变化和字体变化鲁棒性强。在最近几年里，基于连通域的方法已经成为自然场景文本检测领域的主流算法。

Jain *et al.* [47] 提出的方法首先通过颜色聚类的方式将图像分解成多个互不重叠的连通区域，然后根据连通域分析将这些区域组合成文本行并通过几何规则滤除非文本的连通区域。由于使用了大量人工定义的规则和参数，该方法在复杂的场景图像中表现并不出众。利用字符的笔画宽度接近常数的特性，Epshtein *et al.* [9] 提出了一种新的图像算子：笔画宽度变换（SWT），如图Fig. 1，该算子提供了一种简单的方式从梯度图像中恢复字符的笔画宽度并能在复杂的场景图像中高效的提取不同尺度和方向的文本区域。但该方法使用了大量人工定义的规则和参数，而且仅考虑了水平文本。Neumann *et al.* [10] 提出了一种基于最大稳定极值区域（MSER）的文本检测算法。如图Fig. 2，该算法从原始图像中提取最大稳定极值区域作为字符的候选区域，并使用分类器剔除无效的候选区域，最后通过一系列的链接规则将剩余的字符候选区域聚合成文本行区域。然而这些链接规则仅适用于水平或接近水平的文本，对有较大倾斜角度的文本表现较差。SWT和MSER算法是自然场景文本检测的代表性方法，被广泛应用于后续的工作中。

稀疏表示在人脸识别和图像诊断领域的巨大成功为大量研究人员提供了灵感。Zhao *et al.* [52] 从训练样本集合中构建稀疏字典并使用它来判断图像中的特定区域是否包含文本。然而学习到的稀疏字典泛化能力有限，不能处理旋转和尺度变化。Yi *et al.* [28] 提出了一种能检测倾斜文本的方法。首先根据像素点在颜色空间中的分布将图像分成若干区域，然后根据颜色相似性，空间距离和区域的相对尺寸等特征将上述区域结合成连通域，最后通过先验规则将非文本区域滤除。然而该算法前提是图像包含多种特定的颜色，这在复杂的场景图像中是不一定符合的，而且该方法使用了大量人工定义的规则和参数，因此很难泛化到大规模的场景图像集合中。

Yao *et al.* [12] 提出了一种基于SWT且能检测任意方向文本的算法。该算法使用了两级分类方案和两组旋转不变和尺度不变的特征，用于提取字符区域的本质特征。Huang *et al.* [29] 提出了一种基于SWT的新图像算子：笔画特征变换（SFT）。为了解决传统SWT中边沿点的错误匹配问题，SFT引入了边沿点的颜色一致性和约束关系，并取得了更好的区域提取效果。SFT的检测性能在标准数据集上远远高于其他检测方法，但是仅适用于水平文本。在[30]中, Huang *et al* 将最大稳定极值区域（MSER）和卷积神经网络（CNN）结合起来，提出了一种新的文本检测算法。MSER算子用于提取文本候选区域，CNN用于正确识别文本区域并将将字符从候选区域中分离出来。

混合方法

混合方法是指结合了基于纹理和基于连通域的方法。在Liu et al. [54]的算法中，使用精细的边沿检测策略提取所有可能的文本区域的边沿像素，并计算区域轮廓的梯度和几何特性用于验证生成文本候选区域，然后对区域进行纹理分析，将非文本区域从候选区域中剔除。Pan *et al.* [27] 提出的混合方法从多尺度的概率图中提取文本候选区域。算法首先会根据一组由预定义的模式计算到的特征集合（HOG特征等）训练一个分类器，并有分类器估计出图像对应的概率图。然后结合一元的连通属性和二值得上下文关系，使用条件随机场模型（CRF）区分文本和非文本区域。和上述几种算法一样，该方法也只适用于水平文本。

评价准则

在这一节，我们对ICDAR2011和ICDAR2013数据集的算法评估标准进行介绍，在ICDAR2003的评估标准中不能处理一对多（一个Grounding Truth矩形被多个Detection 矩形匹配）和多对一的情况，但这些情况在实际中是经常出现的，如果不考虑这些情况，会过低地估计了算法的性能。ICDAR2013评估标准中对精确度和召回率的定义如下：

G和D分别代表着Grounding Truth矩形集合和Detection 矩形集合，tr属于【0，1】是对召回率的阈值，而tp【0，1】是对准确率的阈值。Tp和tr的常用值时0.4和0.8.

为了判断一对一，多对一和一对多的情况，评估算法使用了精确度矩阵和召回率矩阵，当。。

根据以上的判断规则定义Detection 矩形的匹配函数matchD和Grounding Truth矩形的匹配函数matchG：

Fsc（K）时对出现多对一和一对多情况的惩罚函数，一般设为常值0.8。