基于自适应色彩聚类和上下文信息的自然场景文本检测

邹北骥[1,2],郭建京[1,2],朱承璋[1,2],杨文君[1,2],徐子雯[1,2]
(1. 中南大学信息科学与工程学院,湖南长沙 410083; 2. 中南大学眼科医学影像处理研究中心,湖南长沙 410083)

摘 要: 自然场景文本检测是图像内容分析和理解的重要前提. 本文提出一种基于自适应色彩聚类和上下文信息分析的方法,用于检测自然场景图像文本. 首先,将层次聚类和参数自学习策略结合,设计一种自适应色彩聚类方法,提取图像中的候选字符. 该自适应色彩聚类方法能针对不同图像自动学习权重阈值,有较好的字符召回率. 然后,利用文本中字符成行出现的性质,设计一种基于上下文信息的字符验证策略,既能保证较高字符召回率,也能有效移除非文本字符. 最后,合并字符构建文本行,并通过后处理得到文本检测结果. 在ICDAR2013公共数据集上的实验结果表明:本文分别获得74.17%的召回率,83.40%的准确率和78.52%的F得分. 与其他文本检测方法相比,本文获得了较好的文本检测性能,说明本文方法的优越性.

关键词: 自然场景文本检测; 自适应色彩聚类; 上下文信息; 自学习策略
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2018)06-1436-09
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.06.024

Natural Scene Text Detection Based on Adaptive Color Clustering and Context Information

ZOU Bei-ji[1,2], GUO Jian-jing[1,2], ZHU Cheng-zhang[1,2], YANG Wen-jun[1,2], XU Zi-wen[1,2]
(1. *School of Information Science and Engineering, Central South University, Changsha, Hunan* 410083, *China*;
2. *Center for Ophthalmic Imaging Research, Central South University, Changsha, Hunan* 410083, *China*)

Abstract: Natural scene text detection is an important task for image analysis and understanding. In this paper, a natural scene text detection method is proposed, using adaptive color clustering and context information analysis. Firstly, combining hierarchical clustering with self-learning strategy, we design an adaptive color clustering method, which learns clustering weights automatically and generates high character recall. Then, considering text in images usually containing several characters, we propose a character verification strategy based on image context information, which can guarantee high character recall and remove non-text components at the same time. Finally, characters are merged to text lines, and further post-processing is applied to generate final text detection results. Experiments on the ICDAR2013 publicly available dataset show that we obtain recall of 74.17%, precision of 83.40% and F-score of 78.52%. Compared with other text detection methods, our method obtains better text detection performance, indicating superiority of the proposed method.

Key words: natural scene text detection; adaptive color clustering; context information; self-learning strategy

1 引言

自然场景图像文本包含大量有效信息,提取图像文本是图像内容分析和理解的重要前提,并可广泛应用于车牌检测、无人驾驶、基于内容的图像检索、文本识别和机器人自动导航等领域[1~4]. 然而,自然场景图像背景复杂,图像中文本色彩、字体、尺寸等多样化,且受到不同程度的光照、阴影、拍摄角度的影响,增加了图像文本检测的难度. 目前,已有的场景图像文本检测方法主要分为两类:基于滑动窗口的方法[5~7]和基于连通域的方法[8~11].

收稿日期:2016-12-22;修回日期:2017-02-24;责任编辑:梅志强
基金项目:国家自然科学基金(No.61573380,No.61702559);湖南省科技计划项目(No.2017WK2074);中南大学创新创业师生共创项目(No.2017gczd016)

基于滑动窗口的文本检测方法通常使用多尺度滑动窗口,扫描原始图像,提取候选文本区域,然后结合候选区域色彩、梯度和纹理等特征,利用机器学习的方法进行验证,得到文本检测结果. 文献[5]使用 16 个尺度的滑动窗口扫描图像,获得候选区域,通过提取候选区域的梯度方差、局部能量和小波变换等特征,并结合训练好的 Adaboost 分类器进行验证,得到文本验证结果. 考虑到基于滑动窗口扫描的方法,提取得到的候选区域过多,文献[7]通过区域对称性分析,提取置信度较高的候选区域,并利用训练好的卷积神经网络(Convolutional Neural Network,CNN),验证候选区域,得到文本检测结果. 由于图像中文本尺寸的多样性,基于滑动窗口的方法通常使用多尺度窗口扫描图像,提取候选文本,使得该方法耗时长,产生的候选区域过多,增大了后续文本验证的难度.

基于连通域的方法是目前较为流行的文本检测方法. 该方法进一步分为三个子任务:(1)候选字符提取,(2)候选字符验证,(3)文本区域分析.

候选字符提取通常考虑图像中文本字符包含的像素点具有灰度一致性、色彩一致性、笔画宽度均一性等特征,进而提取特征相似的像素点,构建候选字符连通域. 常见的候选字符提取方法包括最大稳定极值区域(Maximally Stable Extremal Regions,MSER)[8,9]、笔画宽度变换(Stroke Width Transform,SWT)[12,13]和色彩聚类[14~16]方法. 考虑到图像中文本字符内像素点灰度值具有一定的稳定性,MSER 方法提取多个灰度通道下具有极值性质的连通域作为候选字符. 由于传统的 MSER 方法提取的连通域数目过多,Yin 等[9]通过分析重叠连通域之间的层次关系,构建 MSER 树,并通过比较重叠区域灰度值变化的稳定性,采用最小正则变化策略,对 MSER 树进行修剪,从而减少重叠连通域的数目. MSER 方法能够准确提取图像中的文本字符,并且对低分辨率、强噪音的图像也有一定的适应性,然而,提取的重复区域过多,并且对光照不均和色彩连续变化的图像比较敏感.

通常而言,同一文本字符中的像素点笔画宽度具有一致性,而背景像素点不具有该性质,因此,可用笔画宽度性质,区分文本与背景像素点. Epshein 等[13]根据图像边缘,计算像素点的笔画宽度,将笔画宽度相近的像素点,组成连通域提取候选字符. 在文献[13]的基础上,Huang 等[12]将传统 SWT 方法与色彩信息相结合,提出了一种笔画特征变换算子(Stroke Feature Transform,SFT),用于检测图像中的候选字符,与传统 SWT 方法相比,解决了边缘像素点错误匹配的情况,使得像素点的笔画宽度计算更为准确,但是需要设定过多的参数,对复杂背景的图像,提取的候选字符不准确.

基于自适应色彩聚类的方法通过分析图像中像素点的颜色特征,将色彩相近的像素点聚为相同类,构建色彩层,进而标记连通域作为候选字符. 传统的色彩聚类方法通常需要预先设定固定聚类数目,对图像中像素点聚类得到颜色层. 然而,自然场景图像背景复杂多变,对不同图像采用固定聚类数目会造成部分图像聚类结果不理想,进而影响字符提取的性能. 为避免对不同图像设定固定的聚类数目,Yi 等[16]提出了自适应色彩聚类方法,该方法综合运用图像颜色直方图分析和传统 K-means 聚类算法,对图像像素点聚类,标记聚类结果图中的连通域作为候选字符. 为了进一步合理选取初始聚类中心,Wu 等[14]将图像像素点投影到三维色彩空间,将色彩空间进行子区域划分,分析子区域像素点密度,选取密度最大的子区域对应的均值,作为初始聚类中心,并结合 K-means + + 方法,合理选择其他聚类中心,使得聚类结果更准确. 现有的色彩聚类方法能在一定程度上对不同图像聚成不同数目的类别,然而,这些方法仍然需要人工设定相关聚类参数,如颜色距离,无法完全自适应图像的变化,具有一定的局限性.

候选字符验证通常通过对字符和背景区域进行分析,提取一系列易于区分背景和文本的特征,并结合机器学习的方法验证候选区域,移除非文本字符. 文献[8]提取候选字符区域的面积、周长、拐点和欧拉数等特性,结合 Adaboost 和支持向量机分类器,分两个阶段验证字符. 该方法提取特征维度少,字符验证效率高,能实现实时文本检测的效果,但是对于部分特征不明显的字符,验证效果不理想. 文献[14]将候选字符区域的色彩、笔画宽度等几何特征与梯度方向直方图(Histogram of Oriented Gradient,HOG)特征结合验证候选字符,并在字符合并构建文本行时,召回错误移除的字符,提高了字符检测率,对特征不明显的字符也有较好的验证效果.

文本区域分析通常是对验证后保留的字符进行后处理操作. 一般,通过分析字符区域的空间位置,色彩、纹理等特征,将位置、色彩和纹理相近的字符进行合并,构成文本行,然后使用启发式规则和机器学习的方式对文本行进行分词和验证,得到最终的检测结果.

2　本文方法

考虑到现有色彩聚类文本检测和字符验证方法的不足,本文提出一种基于自适应色彩聚类和上下文信息的文本检测方法,主要包含 4 个步骤,其流程如图 1 所示. 首先,对输入图像(图 1(*a*)第一行图像所示)进行预处理,移除高光影响,得到结果如图 1(*a*)第二行图像所示. 然后,结合层次聚类和参数自学习策略,实现自适应色彩聚类方法,分别提取输入图像和去高光图像

的候选字符,结果如图 1(*b*)两幅图像中伪彩色标记的连通域所示. 提取候选字符的几何和梯度方向直方图特征,输入训练好的分类器,移除非文本字符(图 1(*c*)第一行图像所示),并利用图像上下文信息,召回错误移除的字符,得到结果如图 1(*c*)第二行图像所示. 进一步,使用启发式规则合并字符,构建文本行(图 1(*d*)第一行图像所示),并对文本行进行后处理,得文本检测结果如图 1(*d*)第二行图像所示.

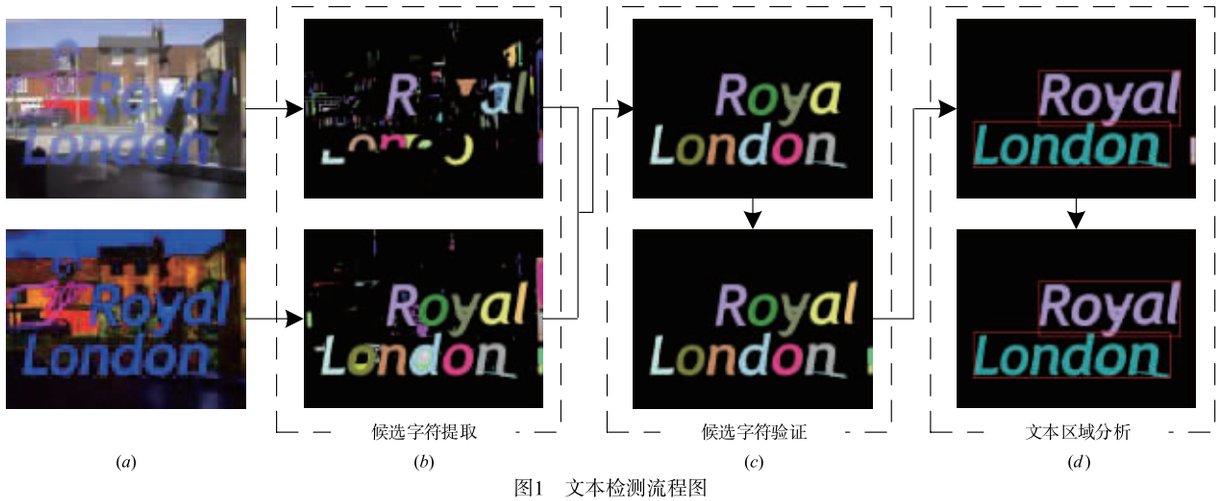


图1 文本检测流程图

2.1 去高光预处理

自然场景图像通过户外拍摄采集得到,受环境影响较大,其中光照不均对图像质量影响严重. 从光照不均的图像中检测文本区域较为困难,这是因为不均匀的光照会使图像失去原有色彩,模糊文本边界,降低文本与背景之间的差异,增加了文本检测的难度. 为减少光照不均对文本检测的影响,本文采用文献[17]提出的基于双边过滤的实时高光去除算法,对原始图像进行去高光处理. 对图 1(*a*)第一行所示受高光影响的图像进行去高光处理,得到结果如图 1(*a*)第二行图像所示,从中可以看出,相对于原始图像,去高光处理后的图像文本色彩更为均一,并与图像背景对比度有较大提高. 因此,本文同时对原始图像和去高光图像进行后续文本检测操作.

2.2 候选字符提取

颜色信息是人眼辨别物体的重要依据,也是区分图像前景与背景的重要特征. 对于自然场景文本图像,相同文本中字符像素点的色彩较为相近,同时与背景像素点的色彩又存在一定的差异,因此,可以通过分析图像中像素点的色彩特征,提取候选字符. 常用的色彩特征聚类方法如 K-Means,通常需要预先设定初始聚类数目,进而将图像像素点聚类到不同的类别. 然而,自然场景图像背景复杂,色彩多变,不同图像颜色复杂程度各不相同,使用固定类别的聚类数目得到的色彩聚类结果不佳,影响字符提取的准确率. 因此,针对不同自然场景图像进行自适应色彩聚类显得尤为重要. 针对字符提取任务,专家学者提出了相应的自适应色彩聚类方法[14~16],虽然能在一定程度上避免聚类类别数对聚类结果的影响,但仍需设定部分阈值,如类别间的颜色距离,进行色彩聚类,泛化能力有限.

为了对自然场景图像进行自适应色彩聚类,提取候选字符,同时避免人为设定部分参数,本文将层次聚类和参数自学习策略相结合,提出了一种自适应色彩聚类方法. 主要的步骤如下:

Step1 提取聚类基本单元. 根据图像中像素点的 R、G、B 值,将其投影到三维色彩空间. 根据文献[14]的方法,将三维色彩空间划分为若干个尺寸一致的小立方体. 提取包含像素点的小立方体,作为层次聚类的基本单元. 计算每个小立方体中像素点的色彩均值,作为聚类基本单元的特征,表示为 $\boldsymbol{c}=(\mu(r),\mu(g),\mu(b))$,其中 $\mu(r)$、$\mu(g)$ 和 $\mu(b)$ 分别为小立方体中像素点的 R、G、B 色彩均值.

Step2 权重向量初始化. 随机初始化权重向量 $\boldsymbol{w}=(w_r,w_g,w_b,w_\theta)$,其中 w_r,w_g,w_b 分别为 $\mu(r)$、$\mu(g)$、$\mu(b)$ 的色彩距离权重,w_θ 为聚类阈值.

Step3 构建层次聚类树. 根据 Step2 中初始化的特征权重向量 $\boldsymbol{w}$,按式(1)计算 Step1 中单元节点两两之间的颜色距离,并合并颜色距离最小的单元节点,构建新的节点. 计算新节点的颜色特征和与其他节点的颜色距离,重复此步骤,由此构建层次聚类树.

$$d_{i,j}=w_r|\mu_i(r)-\mu_j(r)|+w_g|\mu_i(g)-\mu_j(g)|+w_b|\mu_i(b)-\mu_j(b)| \tag{1}$$

其中,$\mu(r)$、$\mu(g)$、$\mu(b)$ 分别为节点中所包含像素点的 R,G,B 色彩均值.

Step4 划分层次聚类树,构建特征向量. 根据初始

设置的 w_0,将 Step3 中的层次聚类树划分为层次聚类森林,如图(2)所示. 其中,蓝色实线划分得到不同的子树,每棵子树下包含不同的单元节点,如图红色虚线框所示. 将同一子树下叶子节点两两之间的颜色距离作为正样本的特征,同时,计算不同子树叶子节点之间的颜色距离作为负样本的特征,由此构建正、负样本的特征向量.

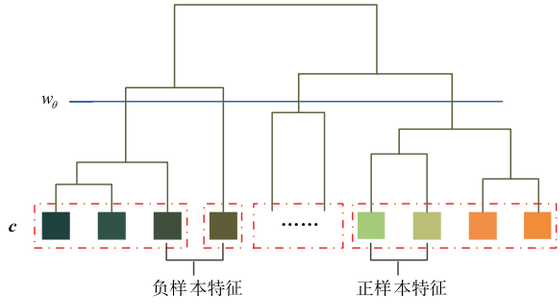


图2　层次聚类示意图

Step5　更新权重向量.

(1)根据权重向量和提取的样本特征,本文采用逻辑回归函数(如式(2)所示)对样本进行类别预测.

$$h_{\boldsymbol{w}}(\boldsymbol{x})=\frac{1}{1+\mathrm{e}^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}} \tag{2}$$

其中,$\boldsymbol{x}$ 为输入向量,由样本特征和截距项 -1 组成,表示为 $\boldsymbol{x}=(|\mu_i(r)-\mu_j(r)|,|\mu_i(g)-\mu_j(g)|,|\mu_i(b)-\mu_j(b)|,-1)$,$h_{\boldsymbol{w}}(\boldsymbol{x})$ 为样本的预测结果. y 为样本真实标签,用 0 和 1 进行表示.

(2)根据样本预测值和其真实标签,构建似然函数,如式(3)所示,其中 $p(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{w})$ 是 $\boldsymbol{x}$ 关于参数 $\boldsymbol{w}$ 的概率密度函数,$\boldsymbol{x}^{(i)}$ 和 $y^{(i)}$ 分别表示第 i 个样本的输入向量和真实标签,n 为样本总量. 为了方便求解似然函数,对式(3)两边取对数,得到对数似然函数,如式(4)所示.

$$\begin{aligned}L(\boldsymbol{w})&=\prod_{i=1}^{n}p(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{w})\\&=\prod_{i=1}^{n}(h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}))^{y^{(i)}}(1-h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}))^{1-y^{(i)}}\end{aligned} \tag{3}$$

$$\begin{aligned}\ell(\boldsymbol{w})&=\log L(\boldsymbol{w})\\&=\sum_{i=1}^{n}y^{(i)}\log h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)})+(1-y^{(i)})\log(1-h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}))\end{aligned} \tag{4}$$

(3)使用随机梯度上升法,最大化对数似然函数 $\ell(\boldsymbol{w})$,由此求解权重向量 $\boldsymbol{w}$. 对式(4)求 $\boldsymbol{w}$ 的梯度,得到结果如式(5)所示,并根据式(6)更新权重向量 $\boldsymbol{w}$,其中 α 为随机梯度上升法的学习速率.

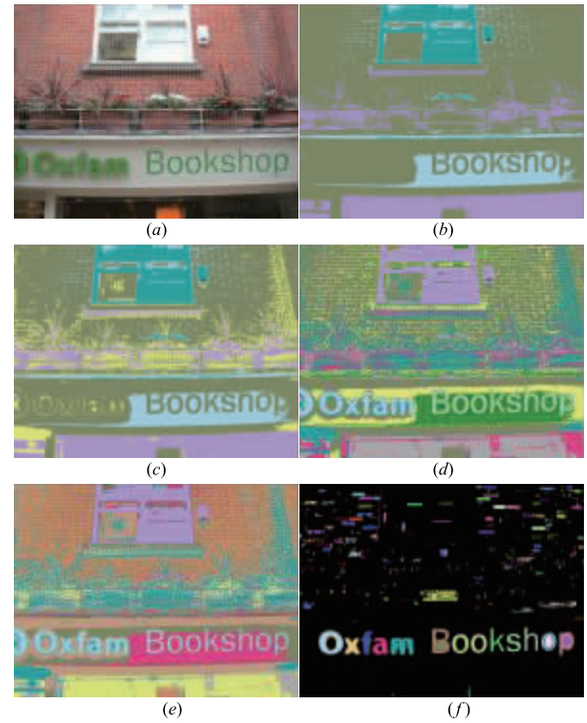
$$\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w})=\frac{\partial\ell(\boldsymbol{w})}{\partial w_j}=(y-h_{\boldsymbol{w}}(\boldsymbol{x}))\boldsymbol{x}_j \tag{5}$$

$$\boldsymbol{w}_j'=\boldsymbol{w}_j+\alpha((y-h_{\boldsymbol{w}}(\boldsymbol{x}))\boldsymbol{x}_j) \tag{6}$$

(4)重复步骤(1)~(3),直到对数似然函数 $\ell(\boldsymbol{w})$ 收敛或者梯度 $\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w})$ 接近为 0 为止.

Step6　构建最佳层次聚类树,提取色彩层,得到候选字符.　根据 Step5 中更新的权重向量 $\boldsymbol{w}$,重复 Step3~Step5,直到 Step5 中对数似然函数 $\ell(w)$ 收敛,得到最佳的权重向量,构建最优层次聚类树. 然后,根据最终的 w_0,划分层次聚类树,得到层次聚类森林. 将层次聚类森林中每一个树包含的基本单元中的像素点提取构建色彩层,得到自适应色彩聚类结果. 标记每个颜色层中的连通域,得到候选字符.

通过上述自适应色彩聚类方法,可以从不同色彩复杂程度的图像中,提取不同数目的颜色层,同时避免人为设定参数,提高了算法的鲁棒性. 图 3 举例说明了本文自适应色彩聚类方法构建色彩层的过程和得到的字符提取结果. 图 3(a)为原始图像,图 3(b)为根据 Step2 中随机初始化的权重向量 $\boldsymbol{w}$,构建得到的色彩层,其中,将同一色彩层中的像素点标记为相同的伪色彩,从中可以看出,部分文本字符与背景无法分离,字符提取效果不佳. 根据本文自适应聚类方法,更新权重向量 $\boldsymbol{w}$,得到分别更新一次、更新三次和最终的权重 $\boldsymbol{w}$ 对应的色彩聚类结果,分别如图 3(c),图 3(d),图 3(e). 从中可以发现,随着权重的更新,图像文本与背景逐渐分离. 提取图 3(e)中的连通域,得到候选字符,如图 3(f).



(a)　(b)　(c)　(d)　(e)　(f)

图3　自适应色彩聚类示意图

2.3 候选字符验证

从图3(*f*)中可以看出,提取得到的候选字符,包含大量的非文本连通域. 本文分别提取候选字符的几何和HOG特征,验证候选字符. 考虑到传统分类器将提取得到的几何和HOG特征相结合,构建特征向量,训练单一分类器,仅仅考虑单个字符的特征,忽略了字符间的上下文关系,容易导致部分文本特征不明显的字符(如'1','i','l')被错误移除. 因此,本文将几何特征和HOG特征分离,分别训练基于几何特征的分类器和基于HOG特征的分类器. 进一步采用上下文特征分析的方法,召回部分错误移除的字符. 主要步骤如下:

Step1 几何特征提取与字符验证. 根据文献[8,10]的方法计算候选区域的几何特征,主要包括笔画宽度、占空比、拐点数等16维特征,输入训练好的几何特征分类器,计算候选字符的得分. 其中,正、负样本得分分别为1和-1.

Step2 HOG特征提取与字符验证. 为了提取候选区域的HOG特征,本文将候选区域尺寸归一化为28×28,采用文献[18]的方法,提取144维HOG特征,输入训练好的HOG特征分类器,计算候选字符的得分. 其中,正、负样本得分分别为1和-1.

Step3 字符置信度评估. 将总得分为2的候选字符视为文本;总得分为-2的候选字符视为背景,直接移除;总得分为0的候选字符视为待校验字符,利用字符与字符之间的上下文信息进一步验证,召回待校验字符集合中的文本.

Step4 文本字符召回. 对Step3中的每一个文本字符C,根据其外接矩形框,沿水平方向往两端扩展至图像边界,沿垂直方向往两端扩展原矩形框高度的0.5倍,构建待召回字符区域R. 从Step3得到的待校验字符中找出位于区域R的连通域,将颜色和高度与字符C相近的连通域视为文本字符.

图4举例说明候选字符验证的步骤. 图4(*a*)为原始图像,对原始图像提取候选字符,根据Step1~Step3计算候选字符得分,并根据候选字符得分,标记为不同的颜色,如图4(*b*)所示. 其中,将得分为2的文本连通域用绿色矩形框标记;将得分为0的待校验连通域用蓝色矩形框标记;将得分为-2的背景连通域用黄色矩形框标记. 根据Step4召回部分错误移除的文本连通域如图4(*c*)所示,从中可以看出,根据字符'g'构建的待召回区域(黑色虚线矩形框所示),能将错误移除的字符'i'和'l'召回. 图4(*d*)为最终的候选字符验证结果,从中可以看出,非文本字符被准确移除,同时文本字符也被完整保留. 由此说明本文提出的基于上下文信息的字符验证策略能保证较高的字符召回率,也能准确移除非文本字符.

(*a*) (*b*) (*c*) (*d*)

图4 候选字符过滤示意图

2.4 文本区域分析

自然场景图像中的文本通常由多个字符连接形成文本行,由此反映图像的语义信息. 常见的文本竞赛评价标准也是基于单词和文本行进行文本检测性能评估,因此本文使用现有的基于启发式规则的方法合并字符构建文本行. 将2.3节中通过验证和召回得到的字符两两组合,形成字符对,将宽高比、水平距离和颜色距离满足一定条件(如式(7)~(9)的字符对视为文本字符对[13,15],合并包含相同连通域的字符对,构建文本行.

$$\frac{\max(w(R_1),w(R_2))}{\min(w(R_1),w(R_2))}+\frac{\max(h(R_1),h(R_2))}{\min(h(R_1),h(R_2))}<3 \tag{7}$$

$$\frac{0.7\times h_d}{w(R_1)+w(R_2)}+\frac{6\times v_d}{h(R_1)+h(R_2)}<2 \tag{8}$$

其中,h_d和v_d分别表示字符区域R_1和R_2两个中心点之间的水平距离和垂直距离.

$$|\text{mean}(R_1)-\text{mean}(R_2)|<80 \tag{9}$$

其中,mean(R)表示字符区域R中像素点的色彩均值.

对于构建得到的文本行,采用式(10)对其进行单词划分.

$$h_d>\alpha*\bar{h}_d+\beta \tag{10}$$

其中,h_d为相邻字符之间的水平间距,$\bar{h}_d$为所有水平间距的均值. α取值为1.5,β为所有水平间距的中位数.

对于文本行分词后得到的仅包含单个字符的单

词,根据 2.3 节 Step1 ~ Step3 的方法,若其计算得到的置信度得分为 2,则将其保留,否则视为负样本进而移除.

3　实验分析

将本文方法在 ICDAR2013 公共数据集上进行实验,并与现有的文本检测方法从召回率、正确率和 F 得分方面进行定量比较.

3.1　数据集

ICDAR2013 数据集[19]是文档分析与识别国际会议(International Conference on Document Analysis and Recognition,ICDAR)公开的公共数据集,也是文本检测和识别竞赛的标准数据集. 该数据集包括 462 张自然场景图像,其中 229 张为训练图像,233 张为测试图像. 这些图像采集于自然场景,图像尺寸从 350 × 200 到 3888 × 2592 大小不等,且受到不同程度的光照、阴影、遮挡、拍摄角度等因素影响.

根据本文提出的候选字符提取方法,提取训练图像中的候选字符,将包含文本的区域视为正样本,反之为负样本,由此得到 37100 个正样本和 39000 个负样本,并根据 2.3 节 Step1 和 Step2 的方法提取特征向量,分别训练基于几何特征和基于 HOG 特征的 Adaboost 分类器.

3.2　评估标准

ICDAR 举办的文本检测和识别竞赛提供了通用的文本检测评估标准以及在线评估系统. 文本检测评估标准包括三个指标:召回率(Recall,R)、准确率(Precision,P)和 F 得分(F-score,F),其分别定义如下:

$$R = \frac{\sum_{i=1}^{|G|} \mathrm{Match}_G(G_i, D, t_r, t_p)}{|G|} \tag{11}$$

$$P = \frac{\sum_{j=1}^{|D|} \mathrm{Match}_D(D_j, G, t_r, t_p)}{|D|} \tag{12}$$

$$F = \frac{2 \times R \times P}{R + P} \tag{}$$

其中,D 和 G 分别表示检测到的文本区域和真实文本区域. t_r 和 t_p 用来控制 D 与 G 的匹配程度,在 ICDAR 竞赛中分别设定为 0.8 和 0.4. Match 为检测文本区域 D 和真实文本区域 G 的匹配函数,主要分为一对一、多对一和不匹配等类型,针对上述匹配类型分别设定的分数为 1、0.8、0.

本文采用 ICDAR 在线评估系统对本文文本检测结果进行定量评估. 此外,由于 ICDAR 在线评估标准主要基于单词级别评估文本检测结果,为了基于字符级别评估本文提出的候选字符提取和候选字符验证方法,本文采用文献[7,14]使用的字符级别评估标准. 主要包含两项指标:平均候选字符提取数目和字符检测率(Detection Rate,DR). 其中,DR 的定义如下:

$$\mathrm{DR} = \frac{\sum_i \sum_{j=1}^{|G_i|} \max_{k=1}^{|D_i|} \mathrm{match}(G_i^{(j)}, D_i^{(k)})}{\sum_i |G_i|} \tag{14}$$

其中,match 为候选字符与真实字符之间的匹配函数,当候选字符与真实字符相交面积大于真实字符面积 80% 以上,且最大高度与最小高度之比小于 1.5h,取值为 1,否则为 0.

3.3　候选字符提取性能分析

为了验证本文候选字符提取方法的有效性,将本文基于自适应色彩聚类方法与 MSER 方法[8]、SWT 方法[13]在 ICDAR2013 数据集上,对候选字符提取性能进行定量比较,结果如表 1 所示.

表 1　ICDAR2013 数据集上不同方法字符检测性能

方法	平均候选字符个数	字符检测率
MSER(Gray + RGB)[8]	4835	0.94
SWT[13]	224	0.77
Ours	1410	0.89

从表 1 中可以看出,MSER 方法在 R、G、B 和 Gray 的 4 个通道上字符检测率为 0.94,平均每张图像提取候选字符的个数为 4835[14]. 平均每幅图像产生过多的候选字符,这是因为 MSER 方法将图像中稳定极值区域全部提取,视为候选字符,由此引入大量的非文本字符和重叠区域. SWT 方法每张图像平均提取 224 个候选字符,但是图像字符检测率只有 0.77[14]. 本文自适应色彩聚类方法的字符检测率为 0.89,接近 MSER 方法的字符检测率,远高于 SWT 方法的字符检测率. 并且,本文方法平均每张图像提取候选字符个数为 1410,远少于 MSER 方法提取的候选字符数量. 由此可以看出,本文字符提取方法优于现有方法.

3.4　候选字符过滤性能分析

将本文提出的基于上下文信息的字符验证策略(GEO + HOG)与单一的几何特征分类器(GEO)、梯度方向直方图特征分类器(HOG)及联合特征分类器(GEO-HOG)对候选字符验证进行定量评估.

表 2 展示了不同特征对 ICDAR2013 数据集上候选字符进行验证的结果. 从中可以得到,采用 GEO 特征验证候选字符,图像中平均候选字符个数从 1410 减少到 42,然而,字符检测率也降低为 0.74. 采用 HOG 特征验证候选字符,图像中平均候选字符个数和字符检测率分别为 70 和 0.76. 由此可以看出,使用单一特征(GEO 或 HOG)进行字符验证,得到的字符验证效果不理想,说明单一特征无法有效区分文本和非文本. 此外,将

GEO 和 HOG 特征结合验证候选字符,字符检测率和平均每幅图像包含的候选字符数量分别为 0.77 和 55,相对于单一特征的字符验证方法性能提升有限. 从表 2 中还可以看出,本文方法字符检测率为 0.79,平均每幅图像中候选字符个数为 39. 本文方法获得了较高的字符检测率,这是因为本文方法综合考虑了字符之间的上下文关系,能有效地将部分错误移除的文本字符召回. 同时,本文方法得到的候选字符个数较少,这是因为本文仅保留 GEO 和 HOG 特征分类器同时判定为正样本的字符,有效移除了大量非文本字符.

表 2 ICDAR2013 数据集不同分类器模型字符过滤性能

分类器	平均候选字符个数	字符检测率
GEO	42	0.74
HOG	70	0.76
GEO-HOG	55	0.77
GEO + HOG	39	0.79

进一步,将本文方法在 ICDAR2003 数据集上进行实验,分别获得 70.04% 的召回率、77.26% 的准确率和 73.47% 的 F 得分. 与 Du 等[20]提出的基于上下文信息的文本检测方法相比,本文方法的文本检测性能取得了较大的提升. 这是因为本文方法采用混合分类器策略验证候选字符,能获得较高的字符准确率,同时利用文本行中字符色彩、位置接近等上下文信息,召回错误移除的文本字符,保证较高的字符召回率.

3.5 文本检测性能分析

图 5 展示了本文方法在 ICDAR2013 公共数据集上文本检测的部分结果. 从图 5(a)中可以看出,对于正常图像,无论是只包含单个字符的文本或者倾斜的文本都能被正确的检测. 图 5(b)展示的是光照不均或文本模糊的图像,从检测结果中可以发现本文方法也能取得不错的文本检测结果,说明本文方法对光照和模糊等干扰因素的鲁棒性. 图 5(c)中图像文本较小,文本边缘与背景色彩差异不明显,且类似于文本的树叶、砖头和指示标较多,从检测的结果可以看出,本文方法不仅能准确检测出文本,而且能有效移除非文本,说明本文提出的自适应色彩聚类方法和基于上下文信息的字符验证策略具有一定的有效性.

将本文方法与最新的文本检测方法在 ICDAR2013 数据集上进行定量比较,结果如表 3 所示. 从表中可以看出,本文方法取得了较高召回率 R、准确率 P 和 F,分别为 74.17%、83.40% 和 78.52%,优于其它文本检测方法. 这是因为本文提出的自适应色彩聚类方法,将层次聚类和参数自学习策略相结合,能够根据图像自身信息,自动学习权重向量,使得聚类效果达到最佳,进而能从复杂背景中准确提取文本字符,获得较高的字符召回率. 此外,考虑到自然场景中的文本通常包含多

(a)

(b)

(c)

图5 ICDAR 2013 数据集部分图像文本检测结果

个字符,本文提出了基于上下文信息的字符验证策略,综合利用字符的几何和梯度特征的同时,结合连通域的空间位置信息,召回错误移除的文本字符,使得本文方法在保证较高字符召回率的前提下,极大可能的移除非文本字符. 与文献[15]提出的基于色彩聚类的文本检测方法相比,本文方法的召回率 R、准确率 P 和 F-score 三个指标分别提高了 9%、5% 和 7%. 相比而言,本文获得了较高的文本检测性能,这是因为文献[15]通过人为设定色彩聚类阈值,对图像中像素点进行聚类,提取候选字符,使得该方法对复杂背景图像的文本检测效果不佳,而本文方法针对图像不同的复杂度,自动学习聚类参数,使得本文的泛化性较好.

表 3 ICDAR2013 不同方法文本检测性能

方法	Year	R(%)	P(%)	F(%)
HUST_MCLAB[7]	2015	74.28	87.68	80.43
Ours	–	74.17	83.40	78.52
Wang et al.[10]	2015	73.86	80.34	76.96
H. Wu[14]	2016	70.00	84.00	76.00
USTB_TexStar[9]	2014	66.45	88.47	75.89
Text_Spotter[8]	2013	64.84	87.51	74.49
Yin et al.[21]	2015	65.11	83.98	73.35
CASIA_NLPR[22]	2014	68.24	78.89	73.18
H. Wu[15]	2015	65.00	78.00	71.00
I2R_NUS	2014	66.17	72.54	69.21
Baseline	2013	35.00	61.00	45.00

此外,为进一步说明本文方法的有效性,将本文方

法在 MSRA-TD500[23] 中英文混合数据集上进行试验,其部分文本检测结果如图 6 所示. 从图中可以看出,本文方法不仅能准确检测英文字符,对中文文本也有一定的检测效果.

图6　中英文混合字符检测结果

4　结论

本文提出一种自然场景文本检测方法,该方法主要针对自然场景文本检测课题中候选字符提取和过滤两项问题,提出了相应的解决方案. 首先,对于候选字符提取,考虑自然场景图像的复杂度较高,本文将层次聚类和参数自学习策略相结合,提出了一种自适应色彩聚类方法,提取图像中的候选字符. 该方法能准确提取不同复杂程度图像中的文本候选区域,与现有的字符提取方法(MSER 和 SWT)相比,本文方法能提取较完整的文本信息,同时获得较少的背景连通域. 然后,对于候选字符过滤,本文考虑图像文本行中通常包含多个字符,提出了一种基于上下文信息的字符验证策略. 该策略采用基于几何特征和 HOG 特征的分类器,分别验证候选字符,能有效移除背景连通域,同时通过分析候选连通域的空间位置信息,召回错误移除的字符,提高了字符验证的准确率. 本文方法在公共数据集 ICDAR2013 上获得了较好的文本检测性能,文本检测的召回率、准确率和 F-score 分别为 74.17%,83.40% 和 78.52%,与现有文本检测相比,具有一定的优越性,此外,对于难度较大的自然场景图像,也有较好的文本检测效果.

本文方法重点考虑水平方向的文本检测,对倾斜度较大的文本检测效果有待提升,在未来的工作中,将针对任意方向的文本检测方法展开研究.

参考文献

[1] Wei B, Yin Z, Jie Y, et al. A novel approach to text detection and extraction from videos by discriminative features and density[J]. Chinese Journal of Electronics, 2014, 23(2): 322 – 328.

[2] Bai X, Yao C, Liu W. Strokelets: A learned multi-scale mid-level representation for scene text recognition[J]. IEEE Transactions on Image Processing, 2016, 25(6): 2789 – 2802.

[3] Wang X, Zha T, Wu C, et al. Text semantics based automatic summarization for chinese videos[J]. Chinese Journal of Electronics, 2015, 24(3): 462 – 467.

[4] 袁海东,马华东,黄晓冬. 基于梯度与粗糙度的视频文本检测与定位[J]. 电子学报,2008,36(8):1660 – 1664.
H. Yuan, H. Ma abd X. Huang. Video text detection and localization based on gradients and coarsenss[J]. Acta Electronica Sinica, 2008, 36(8): 1660 – 1665. (in Chinese)

[5] Lee J J, Lee P H, Lee S W, et al. AdaBoost for text detection in natural scene[A]. 2011 International Conference on Document Analysis and Recognition[C]. Beijing, China: IEEE, 2011. 429 – 434.

[6] Wang T, Wu D J, Coates A, et al. End-to-end text recognition with convolutional neural networks[A]. Proceedings of the 21st International Conference on Pattern Recognition[C]. Tsukuba, Japan: IEEE, 2012. 3304 – 3308.

[7] Zhang Z, Wei S, Yao C, et al. Symmetry-based text line detection in natural scenes[A]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Boston, USA: IEEE, 2015. 2558 – 2567.

[8] Neumann L, Matas J. Real-time scene text localization and recognition[A]. 2012 IEEE Conference on Computer Vision and Pattern Recognition[C]. Providence, USA: IEEE, 2012. 3538 – 3545.

[9] Yin X C, Yin X, Huang K, et al. Robust text detection in natural scene images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(5): 970 – 983.

[10] Wang Q, Lu Y, Sun S. Text detection in nature scene images using two-stage nontext filtering[A]. 2015 13th International Conference on Document Analysis and Recognition (ICDAR)[C]. Nancy, France: IEEE, 2015. 106 – 110.

[11] Yao C, Bai X, Liu W. A unified framework for multioriented text detection and recognition[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2014, 23(11): 4737 – 4749.

[12] Huang W, Lin Z, Yang J, et al. Text localization in natural images using stroke feature transform and text covariance descriptors[A]. 2013 IEEE International Conference on Computer Vision[C]. Sydney, Australia: IEEE, 2013. 1241 – 1248.

[13] Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform[A]. 2010 IEEE Com-

puter Society Conference on Computer Vision and Pattern Recognition[C]. San Francisco, USA: IEEE, 2010. 2963 – 2970.

[14] Wu H, Zou B, Zhao Y-Q, et al. Natural scene text detection by multi-scale adaptive color clustering and non-text filtering[J]. Neurocomputing, 2016, 214: 1011 – 1025.

[15] Wu H, Zou B, Zhao Y-Q, et al. Scene text detection using adaptive color reduction, adjacent character model and hybrid verification strategy[J]. The Visual Computer, 33(1): 1 – 14.

[16] Yi C, Tian Y. Text string detection from natural scenes by structure-based partition and grouping[J]. IEEE Transactions on Image Processing, 2011, 20(9): 2594 – 2605.

[17] Yang Q, Tang J, Ahuja N. Efficient and robust specular highlight removal[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(6): 1304 – 1311.

[18] Dalal N, Triggs B. Histograms of oriented gradients for human detection[A]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)[C]. San Diego, USA: IEEE, 2005. 886 – 893.

[19] Karatzas D, Shafait F, Uchida S, et al. ICDAR 2013 Robust Reading Competition[A]. 2013 12th International Conference on Document Analysis and Recognition[C]. Washington DC, USA: IEEE, 2013. 1484 – 1493.

[20] Du Y, Duan G, Ai H. Context-based text detection in natural scenes[A]. IEEE International Conference on Image Processing[C]. Orlando, USA: IEEE, 2012. 1857 – 1860.

[21] Yin X C, Pei W Y, Zhang J, et al. Multi-orientation scene text detection with adaptive clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1930 – 1937.

[22] Shi C Z, Wang C H, Xiao B H, et al. Scene text recognition using structure-guided character detection and linguistic knowledge[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(7): 1235 – 1250.

[23] Yao C, Bai X, Liu W, et al. Detecting texts of arbitrary orientations in natural images[A]. 2012 IEEE Conference on Computer Vision and Pattern Recognition[C]. Providence, USA: IEEE, 2012. 1083 – 1090.

作者简介

邹北骥 男, 1961 年生于湖南邵阳. 中南大学教授, 博士生导师, 研究方向为图像处理与计算机视觉.

郭建京 男, 1990 年生于湖南祁阳. 硕士研究生, 研究方向为计算机视觉与模式识别.

朱承璋(通信作者) 女, 1979 年出生, 湖南衡阳人, 博士. 中南大学副教授, 研究方向为机器学习与图像处理.

E-mail: Anandawork@126.com