

引用格式:方承志,樊梦雅. 基于改进型卷积神经网络和行特征的文本检测 [J]. 微电子学与计算机,2019,36(8):77-82.
[FANG C Z,FAN M Y. Text detection based on improved convolutional neural network and the feature of text lines [J]. Microelectronics & Computer,2019,36(8): 77-82.]

基于改进型卷积神经网络和行特征的文本检测

方承志,樊梦雅

(南京邮电大学 电子与光学工程学院,江苏 南京 210023)

摘 要: 针对现有的自然场景文本检测算法准确率尚未理想的问题,提出了一种基于改进型卷积神经网络和行特征的文本检测方法. 首先,采用增强的最大稳定极值区域(MSER)提取图像的连通分量,并应用剪枝方法来获取孤立的连通区域;其次,应用改进型卷积神经网络(CNN)对非字符区域进行消除,获得候选字符区域;然后,提出基于行特征构建多方向候选文本行的算法,用于检测任意定向和弯曲的场景文本;最后,应用 C4. 5 决策树算法对候选文本行进行分类. 该算法在 ICDAR2013、ICDAR2015 和 MSER-TD500 数据集上进行实验,实验结果表明,该算法能显著提高自然场景文本检测的准确率和召回率,且适用于任意方向、语言和字体的文本.

关键词: 文本检测;最大稳定极值区域;卷积神经网络;行特征;C4. 5 决策树算法

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2019)08—0077—06

Text detection based on improved convolutional neural network and the feature of text lines

FANG Cheng-zhi,FAN Meng-ya

(College of Electronic and Optical Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210023,China)

Abstract: In order to solve the problem of low precision in existing natural scene text detection algorithm, a text detection algorithm based on improved convolutional neural network and the feature of text lines was proposed. Firstly, the enhanced maximally stable extreme region (MSER) was used to extract connected components of the image and isolated connected regions were obtained by pruning method. Secondly, an improved convolutional neural network (CNN) was used to eliminate non-character regions and obtain candidate character regions. Thirdly, an algorithm based on the feature of text lines for constructing multi-oriented candidate text lines was proposed to detect arbitrary oriented and curved scene text. Finally, C4. 5 decision tree algorithm was applied to classify candidate text lines. Experiments were carried out on ICDAR2013, ICDAR2015 and MSRA-TD500 datasets. Experimental results show that the algorithm can significantly improve the precision and recall of text detection in natural scenes, and is suitable for text in any direction, language and font.

Key words: text detection; maximally stable extremal region (MSER); convolutional neural network (CNN); the feature of text lines; C4. 5 decision tree algorithm

1 引言

尽管在自然场景中的文本检测方面进行了大量

的研究工作,但依然面临着诸多挑战,主要有以下三个方面:(1)文本的多样性;(2)背景的复杂性;(3)光照不均、扭曲、部分遮挡等其他外界因素的干扰.

目前的自然场景文本检测算法主要分为两类：基于滑动窗口的方法和基于连通区域的方法. 基于滑动窗口的方法^[1-3]是用多尺度窗口扫描图像中的所有可能的位置,提取窗口内的特征,使用一个分类器对其进行分类. 目前主流的基于连通区域的方法^[4-6]在速度上展现了很大的优势,基本原理是采用一个底层的滤波器快速分割文本像素和非文本像素,然后将具有相似属性的文本像素连通起来构成连通区域,最后再使用一个分类器对其分类. 最近几年,MSER(Maximally Stable Extreme Regions)^[7]由于尺度旋转不变性、仿射不变性的优势被广泛地应用于文本检测,成为最常用的连通区域方法,其复杂度为 $O(2N\log\log N)$ ^[8],其中 N 为原图像的像素总数.

上述两种方法都需要一个强大的文本分类器对滑动窗口或者连通分量进行文本和非文本的判定. 人工设计的特征的低层属性很难准确地辨别有挑战的分量,比如模糊的文本模式和复杂的背景异常值. CNN(Convolutional Neural Network)^[9]可以避免显式地提取低层特征,隐式地从训练数据中学习文本的高层特征,极大地减少了复杂背景的干扰. 但是如果用 CNN 生成候选区域,却有两个不可忽视的缺点:(1)产生太多的冗余窗口,增加了算法的时间复杂度;(2)忽略了文字本身的低层特征,而且滑动窗口的长宽比一般都是固定的设置几个,对于长宽比浮动较大的文本,无法得到形状很吻合文本的区域. 因此,不同于文献^[10]中直接采用 CNN 得到候选区域的做法,本文结合 MSER 和 CNN 各自的优点,采用基于 CNN 的分类器过滤 MSER 检测出来的连通分量,得到候选字符区域. 受到文献^[11]中有相似笔画宽度的单个字符可以形成一串的启发,本文从文本行的角度进一步提取特征,采用基于行特征的候选字符合并算法,构建多方向的候选文本行. 考虑到复杂背景所带来的干扰,为了进一步减少虚警率,本文采用 C4.5 决策树算法对候选文本行进行文本行和非文本行的判定.

2 本文的算法设计

本文提出的基于改进型卷积神经网络和行特征的文本检测方法具体流程图如图 1 所示.

2.1 基于增强的 MSER 的连通区域提取与剪枝

MSER 一直被认为是最佳的区域检测算子. 但是,传统的 MSER 算法对噪声、模糊比较敏感. 主要存在以下问题:由于图像模糊,提取的字符是相连

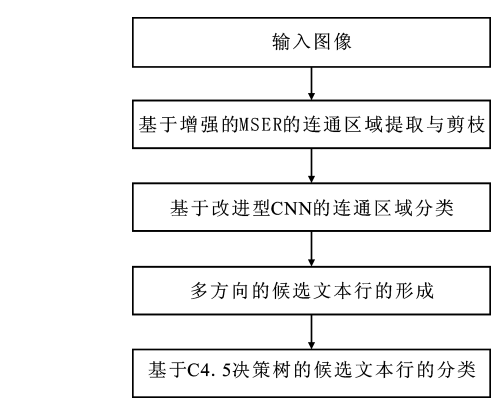


图 1 文本定位流程图

的;字符中间的孔洞没有被剔除;自然场景中包含有字符阴影的文本. 为了增强图像前景和背景的对比度,本文将 Canny 边缘检测算子与 MSER 相结合,得到边缘增强的 MSER 区域^[11]. 首先利用 Canny 算子得到边缘图像并向梯度方向膨胀,得到增强的二值图像,然后将这个二值图像取反,与 MSER 二值图像相与,得到增强后的连通区域. 增强后的 MSER 可以分割因模糊而相连的字符像素,以及字符连通域的孔洞,删除了部分不含文本的极值区域,提高候选字符区域的准确率.

经过 MSER 后,得到的区域会出现嵌套的情况,即最大区域可能包含文本区域,也可能包含非文本区域,甚至可能会导致不同的连通分量相交形成一个大的连通分量,这无疑会增加后续检测过程的计算量. 因为文本区域与背景区域具有显著差异,孤立的对象在文本检测中具有良好的性能,所以在实际操作中,字符不能被其他字符包含或者被包含. 针对这种情况,本文对此进行剪枝操作. 该剪枝算法可以用文献^[12]中的结合平滑度的剪枝操作来修剪 MSER 树,从而可以得到孤立的 MSER 区域.

经过剪枝之后,每一个像素仅属于一个连通区域,这不仅保证了得到的连通区域的独立性,而且不删除一些较小的区域,最大限度的保证了文本的完整性,为后续文本检测提供更加有效的文本连通区域,提高了连通区域分类的准确率,减少了计算复杂度.

2.2 基于改进型 CNN 的连通区域分类

对于一幅给定的自然场景图像,在使用增强的 MSER 和剪枝算法进行区域检测以后,会得到大量的连通体区域,其中包括字符区域和背景区域,因此,需要过滤背景区域,保留文本区域. CNN 检测器表现着良好的性能,本文应用改进型 CNN 过滤大量的背景区域,得到候选字符区域. 主要的流程如图

2 所示。

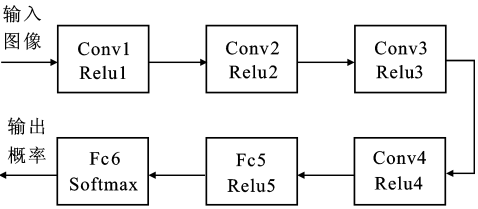


图 2 基于改进型 CNN 的连通区域分类流程图

本文的 CNN 模型共由 6 层构成,其中,前 4 层为卷积层,最后 2 层为全连接层。输入为 32×32 像素的灰度图像。首先,输入图像经过四个特征提取阶段逐层提取得到高层特征表示。在第一个卷积层 C1 中,一共使用了 4 个不同的卷积核,其大小为 5×5 。第二个卷积层 C2 采用了 16 个大小为 3×3 的卷积核。第三个卷积层 C3 和第四个卷积层 C4 的卷积核个数分别为 32 和 64,其大小分别为 3×3 和 2×2 。第五层是个全连接层,神经元数设置为 600。网络的前 5 层后面都跟随一个修正线性单元(ReLU)作为激活函数,目的是给网络引入非线性特征。第六层是网络的输出层,该层也是一个全连接层,采用 Softmax 作为最后输出层的激活函数。Softmax 函数也称为归一化函数,其输出表征了不同类别之间的相对概率,而不是分类标签,其优势主要有两点:(1)衡量分类问题不再是单一的 0 或 1,而是属于 $[0,1]$ 区间的概率值。概率本质上是置信度的反映,丰富了输出结果,从而更有利于理解输入之间彼此的不同。(2)基于不同的场景,可以设置不同的概率输出阈值来改变预测的标签,这样更有利于满足不同的实际应用需求。一般情况下,50%常作为阈值点。但是,实际上会有一些受复杂背景、光照、遮挡等因素影响的字符,其字符概率可能只有 0.49、0.48 等,十分接近分界线 0.5,如果将其判为背景区域,漏检的情况将会较多。因此,为了提高字符检测的召回率,本文将字符概率大于等于 0.4 的连通区域判决为候选字符区域,反之为背景区域。虽然准确率略有下降,但是后续基于行特征的强大的候选字符符合能力,会显著提高文本检测的准确率,足以弥补此时准确率下降的不足。

传统的卷积层^[13]只有两层,这种比较浅的网络只能学习到图像的一些边、角等基本特征,且检测的准确率不高,具有很大的改进空间。网络的深度逐渐加深,能够获得更高的分类精度,但也增加了网络的整体复杂程度,训练复杂度会随着深度增加而急剧

增加,使得网络变得难以优化,很容易过拟合。为了平衡分类准确率和计算复杂度之间的关系,本文进行了大量的实验对比,最终确定了四层卷积层,其足以提取字符和背景的深层特征,同时不会增加太多计算量。而且四层卷积层的设计拥有更多的非线性变换,对特征的学习能力更强。

2.3 多方向的候选文本行的形成

在对连通区域进行分类后,可获得大量的候选字符区域。由于上述过程只考虑了单个字符区域与背景区域的差异,因此,除了得到真正的字符外,还可能会存在一部分和字符很相似的背景区域(如:砖块、窗户等)。但是,真正的字符区域通常是按照一定的规律出现的,即同一文本行的字符具有相似性,如相似的颜色,相似的笔画宽度^[14]等。本文将根据上述规律,在得到的候选字符区域图上进行扫描,对每一个候选字符区域进行特征提取,在其周围寻找任意方向的候选字符区域。如果这两个候选字符具有相似的特征,则可以将其分成一组,进而形成一串。这样就可以有效地将候选字符整合成候选文本行,同时可以进一步滤除大量孤立的类似字符的背景区域,大大地提高了文本检测的准确率。从文本行的角度,而不是单个字符的角度提取特征,另一个好处是,判为背景的字符一样会被周围判为候选字符区域的字符拉入候选文本行中,提高了文本检测的召回率。

如果设置的特征太少、过于简单,那么在某一受限场景下可以达到很好的检测效果,但是鲁棒性较差,因此,本文采用大量特征来进行综合判定。具体特征设置如下:

- (1) 宽度比 $R_w = \max(W_A, W_B) / \min(W_A, W_B)$
- (2) 高度比 $R_h = \max(H_A, H_B) / \min(H_A, H_B)$
- (3) 面积比 $R_s = \max(S_A, S_B) / \min(S_A, S_B)$
- (4) 宽高比比值 $R_q = \max(Q_A, Q_B) / \min(Q_A, Q_B)$
- (5) 占空比比值 $R_{occ} = \max(OC_A, OC_B) / \min(OC_A, OC_B)$
- (6) 紧致度比 $R_{ed} = \max(ED_A, ED_B) / \min(ED_A, ED_B)$
- (7) 颜色比 $R_y = \max(Y_A, Y_B) / \min(Y_A, Y_B)$
- (8) 笔画宽度比 $R_{sw} = \max(SW_A, SW_B) / \min(SW_A, SW_B)$

候选字符整合成候选文本行以后,图像中可能会有少量孤立存在的字符被滤除,但是字符概率极

大,为了避免这种情况,本文也将其保留下来.

2.4 基于 C4.5 决策树的候选文本行的分类

至此,已经得到图像中的候选文本行,但是得到的候选文本行中有可能还有非文本行的存在,为了得到更加精确的文本行区域,要对非文本行进行剔除,根据机器学习算法分类具有更强的鲁棒性,本文采用 C4.5 决策树^[15]算法进一步过滤候选文本行.但是,机器学习算法对训练集有较高的依赖度,因此,本文通过借助不同类特征提取,可以克服机器学习算法对训练集的依赖度,降低漏检率和误检率,增强场景鲁棒性.图 3 给出了 C4.5 决策树算法的基本流程.

Algorithm C4.5(D)

Input: an feature-valued dataset D, feature

Output: Tree

```
1: Tree={}
2: if D is "pure" or stopping criteria met:
3:   terminate
4: for every feature a∈ D:
5:   sort a
6:   for every split point b:
7:     Compute GainRatio
8: c= max GainRatio
9: d, e=Best feature & best split point according to c
10: Tree=Create a decision node that tests d in the root
11: D1=Induced sub-datasets from D based on d & e
12: for every D1:
13:   Tree1=C4.5(D1)
14: Attach Tree1 to the corresponding branch of Tree
```

图 3 C4.5 决策树算法流程

根据同一文本行中的字符具有相似性,比如相似的颜色、相似的笔画宽度等,本文从文本行的角度进一步提取特征,计算同一文本行相邻候选字符相同类别特征的平均值,利用 C4.5 决策树算法对候选文本行进行文本行和非文本行区域分类,能够精确地将非文本行滤除,确定出最终的文本行.本文选用的特征与上一步多方向的候选文本行的形成所选用的特征类别相同,具体特征设置如下:

- (1)宽度比的平均值;
- (2)高度比的平均值;
- (3)面积比的平均值;
- (4)宽高比比值的平均值;
- (5)占空比比值的平均值;
- (6)紧致度比的平均值;
- (7)颜色比的平均值;

(8)笔画宽度比的平均值.

3 实验结果与分析

本文使用了 ICDAR2013、ICDAR2015 和 MSRA-TD500 数据集,分别进行三组实验.为了将该算法与文献[13]的算法进行比较,第一组实验使用了 ICDAR2013 数据集进行评估.ICDAR2013 数据集是最基本的文本检测数据集,其中包含了 229 幅训练图像和 233 幅测试图像,图像中的文本呈水平方向排列,分辨率高,文字清晰可见.为了与文献[11]的算法进行比较,第二组实验使用了 MSRA-TD500 数据集进行评估.这是一个针对多方向、中英文混合的数据集,该数据集包含了 300 幅训练图像和 200 幅测试图像.为了更好地证明本算法的性能,第三组实验采用了 ICDAR2015-偶然场景文本数据集,其包含了 1000 幅训练图像和 500 幅测试图像,该数据集是用穿戴设备随意拍摄的生活场景,不是刻意针对文字拍摄的,图像会出现模糊、遮挡等情况,文字通常是多方向的,因此,检测难度最大.在这组实验中,同样与文献[11]的算法进行比较.本文遵循了数据集各自的标准性能评估准则,来测量所提出方法的性能.所有数据集的最终衡量参数包括准确率 P、召回率 R 和 F 综合指标,定义为:

$$P=C/(C+E) \tag{1}$$

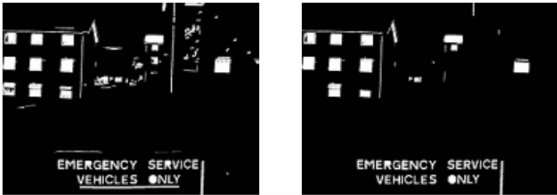
$$R=C/(C+T) \tag{2}$$

$$F=1/(\alpha/P+(1-\alpha)/R) \tag{3}$$

式中,C 为正确检测到的文本区域的个数;E 为错误检测到的文本区域个数;T 为未检测到的文本区域个数;F 是准确率和召回率之间的调和平均数; α 一般设置成 0.5.

算法的中间结果显示在图 4 上.图 4(a)展现了增强的 MSER 和剪枝算法的输出.在图 4(b)上可以看到,许多背景区域都被滤除了,如树叶.所有的文本区域都保留下来了,还有像窗户这样类似字符的区域保留下来了,这表明了本文改进型 CNN 模型是高度可信的.图 4(c)展现了形成的候选文本行.孤立的背景区域没有形成文本行,因为其特征和周围候选字符区域的特征是不相似的.最终的结果显示在图 4(d)中.经过 C4.5 决策树算法,所有的非文本区域被过滤,所有的文本行都保留下来了.

表 1 所示为本文提出的方法与文献[13]的算法在 ICDAR2013 数据集上的实验对比.



(a)增强的 MSER 和剪枝 (b)改进型 CNN 分类



(c)候选文本行形成 (d)C4.5 决策树分类

图 4 中间结果
表 1 用 ICDAR2013 数据集评估

	准确率 P	召回率 R	F 值
本文算法	0.91	0.76	0.83
[13]的算法	0.83	0.71	0.77

从表 1 可以看出,本文的算法在处理水平方向的文本的 P,R 和 F 值分别为 0.91,0.76,0.83,明显优于文献[13],准确率比文献[13]提高了 0.08,主要原因有以下三点:(1)文献[13]进行连通区域分类时,卷积层只有两层,因此,只能提取一些低层次特征,字符的准确率欠佳;而本文采用了四层卷积层,可以提取更加抽象的特征,从而获得更为精确的候选字符。(2)文献[13]进行候选字符合并时,选取的特征太少,过于简单,合并效果不是很理想;而本文提取了八个不同类型的特征,大大地减少了那些类似字符的背景串成文本行的可能性,从而增强了算法的鲁棒性。(3)文献[13]没有用机器学习算法进行候选文本行分类,因此,文本行的准确率欠佳;而本文应用了对于分类问题更加鲁棒的 C4.5 决策树算法,能够有效地将类似文本行的背景区域过滤掉,从而得到准确率更高的文本行。综上所述,本文两级分类器出色的分类性能,以及基于行特征的强大的候选字符合并能力,将非文本区域几乎全部过滤掉,极大地提高了文本检测的准确率,证明了本文所提算法的有效性。

为了验证本文算法的鲁棒性和泛化能力,本文在 MSRA-TD500 数据集上做了进一步实验。表 2 所示为本文提出的自然场景文本检测方法 with 文献[11]的算法在 MSRA-TD500 数据集上的实验对比。

表 2 用 MSRA-TD500 数据集评估

	准确率 P	召回率 R	F 值
本文算法	0.86	0.74	0.80
[11]的算法	0.79	0.72	0.75

从表 2 可以看出,在执行多方向文本时与文献[11]相比也有较好的效果,实验结果说明了本文的方法适用于任意方向的文本,且准确率非常高,召回率也有一些提高,验证了所提算法的有效性。

为了进一步验证算法的有效性,本文在 ICDAR2015 数据集上进行实验,并与文献[11]的算法做对比,结果显示在表 3 中。

表 3 用 ICDAR2015 数据集评估

	准确率 P	召回率 R	F 值
本文算法	0.81	0.63	0.71
[11]的算法	0.72	0.58	0.64

从表 3 可以看出,本文的方法相比文献[11]的算法,准确率提高了 9%,具有明显的优势。因此,算法在处理模糊、透视变形、极端光照、文本字体太小等不利因素影响的多方向文本图像,也有较好的效果,进一步验证了本文算法的有效性。

4 结束语

本文提出了一种基于改进型卷积神经网络和行特征的文本检测算法。应用了增强的 MSER 和剪枝算法得到连通区域。相比以往采用人工设计的特征或浅层网络提取特征,本文使用基于改进型 CNN 的深层特征对连通区域分类,本文从文本行的角度进一步提取特征,采用基于行特征的候选字符合并算法,构建准确率和召回率更高的多方向候选文本行。最后根据机器学习算法分类具有更强的鲁棒性,本文采用 C4.5 决策树算法进行候选文本行的分类和定位,从而提高了最终文本行的准确率。实验结果表明,本文提出的算法准确率非常高,并且可以有效地检测到图像中任意方向的文本,对文本的语言、字体等也有较强的鲁棒性。

参考文献:

[1] CHEN X, YUILLE A L. Detecting and reading text in natural scenes[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2004:366-373.
[2] LEE J J, LEE P H, LEE S W, et al. AdaBoost for Text Detection in Natural Scene[C]// International Conference on Document Analysis and Recognition. IEEE,

- 2011:429-434.
- [3] WANG K,BELONGIE S. Word Spotting in the Wild [C]// European Conference on Computer Vision. Springer-Verlag,2010:591-604.
- [4] BAI B,YIN F,LIU C L. A Fast Stroke-Based Method for Text Detection in Video[C]// Iapr International Workshop on Document Analysis Systems. IEEE, 2012:69-73.
- [5] LIU X,WANG W. Robustly Extracting Captions in Videos Based on Stroke-Like Edges and Spatio-Temporal Analysis[M]. IEEE Press,2012.
- [6] ZHAO Y,LU T,LIAO W. A Robust Color-Independent Text Detection Method from Complex Videos [C]// International Conference on Document Analysis and Recognition. IEEE Computer Society, 2011: 374-378.
- [7] 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.
- [8] 范一华,邓德祥,颜佳. 基于色彩空间的极大稳定极值区域的自然场景文本检测[J]. 计算机应用,2018,38 (1):264-269.
- [9] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.
- [10] 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, 2017, 39 (6): 1137-1149.
- [11] 方承志,黄梅玲. 自然场景中多方向文本的检测[J]. 计算机工程与设计,2018,39(5):1377-1381.
- [12] 方承志,田彪. 结合连通分量规则度与 Adaboost 的文本定位算法[J]. 电视技术,2015,39(21):1-4.
- [13] REN X,CHEN K,YANG X, et al. A novel scene text detection algorithm based on convolutional neural network[C]// Visual Communications and Image Processing (VCIP),2016. IEEE,2016: 1-4.
- [14] 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.
- [15] 田彪. 自然场景图像中的文字检测关键算法研究[D]. 南京:南京邮电大学,2016.

作者简介:

方承志 男,(1976-),博士,副教授. 研究方向为模式识别、嵌入式系统和信号处理.

樊梦雅(通讯作者) 女,(1994-),硕士研究生. 研究方向为图像处理与模式识别. E-mail:956594122@qq.com.