视频和图像文本提取方法综述

蒋梦迪[1] 程江华[1] 陈明辉[2] 库锡树[1]
(国防科学技术大学电子科学与工程学院 长沙 410073)[1] (火箭军驻长沙地区军事代表室 长沙 410073)[2]

摘 要 文本提取在视频和图像中具有重要的应用价值。近年来,大数据时代带来了海量信息检索的迫切需求,大量视频和图像中文本的提取方法涌现出来。回顾了视频和图像中文本提取的算法,从文本提取流程出发,将其分为文本区域检测定位和文本分割两大步骤。在每个步骤中,分析并比较了现有算法的使用范围及相对优缺点,讨论了图像公用数据库,列举了近些年来图像中文本提取的重要应用,指出了当前研究中存在的问题,展望了视频和场景图像文本提取方法的发展趋势。

关键词 视频和图像,文本提取,文本区域检测与定位,文本分割,综述
中图法分类号 TP391.4 **文献标识码** A

Text Extraction in Video and Images: A Review

JIANG Meng-di[1] CHENG Jiang-hua[1] CHEN Ming-hui[2] KU Xi-shu[1]
(College of Electronics Science and Engineering, National University of Defense Technology, Changsha 410073, China)[1]
(Rocket Force in Changsha Area Military Representative Department, Changsha 410073, China)[2]

Abstract Text extraction in video and images has important application value. Big data era brought urgent demands of huge amounts of information retrieval, many text extraction methods have been proposed in recent years. In this paper, we reviewed text extraction methods from video and images. First, we classified the course of text extraction into two steps: text region detection and localization, text segmentation. Then, some text region detection and localization and text segmentation algorithms have been discussed regarding their application fields and their advantages and disadvantages. Finally, we discussed benchmark data and performance evaluation, and pointed out the promising directions for future research.

Keywords Video and images, Text extraction, Text region detection and localization, Text segmentation, Review

1 引言

近年来,视频和图像中的文本检测和识别问题受到越来越多的关注,图像中的文本信息是理解整个图像的重要内容。基于内容的图像索引,是指基于内容给图像贴上标签的过程。图像内容可以分为两个方面:感知内容和语义内容[1]。感知内容包括颜色、强度、形状、纹理和它们的时空变化等诸多属性;语义内容是指物体、事件和它们的关系。针对一系列的视频和图像中低水平感知内容的研究运用见报道,图像中的语义内容(如文本、人脸、车辆、手势等)也引起广泛关注。其中,文本信息吸引了特别的兴趣,这是因为:1)文本对描述图像内容非常有用;2)文本相对于其他的语义内容更容易提取;3)文本提取在基于关键词的图像索引、自动视频记录和安全监控等方面有重要应用;4)光学字符识别(Optical Character Recognition, OCR)软件更为成熟。近些年来,视频和图像中的文本提取应用十分广泛,列举如下。

1)嵌入式应用软件:Google Goggles[2]是一个图像识别软件,它可以将图片翻译成文本信息。Watanabe's[3]翻译相机可以检测自然场景中的文本,并对文本进行识别后将日语翻译成英语。卡内基梅隆大学开发了一个基于PDA的标识识别器[4],可在IOS和Android平台上使用,能立即识别文本并将其翻译成另一种语言[5]。

2)实时车牌识别:通过对道路上的交通监控视频流进行实时处理,提取车牌号码,可实现可疑车辆的跟踪搜索,形成道路监控智能化网络系统,能够更好地满足治安、刑侦、交通管理等业务需求。

3)互联网视频内容安全监控:互联网视频的多样化和复杂性以及内容的质量会对观众的思想和行为产生重要的影响,提取并分析视频图像中的文本,可以有效地理解视频内容包含的语义,从而实现对视频内容的安全监控。

4)基于文本的视频图像检索:图像中的文本信息不仅可以反映内容信息,还可以为其所在的内容片断提供索引和内容标记。文本信息支持基于关键字的检索。提取视频图像中的文本,并对相应的视频进行自动标注,再利用成熟的传统文本检索技术对视频建立索引并进行分类,可实现基于关键字的视频检索。

5)工业自动化:包裹、集装箱、房屋和地图上的文本识别在工业自动化方面有广泛应用。例如,识别信封上的地址可

蒋梦迪(1994-),女,硕士生,主要研究方向为计算机视觉与智能信息处理,E-mail:jiangmengdi11@163.com;程江华(1979-),男,副教授,主要研究方向为视频图像处理及模式识别;陈明辉(1980-),男,工程师,主要研究方向为武器系统研制生产监管;库锡树(1963-),男,教授,主要研究方向为电路与系统。

应用于邮件分类系统；集装箱号码的自动识别提高了物流效率[6]；自动识别房屋号码和地图中的文字有利于地理编码系统的构建[7]。

然而，视频和图像中的文本提取面临着诸多挑战。

1）图像中的文本通常具有多尺寸、多字体、多颜色、多语言和低对比度的特点。

2）背景复杂：自然场景中存在许多与文本结构和外观相似的物体，如建筑物、符号和树叶等。图像中的文本通常嵌入在复杂的背景中，这使得检测提取变得更困难。

3）照度不均匀：由于照明和感知装备不均匀等原因，自然场景图像中经常出现照度不均匀的现象。照度不均匀可导致颜色失真和视觉特征恶化，从而引入错误的检测、分割、识别结果。

4）图像退化：视频图像中的字符的分辨率通常较低，字符质量没有达到利用常规的 OCR 系统进行处理的要求。流行的有损压缩方法，如 MPEG，JPEG 等使得视频图像质量更低。

5）失真：当相机光轴不垂直于文字平面时，会导致视角扭曲，图像中的文本在方向、对准方面存在差异。字符扭曲和非矩形的文本边界框会显著影响文本提取的性能。

上述各类问题的存在，使得视频和图像中的文本提取极具挑战性。近年来，研究人员皓首穷经，探求解决之道，但至今无一套通用的解决方案。尽管前人已对图像中的文本提取进行过综述，但已有综述主要是针对图像中的文本检测这一步骤进行的，尚未出现针对图像中文本提取的整个流程的综述。本文将近年来的视频图像文本提取方法以及相关技术进行系统总结，并结合笔者在该领域的研究实践对这些方法进行了分析和比较，指出当前算法存在的问题，展望该领域的发展趋势。

1.1　视频和图像中的文本

大量视频和图像的文本提取算法已被用于特定的领域，如视频安全监控、实时车牌识别和基于内容的图像/视频索引等。尽管前人已经进行了大量的研究，但是设计一个通用的文本提取系统并不容易。这是因为视频和图像往往具有复杂背景和低对比度，图像中的文本在字体、尺寸、样式、颜色、方向和对齐等方面存在大量变化源，这些变化使得文本的提取变得极其困难。图 1—图 3 给出了一些图像中文本的例子。

页面布局分析通常需要处理文档图像，通过扫描书籍、CD 封面或其他彩色文本得到的图像与文档图像类似（见图 1），这些图像中的文本并不能直接运用于传统的文本图像分析处理。视频图像中的文本可进一步分为人为覆盖在图像上的字幕文本（见图 2）和存在于自然图像中的场景文本（见图 3）。与字幕文本相比，场景文本在方向上可能出现扭曲。此外，场景文本通常受相机的影响，如照度、焦距、相对运动等，因此场景文本更难检测。在进一步研究之前，定义常用的术语和识别常见的文本特征十分重要。自然场景图像中的文本可以在以下属性上表现出许多变化。

1）尺寸：字体大小变化的范围可能不同[8]。

2）对准：自然场景常常在多个方向对准，并且有几何失真[9]。

3）颜色：字符通常具有同样或相似的颜色。这个特性使得基于连通分量的文本检测成为可能[10]。

4）边缘：为使自然场景中的文本容易阅读，文本和背景边界通常具有强烈的边缘[11-12]。

5）失真：由于照相机的角度问题，一些图像中的文本会出现角度扭曲，这会显著影响提取性能[13]。

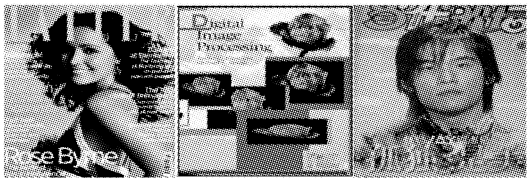


图 1　多颜色文本图像：每一行文本可能有不同颜色

（a）字幕直接覆盖在背景上　（b）字幕方向竖直　（c）字幕与背景有较好的对比度

图 2　字幕文本

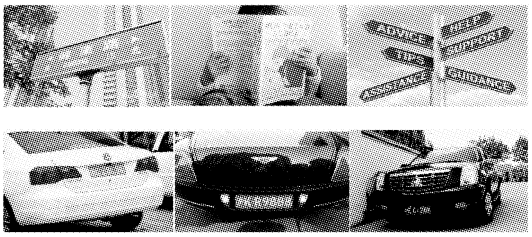


图 3　场景图像：在倾斜角度、视角、模糊、照明和对齐上有变化

1.2　文本提取流程

一个文本提取系统的输入为静态图像或序列图像。图像可以是灰度图像或彩色图像，也可以是压缩或未压缩图像，图像中的文本可以是静止或滚动的。文本提取通常分为以下几个步骤：1）文本检测和定位；2）文本分割和增强；3）文本识别（OCR）。图 4 为其流程图。

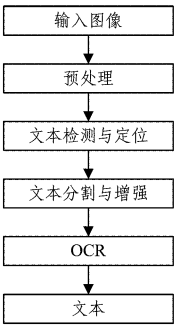


图 4　图像中的文本提取流程图

在大部分文献中，文本检测、定位、分割和增强通常交替使用。然而，在本文中，我们将对这些术语进行区分。文本检测是指在给定的帧中（通常文本检测用于序列图像）测定文本的存在。文本定位是指确定图像中文本的位置并且生成文本边界框的过程。虽然图像中文本的精确位置可以用边界框表示，但是仍然需要将文本从背景中分割出来进行识别，这意味着在输入 OCR 之前，文本图像需要转化为二进制图像并且进行图像增强。文本分割阶段是将文本从背景中分割出来，并提取出字符块精确的轮廓。由于文本区域通常具有低分辨率的特点，并且容易产生噪音，因此需要进行文本图像增强。此后，可以使用 OCR 技术将提取的文本图像转换成纯文本。

1.3 论文结构与安排

上述各类问题的存在，使得视频和图像中的文本提取极具挑战性。本文将近年来的视频和图像中的提取方法以及相关技术进行了系统总结。在 2.1 节，我们详细地回顾了视频图像中文本检测与定位的方法；2.2 节讨论了文本分割与增强。第 3 节讨论了算法的性能评估，并对公共测试数据库进行了回顾。最后讨论了现有视频图像中文本提取方法中存在的问题，并展望了该领域的发展趋势。

2 文本提取的关键技术

2.1 文本区域检测与定位

现存的文本区域检测的方法可以粗略地分为 6 类算法：基于边缘、基于纹理、基于连通分量(Connected Component，CC)、基于笔画、基于深度学习和其他算法。

2.1.1 基于边缘的算法

边缘是文本检测的可靠特征。对于边缘检测，通常首先采用边缘检测器(Canny 和 Laplacian)进行检测，进而确定有高边缘密度和强度的区域，随后用形态学操作从背景中提取文本并剔除非文本区域。Ye 等[14]提出了一种在自然场景图像中提取文本的方法。这种算法基于彩色图像滤波技术，首先获取边缘，随后分析字符间的固有特性。Liu，Samarabandu[11]和 Ou 等[15]运用多尺度边缘检测器检测边缘，多尺度边缘检测器由边缘强度、密度和方向方差形成。这种算法可以自动检测和提取复杂图像中的文本，在字体、尺寸、颜色、方向和对齐存在变化时性能强健，因此有大量的实际应用，如移动机器人导航、车牌检测与识别、目标鉴定、文献检索等。Lyu 等[16]提出一种带有强度分量的改进边缘图来进行文本检测，利用由粗到精的映射聚集检测文本区域，并利用局部阈值法和内部填充提取文本字符串。C. Liu 等[17]考虑到水平、垂直、右上、左上 4 个方向的笔画，在每个方向都生成了一个边缘图，并结合统计学特征，利用 K 均值聚类将图像像素划分为背景和文本候选区域。Kim 等[18]认为文本向背景过渡时亮度和色饱和度数值呈现指数或对数函数的变化，利用这一特点生成颜色转换图并得到候选文本框，然后使用局部二元模型(Local Binary Pattern，LBP)修正结果。Cho 等[19]提出一种基于 Canny 算子的场景文本检测算法，算法考虑到图像边缘与文本之间的相似性，使用双阈值和滞后跟踪检测文本。

基于边缘的算法在背景复杂度不高的情况下比较有效，然而在阴影和照度的影响下，提取好的边缘轮廓非常困难。

2.1.2 基于纹理的算法

基于纹理的方法将文本区域视为一种特殊的纹理，利用图像纹理特征判定像素点或像素块是否属于文本。由于字符通常由多个笔画组成，而存在笔画的区域通常也是整个图像纹理较丰富的区域，因此通过对纹理丰富区域的搜寻即可实现对字符区域的定位。纹理分析的方法通常采用高斯滤波、小波分解、傅里叶变换、离散余弦变换(Discrete Cosine Transform，DCT)和 LBP 等方法来提取纹理特征。典型的算法是在一个特定的区域提取纹理特征，并采用一个分类器(通过机器学习或启发式的方法进行训练)来鉴定区域内是否含有文本。Li 和 Doermann[20]首次提出了基于小波纹理特征的文本定位算法，该算法利用小波系数的平均、二阶、三阶中心矩和神经网络对图像窗口进行分类，并滤除负窗口和孤立的正窗口，连通的正窗口作为文本区域被保留。Zhou 等[21]提出一种多语言的文本检测方法，该方法可以有效地检测出自然场景中各类语种的文本区域。根据书写笔画的规则，该算法选择了 3 种不同的纹理特性来描述多语言文本：梯度方向直方图(Histogram of Gradient，HOG)、平均梯度(Mean of gradients，MG)和 LBP；然后运用一个级联分类器联合这 3 个纹理特性检测定位文本区域。Bertini 等[22]提出基于角点检测的算法，通过视频帧之间角点的相似性来检测文本区域。Sato 等[23]利用一种基于垂直、水平、左对角线、右对角线这 4 个方向的内插滤波器来进行文本检测。Zhong 等[24]在 JPEG/MPEG 压缩域中提出一种基于 DCT 特征的文本定位算法，算法将检测到的水平空间强度变化大的图像块作为文本区域，并通过形态学操作聚集这些区域，最后利用频谱能量阈值进行验证。Goto 等[25]利用 DCT 特征和 Fisher 判别分析来定位场景图像。Kim 等[37]提出一种基于 SVM 和纹理模板的算法进行文本定位，被分类为正值的像素通过均值移位算法合并成文本区域。Pan 等[26]提出一种新的快速文本区域检测定位算法，该算法使用了基于学习的区域滤波和基于由粗到精的验证算法。不同于仅仅使用基于学习的分类器进行滤波和分类，该算法选择有区别的特性，分别采用一个增强的分类器和一个多项式分类器进行粗区域滤波和细区域的验证。在验证阶段，作者评估了 5 个被广泛使用的特性：HOG，LBP，DCT，Gabor 和小波。Wu 等[27]提出了一种基于 K-means 的算法来识别文本像素，该方法将文本看成一种特殊的纹理，并在 3 个不同尺度上使用 9 个二阶高斯导数来寻找可能的文本区域。Zhao 等[28]对图像进行小波变换，边缘检测后采用滑动窗口将图像分成小块，将一种新的稀疏表示模型用于纹理分割和特征提取，再利用学习型判别字典对候选文本区域进行修正。Li 等[29]基于 Harris 角点对文本进行检测，生成角点响应图，利用基于块的阈值法得到候选文本区域，进行连通区域分析后用投影法得到文本行。

由于文本区域相比于非文本区域有特殊的纹理特性，这些算法在复杂背景下可以准确地检测定位文本区域。然而，算法的运行速度相对较慢，且对文本的对齐和方向敏感。

2.1.3 基于连通区域的算法

基于连通区域的算法采用自底向上的结构，将图像中的小区域合并成连续的较大区域，直到图像中所有区域被识别。在后期阶段，通常需要进行几何分析来识别文本区域，并聚集这些文本区域来定位文本。基于连通区域的方法通过边缘检测或颜色聚类直接分割候选文本区域。非文本区域通过启发式规则或分类器进行修剪。Zhang 等[30]运用条件随机场(Conditional Random Field，CRF)给连通区域贴上“文本”和“非文本”的标签。Pan 等[31]也运用了 CRF 模型，在文章中提出一种两步迭代的 CRF 算法，即置信度推理阶段和 OCR 滤波阶段。第一个 CRF 迭代旨在找出确定的文本连通区域，并将不确定的连通区域送入第二个迭代；第二个迭代通过 OCR 判定不确定连通区域，并过滤虚警连通区域。同上述两个方法类似，Wang 等[32]提出一种基于连通区域的由粗到精的算法来检测定位场景图像中的文本。算法将彩色图像分隔成均匀的颜色层，利用块邻接图(Block Adjacency Graph，BAG)分析颜色层中的每个连通区域块。在粗定位阶段，提出一种调整与分析的方案来定位所有颜色层中可能的文本区域。基于

区域的方法通常假设文本区域的像素都有相同的颜色，根据字符像素颜色的一致性和字符颜色与背景存在较大的对比度等特征对图像进行分割。Agnihotri 等[33]采用字符红色的特性来获得文本和背景间的高对比边缘。Hua 等[34]通过检测高对比度视频帧中的“统一颜色”块来检测定位文本区域。Kim 等[35]提出一种基于 RGB 空间中欧氏距离的颜色聚类方法，并利用 64 个聚集颜色通道进行文本检测。由于编码压缩会导致图像退化，并且背景与文本通常具有低对比度，图像中文本很少由相同的颜色构成，某些情况下字符和背景有相似的颜色，颜色并不是一个稳定的特征，此时字符和背景目标很难被区分，因此这种方法的鲁棒性较差，不适用于处理复杂背景下的文本区域检测。一个新的趋势是用统计模型[36,38]来实现连通区域算法，例如在成对空间特征上使用 AdaBoost 分类器来学习连通区域算法模型，统计模型的使用显著提高了连通区域算法的适应性。

分割的候选文本区域的数量相对较少，基于连通区域的算法具有计算复杂度低的优点，并且定位出的文本区域可直接进行识别。然而，基于连通区域的算法需要事先知道文本位置和尺寸等先验知识，这在实际应用中通常是无法满足的。另外，由于在进行分析比较时背景中的非文本连通区域很容易与文本区域混淆，因此设计一个快速且可靠的连通区域分类器十分困难。

2.1.4　基于笔画特征的算法

作为文本字符串的基本元素，笔画为自然场景中的文本提供了强健的检测功能。文本可以看作是由各方向笔画元素结合构成的模型，通过笔画元素的组合与分布能够提取文本的特征。一个区分文本与场景中其他元素的特征是其近乎恒定的笔画宽度，这个特征可以用来检测包含文本的区域。Jung C 等人[42]提出笔画滤波器（Stroke filter，SF）的概念。文章认为，文本的边缘（梯度）特征、连通区域、纹理、投影等为文本的外部特征，而笔画为文本的内在特征，因此笔画不仅适用于所有语言，还能检测手写体文本。文章定义了类笔画特征，如图 5 所示。

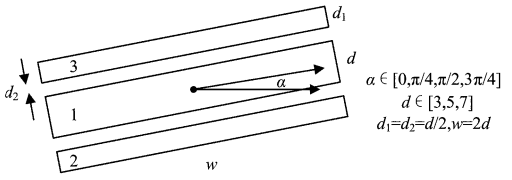


图 5　类笔画特征[42]

Gui 等[43]提出了一种基于笔画滤波器的笔画响应图，即分别提取细笔画和粗笔画来定义不同的区域，最后用基于支持向量机 SVM 的方向梯度直方图 HOG 分类器消除噪声并检测结果。Epthtein[44]进一步探究了这种笔画特征，根据道路检测和血管检测的方法提出了笔画宽度变换 SWT（Stroke Width Transform）。这种图像文本定位算法对文本的大小、方向、颜色、字体和语种不敏感，是近几年来很热门的一种算法。该算法首先通过 Canny 算子进行边缘检测，提取图像的边缘及边缘梯度方向；然后遍历边缘图像的每一个像素，根据边缘像素的梯度方向，查找其梯度方向相反、角度大致一样的像素形成像素对，像素对间的宽度即为当前像素的笔画宽度；最后根据笔画宽度的变化来检测文本。笔画宽度的定义如图 6 所示。基于笔画特征的算法对高分辨率场景文本检测定位显示出很强的竞争力[44-45]，尤其在与适当的学习方法[46]或时空分析[47]相结合后，性能更好。Mosleh 等人[48]引入基于带状的边缘检测器来改进 SWT，该边缘检测器增强了文本边缘并有效地消除了噪声和叶面边缘，适用于低分辨率文本检测定位。

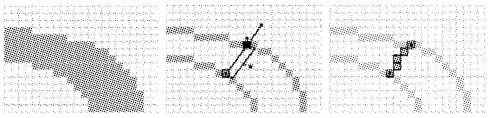


图 6　笔画宽度的定义[44]

SWT 算法能够检测大部分文本区域，但当图像背景复杂时会产生严重的虚警，如树叶、条纹、灌木丛、标志、房屋等往往会被误认为候选文本区域。

2.1.5　基于深度学习的算法

深度学习是机器学习研究中的一个新领域，其动机在于建立、模拟人脑进行分析学习的神经网络，模仿人脑的机制来解释数据。深度学习是无监督学习的一种，采用了神经网络的分层结构，通过建立类似于人脑的分层模型结构，对输入数据逐级提取从底层到高层的特征，从而能很好地建立从底层信号到高层语义的映射关系。近年来，深度学习模型已被证明对自然场景图像文本识别具有强大的功能。

为了充分利用整个场景图像中丰富的信息，Yao 等[49]提出将场景文本检测作为语义分割问题，以整体方式检测文本。算法直接在整体图像上运行，并产生全局像素预测图，随后使用一个完全卷积网络（Fully Convolutional Network，FCN）检测文本。算法可同时检测场景图像中的水平、多方向和扭曲的文本。

Zhong 等[50]在文本区域定义了一个名为深度文本（Deep Text，DT）的统一框架，并通过一个完全的卷积神经网络（Convolutional Neural Network，CNN）对场景图像文本进行检测。首先，作者提出了起始区域提议网络（Region Proposal Network，RPN），并设计了一套文本特征先验边界框来提高召回率；然后，提出一个强大的文本检测网络，网络嵌入了模糊文本类别（Ambiguous Text Category，ATC）信息和多级感兴趣的区域集合（Multilevel Region-of-interest Pooling，MLRP）；最后，使用迭代边界框投票方案提高召回率，并引入一个过滤算法来删除每个文本实例中冗余的内部和外部边界框。

Zhang 等[51]提出一种新颖的场景文本检测算法。不同于传统利用单个字符或笔画的特性，该算法基于字符组的对称性，从自然场景图像中直接提取文本行。

He 等[52]提出一种新颖的基于级联卷积文本网络（Cascaded Convolutional Text Network，CCTN）的场景文本检测算法。CCTN 连接了两个自定义的卷积网络，用于由粗到精的文本定位。该算法对于多语言、多方向的文本具有强健的性能。同年，He 等[53]提出一种针对文本的卷积神经网络（Text-attentional Convolutional Neural Network，Text-CNN）算法来检测场景图像文本，该算法特别关注文本区域的特征，开发出一种新的学习机制，通过多层次丰富的监督信息对 Text-CNN 进行训练。监督信息包含文本区域掩码、字符标签和文本/非文本的二值化信息，这些信息使得 Text-CNN 具有强大的检测模糊退化文本的性能，并且增强了复杂背景图像的鲁棒性。

深度学习模型不仅大幅提高了图像中文本识别的精度，而且也避免了需要消耗大量的时间进行人工特征的提取，使得在线运算效率大大提升。然而，深度学习需要选取样本进行训练，因此训练样本集与测试样本集的相似度不高时所取得的效果也不够理想。

2.1.6 *其他算法*

由于图像中的文本存在大量的变化源，上述单独的算法通常在特定的条件下失效。为处理这些变化源，一些研究者提出了一些综合性算法[39-42]。

现存的大部分算法都是针对水平方向文本的检测，导致大部分非水平方向的文本区域未被检测出来而产生了严重的误报。Yao 和 Bai[54]构建了一个能够有效检测自然场景图像中任意方向文本的实际检测系统(场景图像中任意方向的文本样本如图 7 所示)。作者利用 SWT 的旋转不变特性和一个二级分类方案来区别文本和非文本，因此这个系统能够有效地检测任意方向的文本。Pan 等[55]联合了基于区域和基于连通分量的方法。首先，设计一个区域检测器来估测图像金字塔中每一层的文本区域，并利用尺度自适应二值化生成文本区域；然后，在区域分析阶段利用 CRF 模型滤除非文本区域；最后，通过最小跨越树聚集文本。Toan Dinh Nguyen 等[56]提出了一种非常新颖的算法，即利用二维张量投票来鉴定文本区域和非文本区域。通过张量投票来提取文本行信息，降低了基于区域的文本检测算法中的误报率。

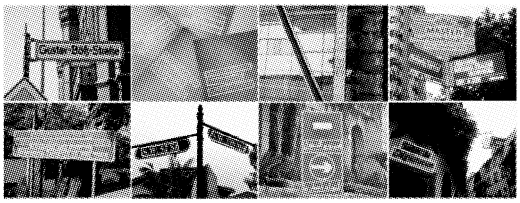
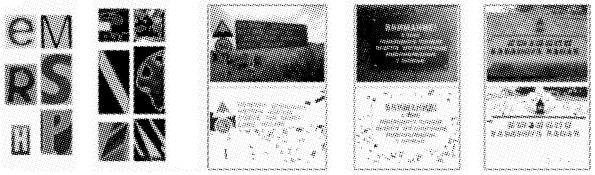


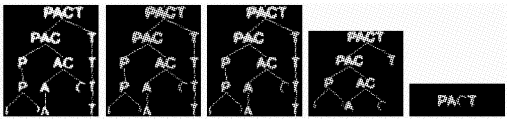
图 7 场景图像中任意方向的文本检测结果样本[54]

基于最大稳定极值区域(Maximally Stable Extremal Region，MSER)的文本定位已被广泛研究[38,57,59]。这种算法的主要优点在于使用 MSER 作为文本候选区域的有效性。文本区域与背景通常具有显著的颜色对比度，且倾向于形成均匀的颜色区域，因此自适应检测稳定颜色区域的 MSER 算法为定位文本提供了可行的解决方案[60-61]。Neumann 和 Matas[57]第一次将 MSER 引入场景图像文本检测，高效实现了具有尺度和旋转不变性的特征检测，鲁棒性良好。该算法在 Char74k 数据库中，实现了 72% 的识别率；在 ICDAR 20003 文本检测数据库中，查准率为 59%，f 指数为 0.57，实验结果如图 8 所示。Yin 等[61]采用修减算法选择适当的 MSER 作为文本候选区域，并通过混合特征对候选进行验证，算法在低对比度、复杂背景和字体变化的条件下性能良好，实验结果如图 9 所示。Huang 等[62]提出一种结合 MSER 和 CNN 的新算法来检测场景图像文本。算法首先用 MSER 检测文本候选区域，然后采用 CNN 分类器来识别正确的候选区域，并分割连接在一起的多个字符。Lluis 等[63]提出一种专门为文本图像设计的对象建议(Object Proposals，OP)算法。该算法首先使用 MSER 实现图像的初分割；然后通过自定义的距离公式合并初始区域，复杂场景需要多种距离策略来进行计算；最后生成候选区域集合，并利用一个弱分类器进行排序，最终得到文本区域。该算法分别在 SVT 和 ICDAR'13 数据库进行了测试，性能优越。

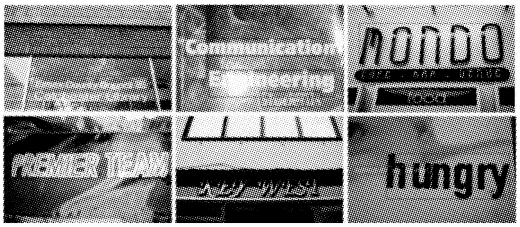


(a)字符样本 (b)非字符样本 (c)亚美尼亚语检测结果 (d)俄语检测结果 (e)卡纳达语检测结果

图 8 文献[57]中的实验结果



(a)MSER 修剪



(b)在数据库 ICDAR 2011 中的实验结果示例

图 9 文献[61]中的实验结果

2.2 文本字符的增强与分割

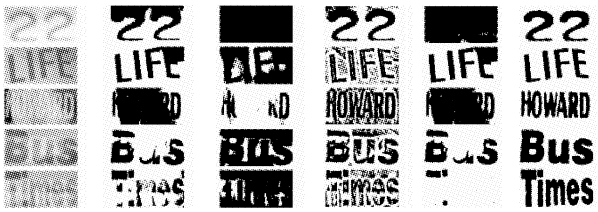
视频和场景图像常常具有低分辨率、模糊和透视失真等特点，加上内容和背景交互的布局复杂，检测出来的文本区域并不能直接用于 OCR 系统进行字符识别，因此需要进行文本行分割和字符分割来获得字符的精确边界，分割已被确定为最具有挑战性的问题之一。大部分检测出来的文本区域具有简单背景和高对比度，能够准确地定位和提取字符边缘轮廓，然而低质量的文本却很难提取。下文将专注于复杂背景下的文本增强与分割。

文本分割也即文本二值化，旨在提取文本像素并删除背景像素。相关算法有阈值算法[64]、概率模型[65-69]和聚类算法[74-75]等。

阈值算法可分为全局阈值法和局部阈值法。全局阈值法根据图像的直方图或灰度的空间分布确定一个阈值，并根据此阈值将灰度图像转化为二值图像。典型的全局阈值法有 Otsu 算法[70]、最大熵算法、迭代算法等。全局阈值法简单，对于目标和背景明显分离、直方图分布呈双峰的图像效果良好；但由于文本图像一般都存在背景复杂、光照不均匀的特点，因此单一的全局阈值法很难得到理想的分割效果，会出现细节丢失等现象。局部阈值法通过定义像素点的邻域，由邻域计算模板实现像素点灰度与邻域点的比较。典型的局部阈值法包括 Bernsen 算法[71]、Niblack 算法[72]、Sauvola 算法[73]等。局部阈值法虽然能够根据局部特性自适应选取阈值，但由于过分夸大图像细节，会造成伪影、断笔等现象。由于文本边界处的像素通常与背景融合，因此很难为退化的文本图像选择可靠的阈值。

当对文本中大量前景像素进行采样后，可以采用高斯混合模型对文本图像进行二值化操作[65-67]。受到马尔科夫随机场(Markov Random Field，MRF)模型在图像分割中成功应用的启发，Mishra 等[68]将 MRF 运用到文本图像二值化中，

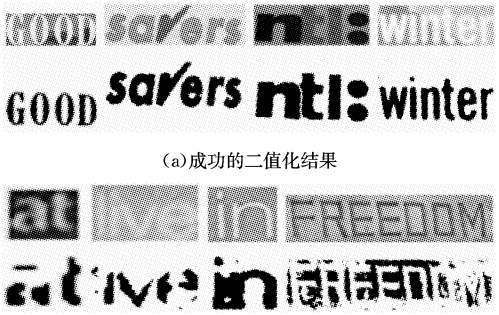
作者将图像中的每个像素表示为 MRF 中的随机变量，并在这些变量中引入新的能量函数，这里的能量函数采用高斯混合模型。最后，每个变量由能量函数标记为前景或背景，实验结果如图 10 所示。Lee 等[69]提出一种分为两步的 CRF 场景图像二值化方法，算法基于层次空间结构来标记连续的文本区域，进而分割字符。



(a)原图　(b)Otsu　(c)Sauvola　(d)Niblack　(e)Kittler　(g)本文算法

图 10　文献[68]中的实验结果

在提取视频中的退化文本时，更倾向于聚类算法[74-75]。Thillou 等[74]使用多个颜色度量和聚类方法来提取文本像素，作者使用从 Log-Gabor 滤波器获取的空间信息来补充颜色度量。Wakahara 等[75]提出一个基于 K 均值聚类和支持向量机 SVM 的算法，其能够从自然场景中提取严重退化的多彩色字符的文本。该算法分为 4 步：1)对 HSI 颜色空间中给定的图像像素进行 K 均值聚类，初步的二值化图像总数为 2^k-1；2)根据字符的长宽比将二值化图像划分为“单个字符”图像序列；3)使用 SVM 来确定每个“单个字符”图像是否表示一个字符；4)选择具有最大平均值的单个二值化图像作为最优结果。算法在 ICDAR 2003 中的字符识别率达到 80.8%，实验结果如图 11 所示。

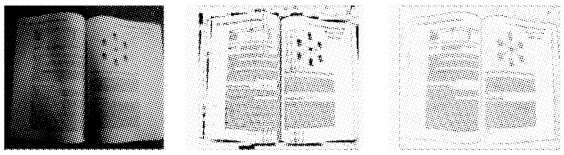


(a)成功的二值化结果

(b)失败的二值化结果

图 11　文献[75]中的实验结果

Zhu 等[76]提出了一种利用 CNN 和双峰图像增强的算法来分割文本，该算法在 ICDAR03 中的字符识别率为86.96%；作者在后续研究中将寻求一个能够同时增强图像和降低噪声的适当方法。Zhou 等[77]提出一种改进的自适应文档图像二值化方法：首先，采用基于局部统计的维纳滤波器给图像降噪，并进行第一次前景区域的粗略估计；然后，通过相邻像素插值计算背景像素值；最后，通过计算预处理图像的背景获得最终的阈值。该算法在照度不均匀时具有良好的鲁棒性，能够较少地丢失笔划并有效地保留边缘信息，实验结果如图 12 所示。Anand Mishra 等[78]提出一个基于自下而上和自上而下线索的框架来提取街道图像中的文本。由于自然场景图像背景复杂，很难直接从背景中分割字符，因此该算法使用滑动窗口检测可能的字符，并将检测结果作为自下而上的信息，而自上而下的信息来自于大型字典的统计信息。自上而下和自下而上的信息通过 CRF 集成于同一模型中。该算法的优点之一在于能够容忍字符检测中的错误，如图 13 所示，两个“o”之间的区域被识别为“x”，然而根据先验信息，“oor”的可能性更高，故该词最终被确认为“door”。



(a)原图　(b)B. Gato 算法结果　(c)本文算法结果

图 12　文献[77]中扭曲的印刷文本图像二值化

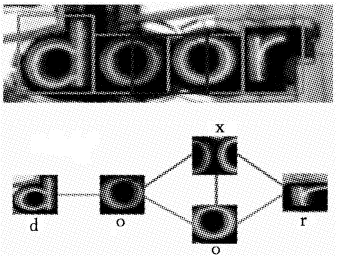


图 13　文献 [78]中自上而下和自下而上模型下的场景文本识别

Liu 和 Wang[79]提出一种基于自适应阈值的新算法，将图像二值化视为一个优化问题，通过分割 Otsu 阈值获得最佳阈值，并消除边界效应进一步优化二值化。实验表明，这种算法能够较好地保持原始边缘特性，特别是在具有丰富的边缘信息的图像中获得了更好的二值化优化效果。Jiang 等[80]描述了一个低质量文本的二值化算法。算法在灰度化图像上进行膨胀和腐蚀操作，并结合文献[81]中提出的二值化算法和矩形区域小邻域进行二值化，实验结果如图 14 所示。Le 和 Lee[82]提出了两种关于扭曲文本的二值化方法，一个使用映射函数，另一个使用双二次变换函数。算法利用 Hough 变换和贝塞尔曲线近似法检测标记边界线。实验结果表明，该算法可以正确恢复原始标签的矩形区域。

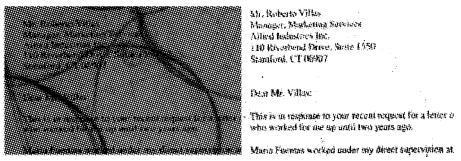


图 14　文献[80]中的二值化结果

3　性能评估

文本检测性能评估是衡量算法优劣的标准。下面首先分别从性能指标和算法测试所用的公共数据库两方面进行讨论，随后讨论了现有视频和图像中文本提取算法中存在的问题。

3.1　性能指标

文本检测中主要的性能指标有查全率 r (recall)、查准率 p(precision)和 f 指数(F measure)等。这些指标来源于信息检索中的评价参数，定义如下：

$$r=\frac{C}{T} \tag{1}$$

$$p=\frac{C}{E} \tag{2}$$

$$f=\frac{1}{\frac{\alpha}{p}+\frac{1-\alpha}{r}} \tag{3}$$

其中，C 为正确检测出来的文本区域的数量，T 为实际文本区域的数量，E 为检测出来的文本区域数量(包含虚警区域)。f 指数是查全率和查准率的加权调和平均，其中 α 为加权因子。

表 1 按时间顺序列出了近年来在 ICDAR 2003/2005/2011/2013 数据库进行测试的一些参考文献。表中数据数值越高则性能越好，每一项检测指标最好的结果已用黑体加粗。

表 1 文本提取简要总结

作者	年份	查准率 p	查全率 r	f 指数	特点
Hu 和 He[43]	2008	0.58	0.74	0.65DW	边缘检测，文本提取
Bui 等[39]	2009	0.787	0.734		地形图、卷积神经网络、文本检测
Pan 等[26]	2010	0.66	0.70	0.68	文本检测、特征提取、由粗到精
Lee 等[34]	2010	0.69	0.60	0.64	边缘约束、文本检测
Minett 等[40]	2010	0.63	0.61	0.61	文本检测、多分辨率、图像分割
Pan 等[31]	2011	0.674	0.697	0.685	条件随机场、连通区域分析
Pan 等[55]	2011	0.68	0.67	0.67	文本检测、笔画分割、条件随机场
Epshtein 等[44]	2011	0.73	0.60	0.66	文本检测、笔画宽度变换
Zhou 等[21]	2011	0.37	**0.88**	0.53	多语言、场景文本检测、HOG、MG、LBP
Neuman 等[57]	2012	0.647	0.731	0.687	极值区域、文本定位
Palaiahna 等[83]	2012	0.72	0.87	0.78	贝叶斯分类器、边界生长、文本检测
Koo 等[38]	2013	0.764	0.619	0.684	连通区域聚类、MSER、文本检测
Lee 等[69]	2013	0.666	0.619	0.713	TCRF、颜色聚类、文本检测
Wei 等[84]	2014	0.87	0.79	0.762	边缘分布熵、SVM、文本提取
Yin 等[61]	2014	0.863	0.689	0.78	MSER、单链聚类、文本检测
Huang 等[62]	2014	0.88	0.71		MSER、CNN、文本检测
Zhang 等[51]	2015	0.88	0.74	0.80	对称检测器、LBP、角度和距离约束、CNN
Zhong 等[50]	2015	0.85	0.81	0.83	CNN、RPN、ATC、MLRP、文本检测
Yao 等[49]	2016	0.889	0.802	**0.843**	整体性、多通道预测、FCN、文本检测
Cho 等[19]	2016	0.863	0.785	0.822	Canny 算子、ERs、双阈值分类
He 等[52]	2016	0.88	0.79	0.84	CCTN、多层融合、由粗到精、文本检测
He 等[53]	2016	**0.91**	0.74	0.82	Text-CNN、MSER、文本检测

3.2 公共数据库

场景文本检测有丰富的公用数据库，典型的数据库有 ICDAR，Char74k，KAIST，SVT，MSRA-TD500，OSTD，NEOCR，COCO-text，DOST，FSNS 等。本文总结一些典型图像文本数据库，具体信息如表 2。图 15 展示了上述数据库中的一些图像样本。

ICDAR'03[85]是第一个正式发布的基准数据库，用于场景文本检测和识别。该数据库包含 509 个完全注释的文本图像，其中 258 个图像用于训练，251 个图像用于测试。ICDAR'03/05 数据库已广泛地运用于文本检测中，但其仍存在两个主要的缺点：数据库中大部分文本是水平的，数据库中所有的文本都是英语。ICDAR'11[89]和 ICDAR'13[91]数据库在 ICDAR'03/05 数据库上进行了扩展，包含多方向和扭曲的文本图像、视频文本图像、网页图像和电子邮件等。其中，ICDAR'13 除英文外，还有西班牙语和法语。ICDAR'03 / 05 和 ICDAR'11 / 13 数据库主要为场景文本，包含文本定位、字符分割和字符识别等任务。

Chars74k 数据库[86]适用于自然场景图像中的字符识别，图像中的文本均为水平方向文本，包含英语和加拿大语。

SVT 数据库[9]从 Google 街景视图中获取，包含 350 个完整图像。该数据库中的图像文本具有很高的变异性，且通常具有低分辨率。

KAIST 场景文本数据库[8]由 3000 幅不同环境下的图像构成，包含不同光照条件下(自然光、强烈的人为照明等)的室内和室外图像。数据库中有高分辨率数字相机拍摄的图像，也有低分辨率手机图像，所有图像的像素尺寸都为 640×480，包含英语、韩语等混合语言。

NEOCR 数据库[87]中的样本为自然场景中的多方向文本，包含 659 个真实世界的图像和 5238 个注释边界框。其涵盖多种语言，如英语、匈牙利语、俄语、土耳其语和捷克语等。

面向场景文本数据库 OSTD 由 Yi 等[88]提出，该数据库包含 89 个标志图像、室内场景和街景图像，可以用于自然场景中的多方向文本检测算法。

MSAR-TD500 数据库[90]是评估自然场景中多方向文本检测算法的基准，包含 500 个具有复杂背景的水平或倾斜的自然场景图像。图像是采用袖珍相机从室内(办公室和商场)和户外(街道)场景中拍摄的，图像分辨率从 1296×864 到 1920×1280 不等。

DOST 数据库[92]集中于城市中心场景图像文本的检测与识别。在此数据库中，将开启 5 项任务：视频文本定位、静态图像文本定位、剪裁字符识别、视频端到端识别、静态图像端到端识别。DOST 数据库包含使用全向摄像头在大阪市区购物街拍摄的视频(连续图像)，数据库中的序列图像有助于鼓励开发一种利用时间信息的新型文本检测识别技术。DOST 数据库的另一个重要特征是它包含非拉丁文本，由于图像在日本被捕获，因此数据库包含大量的日语。

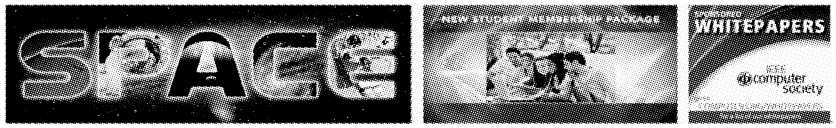
FSNS 数据库[94]包含 100 多万张法国街道名称标志图像，每张图像包含相同街道标志的 4 个视图，路标中的文字可跨越 3 行。数据库面临的主要挑战为：端到端多行场景文本识别、大规模的训练样本，以及利用多个视图提高识别精度。

COCO-Text 数据库[93]是基于 MS COCO 数据库的一种新的大规模数据库，包含复杂的自然场景图像，一共有 63686 张图像，145859 个文本实例，3 个精细的文本属性。文本实例分为机器印刷和手写文字、清晰和模式的文字、英文和非英文样本。数据库围绕 3 个任务进行构造：文本定位、字符识别、端到端识别。

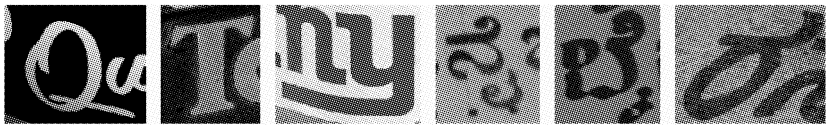
表 2　公用数据库

数据库(年份)	属性	图像数量 (训练/测试)	语言	文本方向
ICDAR'03(2003)[85]	场景文本	2276(1110/1156)	英语	水平
Chars74k(2009)[86]	图形和场景文本	74107	英语、加拿大语	水平
KAIST(2010)[8]	场景文本	3000	英语、韩语	扭曲
SVT(2010)[9]	场景文本	350(100/250)	英语	水平
NEOCR(2011)[87]	场景文本	659	8 个语种	扭曲、多方向
OSTD(2011)[88]	场景文本	89	英语	多方向
ICDAR'11(2011)[89]	场景文本	484(229/255)	英语	水平
	图形文本	522(420/102)	英语	扭曲
MSAR-TD500(2012)[90]	场景文本	500(300/200)	英语、中文	多方向
ICDAR'13(2013)[91]	场景文本	463(229/233)	英语	水平
	图形文本	551(410/141)	英语	多方向
	视频文本	28(13/15)	英语、法语、西班牙语	多方向
DOST(2016)[92]	场景文本	32147	日语、英语	多方向
COCO-Text(2016)[93]	场景文本	63686	英语	多方向
FSNS(2016)[94]	场景文本	＞1000000	法语、英语	多方向

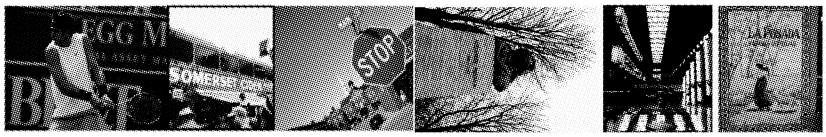
(a) ICDAR'03/05



(b) ICDAR'11/13



(c) Chars74k



(d) COCO-Text　　(e) NEOCR

(f) KAIST　　(g) DOST

(h) MSAR-TD500　　(i) SVT

图 15　公用数据库中的样本图像

3.3　存在的问题

视频和图像中文本提取现存的技术状态和所需性能要求之间的差距仍待解决。虽然相关研究已取得很大进展，但仍有许多进步空间。下面总结当前研究普遍存在的问题，并展望发展趋势。

1）实时检测与识别：视频中文本检测具有重要应用，现存

大多算法都是对视频进行逐帧文本检测，忽略了帧与帧之间的时间线索，效率低下，达不到实时检测与识别的性能要求。一种改进的方法是将文本检测与识别算法和文本跟踪算法结合，这样不仅可以提高检测和识别的精度，还能提高实时性能。

2)多语言文本识别：图像中不同语言文本具有不同问题。现存大部分算法都是针对英语设计的，对于其他语种性能有明显下降。由于字符种类多、结构复杂且存在变异性，中国、日本、韩国等东亚国家的文本识别被认为是一个非常困难的问题。使用参数固定的单一算法难以识别所有的语言文本实现。一种可能的解决方案是对每种语言模型使用通用的训练方法。

3)场景文本识别：自然场景中的图像文本多存在退化、扭曲、字体变化和背景复杂等情况。许多算法可以解决单个问题，但很少有算法能解决综合问题。为了解决场景文本识别的一般性问题，必须进一步设计和学习文本的不变特征。

4)端到端识别：与干净的文档图像优良的文本提取性能相比，端到端图像文本识别性能仍远远落后。算法的改进不仅来自更强大的字符识别模型，更应来自于设计良好的信息共享、反馈和优化策略。近年来兴起的大规模深度学习采用了神经网络的分层结构，大幅度提高了图像文本检测与识别的性能。在未来，深度学习与优化的分割、识别算法和高度整合的语言模型相结合可以进一步提升性能。

结束语 随着多媒体技术的发展，视频和图像已经成为当前最为流行的一种媒体表现形式，视频和图像中的文本提取在信息检索应用领域具有不可替代的地位。本文主要关注近年来视频和图像文本提取方法的进展，从文本检测定位、文本分割与增强两个子步骤进行算法的归纳总结。最后，讨论性能评估、图像公用数据库和存在的问题。尽管上文总结了很多图像中文本检测定位、分割增强的算法，但至今仍没有一个针对用户的完整、有效的自动化文本提取系统。场景图像的文本提取仍存在诸多困难，实际运用中的强烈要求需要我们在这一方向上投入更多的精力。

参 考 文 献

[1] 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.

[2] Googlegoggles[OL]. http://www.google.com/mobile/goggles/#text, 2011.

[3] VALIZADEH M, ARMANFARD N, KOMEILI M, et al. A novel hybrid algorithm for binarization of badly illuminated document images[C]// 2009 14th International CSI Computer Conference(CSICC 2009). IEEE, 2009: 121-126.

[4] CHEN X, YANG J, ZHANG J, et al. Automatic detection and recognition of signs from natural scenes[J]. IEEE Transactions on Image Processing, 2004, 13(1): 87-99.

[5] BISSACCO A, CUMMINS M, NETZER Y, et al. PhotoOCR: Reading text in uncontrolled conditions [C]// IEEE International Conference on Computer Vision. IEEE Computer Society, 2013: 785-792.

[6] HE Z, LIU J, MA H, et al. A new automatic extraction method of container identity codes [J]. IEEE Transactions on Intelligent Transportation Systems, 2005, 6(1): 72-78.

[7] SERMANET P, CHINTALA S, LECUN Y. Convolutional neural networks applied to house numbers digit classification [C]// International Conference on Pattern Recognition. IEEE, 2012: 3288-3291.

[8] LEE S H, MIN S C, JUNG K, et al. Scene text extraction with edge constraint and text collinearity [C]// International Conference on Pattern Recognition(ICPR 2010). Istanbul, Turkey, DBLP, 2010: 3983-3986.

[9] WANG K, BELONGIE S. Word spotting in the wild [C]// Proceeding of European Conference on Computer Vision(ECCV). Heraklion, Crete, Greece, 2010: 591-604.

[10] WANG K, BABENKO B, BELONGIE S. End-to-end scene text recognition [C]// International Conference on Computer Vision (ICCV 2011). IEEE, 2011: 1457-1464.

[11] LIU X, SAMARABANDU J. Multiscale edge-based text extraction from complex images [C]// International Conference on Multimedia and Expo. IEEE, 2006: 1721-1724.

[12] SHIVAKUMARA P, PHAN T Q, TAN C L. A gradient difference based technique for video text detection [C]// 2009 10th International Conference on Document Analysis and Recognition (ICDAR'09). IEEE, 2009: 156-160.

[13] SHIVAKUMARA P, HUANG W, TAN C L. An efficient edge based technique for text detection in video frames [C]// Eighth IAPR International Workshop on Document Analysis Systems (DAS'08). IEEE, 2008: 307-314.

[14] YE Q, JIAO J, HUANG J, et al. Text detection and restoration in natural scene images [J]. Journal of Visual Communication and Image Representation, 2007, 18(6): 504-513.

[15] OU W, ZHU J, LIU C. Text location in natural scene [J]. Journal of Chinese Information Processing, 2004, 18(5): 42-43.

[16] LYU M R, SONG J, CAI M. A comprehensive method for mutilingual video text detection, localization, and extraction [J]. IEEE Trans. on Circuit and Systems for Video Technology, 2005, 15(2): 243-255.

[17] LIU X, SAMARABANDU J. Multisccale edge-based text extraction from complex images [C]// International Conference on Multimedia and Expo. IEEE, 2006: 1721-1724.

[18] KIM W J, KIM C. A new approach for overlay text detection and extraction from complex video scene [J]. IEEE Transactions on Image Processing, 2009, 18(2): 401-411.

[19] CHO H, SUNG M, JUN B. Canny Text Detector: Fast and robust scene text localization algorithm [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3566-3573.

[20] LI H, DOERMANN D, KIA O. Automatic text detection and tracking in digital video [J]. IEEE Transactions on Image Processing, 1998, 9(1): 147-156.

[21] ZHOU G, LIU Y, MENG Q, et al. Detecting multilingual text in natural scene [C]// International Symposium on Access Spaces (ISAS 2011). IEEE, 2011: 116-120.

[22] BERTINI M, COLOMBO C, BIMBO A D. Automatic caption localization in videos using salient points [C]// IEEE International Conference on Multimedia and Expo. 2001: 68-71.

[23] SATO T, KANADE T, HUGHES E K, et al. Video OCR for digital news archive [C]// International Workshop on Content-Based Access of Image and Video Libraries. IEEE, 1998: 52-60.

[24] ZHONG Y, ZHANG H, JAN A K. Automatic caption localization in compressed video [J]. IEEE Transactions on Pattern

Analysis & Machine Intelligence，2000，22(4)：385-392.

[25] GOTO H，TANAKA M. Text-Tracking wearable camera system for the blind [C]// International Conference on Document Analysis and Recognition. IEEE Computer Society，2009：141-145.

[26] PAN Y F，LIU C L，HOU X. Fast scene text localization by learning-based filtering and verification [C]// 17th IEEE International Conference on Image Processing (ICIP 2010). IEEE，2010：2269-2272.

[27] WU V，MANMATHA R，RISEMAN E M. Digital Libraries by recognition of superimposed caption Multimedia Systems [J]. Proc of 2nd ACM International Conference，1999，7(5)：385-395.

[28] ZHAO M，LI S T，KWOK J. Text detection in images using sparse representation with discriminative dictionaries [J]. Image and Vision Computing，2010，28：1590-1599.

[29] SUN L，LIU G Z，JAN X M，et al. A novel text detection and localization method based on corner response[C]// Proc of ICME. 2009：90-393.

[30] ZHANG H，LIU C，YANG C，et al. An improved scene text extraction method using conditional random field and optical character recognition[C]// International Conference on Document Analysis and Recognition(ICDAR 2011). IEEE，2011：708-712.

[31] PAN Y，HOU X，LIU C. A hybrid approach to detect and localize texts in natural scene images [J]. IEEE Transactions on Image Processing，2011，20：800-813.

[32] WANG K，KANGAS J A. Character location in scene images from digital camera [J]. Pattern Recognition，2003，36(10)：2287-2299.

[33] AGNIHTORI L，DIMITROVA N. Text detection for video analysis [C]// International Workshop on Content-Based Access of Image and Video Libraries. IEEE，1999：109-113.

[34] HUA X S，YIN P，ZHANG H J. Efficient video text recognition using multiple frame integration [J]. Proceedings of International Conference Image Processing，2004，11(2)：22-25.

[35] KIM K C K. Scene text extraction in natural scene images using hierarchical feature combining and verification [C]// Proceedings of International Conference Pattern Recognition (ICPR 2004). 2004：679-682.

[36] PAN Y F，HOU X，LIU C L. A hybrid approach to detect and localize texts in natural scene images[J]. IEEE Transactions on Image Processing，2011，20(3)：800-813.

[37] KIM K I，JUNG K，KIM J H. Texture-Based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2003，25(12)：1631-1639.

[38] KOO H I，KIM D H. Scene text detection via connected component clustering and nontext filtering [M]. IEEE Press，2013.

[39] BUI T D，PAN W，SUEN C Y. Text detection from natural scene images using topographic maps and sparse representations [C]// International Conference on Image Processing. IEEE，2009.

[40] LEE S H，CHO M S，JUNG K，et al. Scene text extraction with edge constraint and text collinearity[C]// Proceedings of 20th International Conference Pattern Recognition (ICPR 2010). 2010：3983-3986.

[41] MINETTO R，THOME N，CORD M，et al. A multiresolution system for text detection in complex detection in complex visual scenes[C]// Proceedings of 17th International Conference on Image Processing Snoopertext. IEEE，2010：3861-3864.

[42] JUNG C，LIU Q，KIM J. A stroke filter and its application for text localization [J]. Pattern Recognition Letters，2009，30(2)：114-122.

[43] GUI T Y，SUN J，NAOI S. A fast caption detection method for low quality video images [C]// International Workshop on Document Analysis Systems(IAPR 2012). 2012.

[44] EPSHTEIN B，OFEK E，WEXLER Y. Detecting text in nature scenes with Stroke Width Transform[C]// Proceedings of Computer Vision and Pattern Recognition (CVPR 2010). IEEE，2010：2963-2970.

[45] CHOWDHURY A R，BHATTACHARYA U，PARUI S K. Scene text detection using sparse stroke information and MLP [C]// International Conference on Pattern Recognition. 2012：294-297.

[46] ZHOU Y，LU T，LIAO W. A robust color-independent text detection method from complex videos [C]// International Conference on Document Analysis and Recognition. IEEE，2011：374-378.

[47] 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.

[48] 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.

[49] YAO C，BAI X，SANG N，et al. Scene text detection via holistic，multi-channel prediction[J]. arXiv：1606. 09002，2016.

[50] ZHONG Z，JIN L，ZHANG S，et al. DeepText：A unified framework for text proposal generation and text detection in natural images[J]. Architecture Science，2015(12)：1-18.

[51] ZHANG Z，SHEN W，YAO C，et al. Symmetry-based text line detection in natural scenes[C]// Computer Vision and Pattern Recognition. IEEE，2015：2558-2567.

[52] HE T，HUANG W，QIAO Y，et al. Accurate text localization in natural image with cascaded convolutional text network [J]. Computer Vision and Pattern Recognition，arXiv：1603. 09423，2016.

[53] HE T，HUANG W，QIAO Y，et al. Text-Attentional convolutional neural network for scene text detection[J]. IEEE Transactions on Image Processing，2016，25(6)：2529-2541.

[54] YAO C，BAI X. Detecting texts of arbitrary orientations in natural images[C]// Proceedings of Computer Vision and Pattern Recognition(CVPR 2012). IEEE，2012：1083-1090.

[55] PAN Y F，ZHU Y，SUN J，et al. Improving scene text detection by scale adaptive segmentation and weighted CRF verification [C]// International Conference on Document Analysis and Recognition(ICDAR 2011). IEEE，2011：759-763.

[56] NGUYEN T D，PARK J，LEE G. Tensor voting based text localization in natural scene images [J]. IEEE Signal Processing Letters，2010，17(7)：639-642.

[57] NEUMANN L，MATAS J. A method for text localization and recognition in real world images [C]// Computer Vision-ACCV 2010. New Zealand，2010：770-783.

[58] EITEL A，SPRINGENBERG J T，SPINELLO L，et al. Multimodal deep learning for robust RGB-D object recognition[C]// International Conference on Intelligent Robots and Systems. IEEE，2015：681-687.

[59] SHI C，WANG C，XIAO B，et al. Scene text recognition using part-based tree-structured character detection [C]// Computer

Vision and Pattern Recognition. IEEE,2013:2961-2968.
[60] YE Q,DOERMANN D. Scene text detection via integrated discrimination of component appearance and consensus[M]// Camera-Based Document Analysis and Recognition. 2014:47-59.
[61] YIN X C,YIN X,HUANG K,et al. Robust text detection in natural scene images[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2014,36(5):970-983
[62] HUANG W,QIAO Y,TANG X. Robust scene text detection with convolution neural network induced MSER trees[C]// Proceeding of European Conference on Computer Vision(ECCV). 2014:497-511.
[63] GOMEZ L,KARATZAS D. Object proposals for text extraction in the wild[C]// ICDAR. IEEE,2015:1786-1812.
[64] ZHOU Z,LI L,TAN C L. Edge based binarization for video text images [C]// International Conference on Pattern Recognition (ICPR 2010). Istanbul,Turkey,2010:133-136.
[65] YE Q,GAO W,HUANG Q. Automatic text segmentation from complex background [C]// International Conference on Image Processing. IEEE,2004:2905-2908.
[66] WANG K,BABENKO B,BELONGIE S. End-to-end scene text recognition [C]// IEEE International Conference on Computer Vision(ICCV 2011). Barcelona,Spain,2011:1457-1464.
[67] WEINMAN J J,BUTLER Z,KNOLL D,et al. Toward integrated scene text reading [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2014,36(2):375-87.
[68] MISHRA A,ALAHARI K,JAWAHAR C V. An MRF model for binarization of natural scene text [C]// International Conference on Document Analysis and Recognition. IEEE Computer So-ciety,2011:11-16.
[69] LEE S,KIM J H. Integrating multiple character proposals for robust scene text extraction [J]. Image and Vision Computing, 2013,31(11):823-840.
[70] OHTSU N. A threshold selection method from gray-level histograms [J]. IEEE Transactions on Systems Man & Cybernetics, 2007,9(1):62-66.
[71] BERNSEN J. Dynamic thresholding of gray-level images[C]// International Conference on Pattern Recognition. 1986.
[72] NIBLACK W. An introduction to digital image processing [M]. Strandberg Publishing Company,1985.
[73] SAUVOLA J,PIETIKÄINEN M. Adaptive document image binarization [J]. Pattern Recognition,2000,33(2):225-236.
[74] MANCAS-THILLOU C,GOSSELIN B. Color text extraction with selective metric-based clustering [J]. Computer Vision & Image Understanding,2007,107(1):97-107.
[75] KITA K,WAKAHARA T. Binarization of color characters in scene images using k-means clustering and support vector machines [C]// 20th International Conference on Pattern Recognition(ICPR 2010). IEEE,2010:3183-3186.
[76] ZHU Y,SUN J,NAOI S. Recognizing natural scene characters by convolutional neural network and bimodal image enhancement [C]// International Conference on Camera-Based Document Analysis and Recognition. 2012:69-82.
[77] ZHOU S,LIU C,CUI Z. An improved adaptive document image binarization method [C]// 2nd International Congress on Image and Signal Processing(CISP'09). IEEE,2009:1-5.
[78] MISHRA A,ALAHARI K,JAWAHAR C V. Top-down and bottom-up cues for scene text recognition [C]// Proceedings of Computer Vision and Pattern Recognition(CVPR 2012). IEEE, 2012:2687-2694.
[79] LIU J,WANG C. An algorithm for image binarization based on adaptive threshold [C]// Chinese Control and Decision Conference(CCDC 2009). IEEE,2009:3958-3962.
[80] JIANG L,CHEN K,YAN S,et al. Adaptive binarization for degraded document images [C]// International Conference on Information Engineering and Computer Science (ICIECS 2009). IEEE,2009:1-4.
[81] SAUVOLA J,PIETIKÄINEN M. Adaptive document image binarization [J]. Pattern Recognition,2000,33(2):225-236.
[82] LE H P,LEE G S. Text correction in distorted label images by applying biquadratic transformation [C]// International Conference on Signal and Image Processing Applications (ICSIPA). IEEE,2009:326-329.
[83] SHIVAKUMARA P,SREEDHAR R P,PHAN T Q,et al. Multioriented Video scene text detection through bayesian classification and boundary growing [J]. IEEE Transaction on Circuits and Systems for Video Technology,2012,22(8):1227-1235.
[84] WEI B G,ZHANG Y. A nobel approach to text detection and extraction from videos by discriminative features and density [J]. Chinese Journal of Electronics,2014,23(2):322-327.
[85] LUCAS S M,PANARETOS A,SOSA L,et al. ICDAR 2003 robust reading competitions [C]// International Conference on Document Analysis and Recognition,2003(DBLP). 2003:682-687.
[86] CAMPOS T E D,BABU B R,VARMA M. Character recognition in natural images[C]// Proceedings of the Fourth International Conference on Computer Vision Theory and Applications. Lisboa,Portugal,2009:273-280.
[87] NAGY R,DICKER A,MEYER-WEGENER K. NEOCR:A configurable dataset for natural image text recognition[C]// International Conference on Camera-Based Document Analysis and Recognition. Springer-Verlag,2011:150-163.
[88] YI C,TIAN Y L. Text string detection from natural scenes by structure-based partition and grouping [M]. IEEE Press,2011.
[89] KARATZAS D,MESTRE S R,MAS J,et al. ICDAR 2011 Robust reading competition-challenge 1:reading text in born-digital images (Web and Email)[C]// International Conference on Document Analysis and Recognition. IEEE Computer Society, 2011:1485-1490.
[90] YAO C. Detecting texts of arbitrary orientations in natural images[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2012:1083-1090.
[91] KARATZAS D,SHAFAIT F,UCHIDA S,et al. ICDAR 2013 robust reading competition [C]// International Conference on Document Analysis and Recognition. IEEE Computer Society, 2013:1484-1493.
[92] IWAMURA M,MATSUDA T,MORIMOTO N,et al. Downtown osaka scene text dataset[M]// Computer Vision-ECCV 2016 Workshops. Springer International Publishing,2016.
[93] VEIT A,MATERA T,NEUMANN L,et al. COCO-Text:Dataset and benchmark for text detection and recognition in natural images[J]. Computer Vision and Pattern Recognition, arXiv: 1601.07140,2016.
[94] SMITH R,GU C,LEE D S,et al. End-to-End interpretation of the french street name signs dataset[M]// Computer Vision-ECCV 2016 Workshops. 2016:411-426.