计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(6):231−235 [doi: 10.15888/j.cnki.csa.006395]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

卷积深度置信网络的场景文本检测①

王　林, 张晓锋

(西安理工大学 自动化与信息工程学院, 西安 710048)
通讯作者: 张晓锋, E-mail: 792094891@qq.com

摘　要: 自然场景中的文本检测对于视频、图像和图片等海量信息的检索管理具有重要意义. 针对自然场景中的文本检测面临着图像背景复杂、分辨率低和分布随意的问题, 提出一种场景文本检测的方法. 该方法将最大稳定极值区域算法与卷积深度置信网络进行结合, 把从最大稳定极值区域中提取出来的候选文本区域输入到卷积深度置信网络中进行特征提取, 由 Softmax 分类器对提取的特征进行分类. 该方法在 ICDAR 数据集和 SVT 数据集上进行实验, 实验结果表明该方法有助于提高场景文本检测的精确率及召回率.

关键词: 场景文本检测; 特征提取; 候选文本区域; 最大稳定极值区域算法; 卷积深度置信网络

引用格式: 王林,张晓锋.卷积深度置信网络的场景文本检测.计算机系统应用,2018,27(6):231–235. http://www.c-s-a.org.cn/1003-3254/6395.html

Scene Text Detection in Convolutional Deep Belief Networks

WANG Lin, ZHANG Xiao-Feng

(College of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: Text detection in the natural scenes is of great significance to the retrieval and management of large amounts of information such as video, images, and pictures. Depending on the complex background, low resolution and random distribution of the text detection in natural scenes, a scene text detection method was proposed, which combined the maximum stable extremal region algorithm and convolutional deep belief networks. In this method, candidate text region extracted from the maximally stable extremal region entered into the convolutional deep belief network for feature extraction. Then these features were classified by Softmax classifier. Experiments were carried out on ICDAR datasets and SVT datasets, and the experiment results show that the proposed method is helpful for improving the precision and recall rate of scene text detection.

Key words: scene text detection; feature extraction; candidate text region; maximum stable extremum region algorithm; convolutional deep belief networks

随着智能硬件的普及, 通过手机、平板和数码相机等移动可穿戴设备的终端摄像头获取、处理和分享信息已经逐渐成为客观的发展趋势. 自然场景中的文本检测是检测图像中是否含有文本信息, 并确定文本信息的位置. 通过文本信息来对场景进行理解, 将有助于我们对日夜增加的视频、图像和图片等海量信息的检索管理等. 因此, 本文主要集中在检测自然场景中的文本信息.

目前, 自然场景中的文本检测有两种经典模型: 卷积神经网络[1]和深度置信网络[2], 卷积神经网络 (Convolution Neural Network, CNN) 是一个多层的神经网络, 每层由个二维平面组成, 而每个平面又由多个独立的神经元组成. 卷积神经网络可以看成是卷积层和子采样层两种结构交替连接而成的. 卷积神经网络对

① 基金项目: 陕西省科技计划重点项目 (2017ZDCXL-GY-05-03)
收稿时间: 2017-10-12; 修改时间: 2017-11-03; 采用时间: 2017-11-10; csa 在线出版时间: 2018-05-28

图像的位移、缩放及其他旋转等变化具有良好的适应性, 但是忽略了图像中的高阶统计特征. 相应地, 深度置信网络 (Deep Belief Network, DBN) 是一种由多个受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM) 叠加而成的深度学习结构, 两者的区别在与卷积的使用. 对于深度置信网络模型而言, 虽然它在提取图像高阶特征方面具有很好的性能, 但忽略了图像的局部不变性, 对外部变化较为敏感.

Lee 等人[3]提出了卷积深度置信网络 (Convolutional Deep Belief Network, CDBN), 该网络由卷积受限玻尔兹曼机 (Restricted Boltzmann Machine, CRBM) 为基础堆叠而成的, 是一个分层的概率生成模型. 该网络结合了深度置信网络在图像高阶特征方面具有的良好性能和卷积神经网络对图像的位移、缩放及其他旋转等变化具有很好的适应性, 解决了对于扩展图像到原尺寸, 以及图像特征会因输入局部变换而变换的问题. Huang[4]利用卷积深度置信网络模型 CDBN 和局部二进制模式 LBP 相结合所形成的深度学习方法, 更好的学习到高分辨率图像中的特征, 实验结果表明该方法在真实世界的人脸验证数据库上实现了最新的结果. Wicht[5]利用卷积深度置信网络模型 CDBN 识别包含手写和打印数字的数独拼图, 实验结果表明当考虑检测误差时, 识别精确率达到 92%; 当不考虑检测误差时, 识别精确率提高到 97.7%. 何灼彬[6]利用卷积深度置信网络模型 CDBN 进行歌手识别, 实验结果表明该模型在声音识别分类表现上具有一定的优势. Ren 等[7]提出利用卷积深度置信网络模型 CDBN 对脑电信号特征提取, 与其他提取方法相比, 利用卷积深度置信网络学习的特征具有更好的性能. 祝军[8]利用卷积深度置信网络模型 CDBN 进行场景图像分类识别, 实验结果表明该模型在场景图像分类识别中取得较好的效果.

综上所说, 卷积深度置信网络因结合了深度置信网络在图像高阶特征方面具有的良好性能和卷积神经网络对图像的位移、缩放及其他旋转等变化具有很好的适应性, 已广泛应用于图像分类、语音识别和人脸识别[9]等领域, 但是目前尚未发现有研究将卷积深度置信网络应用于自然场景中的文本检测领域. 因此, 本文考虑将卷积深度置信网络模型应用到自然场景中文本检测中, 旨在解决图像背景复杂、分辨率低和文本分布随意的问题, 从而提高文本检测的精确率以及召回率.

1 卷积深度置信网络

2011 年, Lee 提出了卷积深度置信网络 CDBN, 该卷积深度置信网络有多个卷积受限玻尔兹曼机 CRBM 堆叠而成, 这种结构的层与层之间引入了一种最新的操作, 即概率型最大池化 (Probabilistic Max-pooling)[3], 如图 1 所示. 一般而言, 要获取高层的特征描述需要更多的区域信息, 通过用最大值池化特征表示, 能够使得高层特征描述对输入的微小变化具有良好的不变性, 同时能够减少计算复杂度.

在本文中 CDBN 模型的输入层设置为 28×28×3 大小 (即将输入可以看成 3 个大小为 28×28 的映射层), 第一隐含层中的卷积层包含 6 个特征映射, 卷积核大小均为 7×7, 池化层的池化区域为 2×2, 第二个隐含层的卷积层包含 8 个特征映射, 卷积核的大小为 5×5, 池化层的池化区域为 2×2, 最后将模型的输出单元组合成长度为一维的向量. 学习速率为 0.05, 模型的激活函数采用 sigmoid 函数, 第一层的稀疏系数为 0.02, 第二层为 0.03. 采用 Dropout 方法对隐含层以 50% 的概率进行随机丢取. 最后的分类器采用 Softmax.

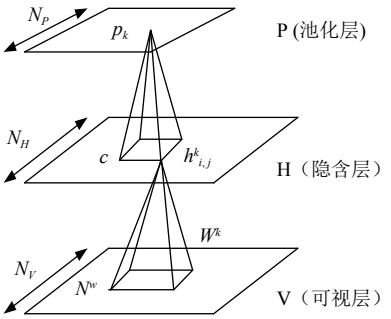

图 1 一个概率 max-pooling 卷积 CRBM 结构示意图

2 自然场景文本检测方法

一个典型的自然场景文本检测主要流程如图 2 所示, 简单描述自然场景文本检测的主要步骤[10–12]:

1) 最大稳定极值区域 (Maximally Stable Extremal Regions, MSERs)[13]文本定位: 假定同一个区域成分的某些相似特征 (颜色、亮度和笔划宽度的特征) 差别较大, 并且与背景的特征也存在较大区别的前提下, 采用自底向上的方法在图像中把连通成分作提取处理, 获取文本候选区域.

2) 预处理: 对最大稳定极值区域 MSER 提取的文本候选区域进行裁剪分割, 过滤掉一些很长很细的

MSER 区域 (很长很细的 MSER 区域不可能是文本区域), 把不规整的 MSER 区域统一规范成 28×28 的输入图像如图 3 所示, 并在整理好的 28×28 输入图像上添加 Ground truth 矩形框.

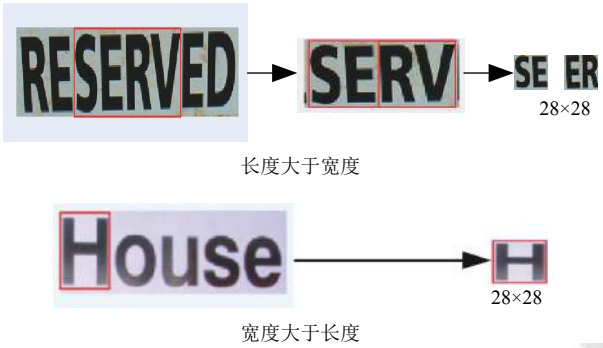

图 2 MSER 区域统一规范成 28×28 的输入图像

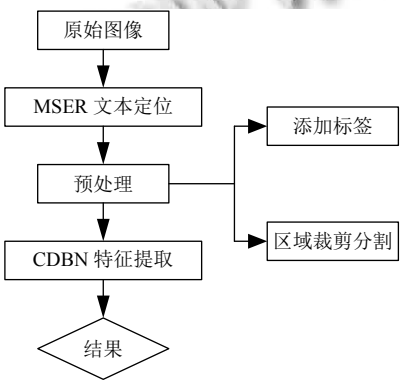

图 3 自然场景文本检测主要流程

3) CDBN 特征提取: 将从最大稳定极值区域 MSER 中提取出来的候选文本区域经过预处理后输入到卷积深度置信网络中进行训练, 从训练最大稳定极值区域数据中进行学习更多隐藏特征, 对候选文本区域进行验证, 进而过滤掉大量的非文本的 MSER 区域.

3 结果与分析

下面通过数值实验来验证本文所提出的场景文本检测方法性能, 将本文的方法和其他方法进行比较. 本文使用一些公开的自然场景文本检测的数据集, 包括 ICDAR2011 鲁棒阅读竞赛 (Robust Reading Competition) 数据集[14], 和街景 (Street View Text, SVT) 数据集[15]. 数据集中的图片是彩色的, 尺寸在 307×93 到 1280×960 内. 本实验的文本检测输出结果为单词级别的矩形框, 与数据集的 Ground truth 匹配. 对于文本检测任务而言, 有两个重要的评价指标[1]: 精确率 (使用 p 表示) 和召回率 (使用 r 表示). 其中 p 用来反映检测出的单词在 Ground truth 被标记的比例, 而 r 则用来表示 Ground truth 里标记的单词被检测出的比例 p 和 r 通过计算 Ground truth 矩形框和检测到的矩形框之间的差异得到.

3.1 开发与实验环境

硬件环境: 64 位 Intel(R)Core(TM)i7-4790 3.6 GHz CPU, 4 G RAM.

软件环境: Windows 8.1 旗舰版, Matlab R2016b.

本文实验在 Visual Studio 2013 和 Opencv 2.4.8 环境中进行了数据准备和在 Matlab R2016b 环境中进行了基于稀疏自动编码的文本检测.

3.2 数据集和评估方法

① ICDAR2011 数据集

ICDAR 2011 数据集包含 484 张图片, 其中训练集包 229 张 (848 个单词), 测试集包含 255 张 (1189 个单词, 6393 个字符). ICDAR 2011 数据集的评价协议考虑三种匹配情况: 一对一、一对多和没有匹配. 相应地, 其精确率和召回率的计算方式如下:

$$Precision = \frac{\sum_i^N \sum_j^{|D^i|} M_D\left(D_j^i, G^i\right)}{\sum_i^N |D^i|} \tag{1}$$

$$Recall = \frac{\sum_i^N \sum_j^{|G^i|} M_G\left(G_j^i, D^i\right)}{\sum_i^N |D^i|} \tag{2}$$

$$F-measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

其中, N 是数据集中图像的总数, $|D^i|$和$|G^i|$分别是第 i 个图像中的检测到矩形数和真实矩形数. $M_D\left(D_j^i, G^i\right)$和$M_G\left(G_j^i, D^i\right)$分别是检测矩形$D_j$和真实矩形$G_i$的匹配分数. 对于一对一匹配, 它们的值设置为 1, 对于一对多的匹配, 它们的值为 0.8, 不匹配的值为 0. 当它们的重叠比率高于定义的阈值时, 两个矩形认为是匹配的, 即重叠率越高, 检测率越高.

② SVT 数据集

SVT 数据集从 Google 街景中搜集的, 图像背景多为街道, 其中包含的文本信息主要是商业名称, 建筑名称等. 由于其图像是通过移动的车辆拍摄获得, 所以不可避免地会产生运动模糊以及形变, 而且图像的分辨率较低, 文本字体差异明显[16]. 共包含 350 张, 其中

101 张用作训练集 (257 个单词), 249 张用作测试集 (674 个单词, 3796 个字符). 对于 SVT 数据集, 使用与 ICDAR2011 数据集相同的评价协议.

3.3 实验和结果

① ICDAR 数据集实验结果

为了评价本文两个方法的有效性, 首先在 ICDAR 数据集上与其它较好的方法进行比较. 表 1 是在 ICDAR2011 数据集上的文本检测对比结果. 可以看到, MSER-CDBN 方法的精确率和召回率都取得改善, 提高了 1.45%-2.18% 并且 *F-measure* 分数超过了 78.63%. 由于 MSER-CDBN 使用了对复杂图像更加鲁棒的候选字符提取算法 MSER 和可以更好学习特征的 CDBN 模型, 因此识别精确率和召回率都得到提高.

表 1 ICDAR2011 数据集上实验对比结果

方法	精确率 (%)	召回率 (%)	*F-measure*(%)
MSER-CDBN	89.48	73.23	80.54
MSER-CNN[1]	88.03	71.05	78.63
文献[11]	86.32	67.46	75.73
Neumann and Matas[1]	85.42	67.39	75.34

为了提高模型检测精确率, 一个非常重要的策略就是引入随机噪声. 为了验证随机噪声引入与否的影响, 在其他条件不变的情况下, 引入随机噪声和不引入随机噪声的实验结果对比, 如表 2 所示.

表 2 ICDAR2011 数据集上引入随机噪声和不引入随机噪声的实验结果对比

数据集	精确率 (%)	
	不引入随机噪声	引入随机噪声
500	80.44	82.63
750	83.32	84.26
1000	84.45	86.92
1250	86.38	88.29
1500	87.42	89.46

由表 2 可知加入了噪声后的 CDBN 学习到的特征比较好, ICDAR2011 数据集上的精确率提高到了 89.49%, 可以看出随着训练次数的增加, 文本检测的精确率也在提高, 对于那些误判的文本进行归类发现很大一部分是由于复杂的背景造成的, 为此, 本文给输入数据加入噪声, 利用污染后的数据进行特征学习, 和原先的数据进行对比发现, 精确率有所提高. 图 4 显示了 MSER-CDBN 方法在 ICDAR2011 数据集上的部分检测结果.

② SVT 数据集实验结果

SVT 数据集比 ICDAR2011 数据集更为复杂, 拥有更多的字体变化, 而且图像常常包含大量的噪声信息. 在 SVT 数据集上对比方法比较少, 本文选择了两个代表性的方法用于对比实验. 这里需要注意的是下列方法均采用 ICDAR 2011 官方的评价协议.

图 4 MSER-CDBN 在 ICDAR2011 数据集上实验示例

可以从表 3 看到, 本文的方法相比 MSER-CNN 和 SWT-DBN 的方法, 精确率提高了 2.53%, 召回率与 SWT-DBN 方法相比提高了 9.24%, 主要得益于本文方法使用了深度学习模型, 将从最大稳定极值区域中提取出来的候选文本区域经过预处理后输入到卷积深度置信网络中进行训练, 从训练最大稳定极值区域数据中学习更多隐藏特征, 进而过滤掉大量的非文本的 MSER 区域. 图 5 显示了 MSER-CDBN 方法在 SVT 数据集上的部分识别结果.

表 3 SVT 数据集上实验对比结果

方法	精确率 (%)	召回率 (%)	*F-measure*(%)
MSER-CDBN	41.43	72.64	52.76
MSER-CNN[1]	38.90	74.03	51. 00
SWT-DBN[12]	34.12	60.08	43.52
文献[13]	67.05	29.12	40.60

4 结论与展望

由于卷积深度置信网络结合了深度置信网络在图像高阶特征方面具有的良好性能和卷积神经网络对图像的位移、缩放及其他旋转等变化具有很好的适应性,

本文将该模型和最大稳定极值区域算法相结合用于场景文本检测解决了图像背景复杂、分辨率低和分布随意的问题. 本文在 ICADR 和 SVT 数据集上进行实验, 结果表明与其它场景文本检测算法相比本文的算法在检测精确率和召回率上有了提高.

图 5 MSER-CDBN 在 SVT 数据集上的实验示例

参考文献

1 Huang WL, Qiao Y, Tang XO. Robust scene text detection with convolution neural network induced MSER trees. Computer Vision(ECCV 2014). Cham: Springer, 2014. 497–511.

2 Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform. Proceedings of Computer Vision and Pattern Recognition. San Francisco, CA, USA. 2010. 2963–2970.

3 Lee H, Grosse R, Ranganath R, *et al*. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, QC, Canada. 2009. 609–616.

4 Huang GB, Lee H, Learned-Miller E. Learning hierarchical representations for face verification with convolutional deep belief networks. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA. 2012. 2518–2525.

5 Wicht B, Henneberty J. Mixed handwritten and printed digit recognition in Sudoku with Convolutional Deep Belief Network. Proceedings of the 13th International Conference on Document Analysis and Recognition. Tunis, Tunisia. 2015. 861–865.

6 何灼彬. 基于卷积深度置信网络的歌手识别[硕士学位论文]. 广州: 华南理工大学, 2015. 38–48.

7 Ren YF, Wu Y. Convolutional deep belief networks for feature extraction of EEG signal. Proceedings of International Joint Conference on Neural Networks. Beijing, China. 2014. 2850–2853.

8 祝军, 赵杰煜, 董振宇. 融合显著信息的层次特征学习图像分类. 计算机研究与发展, 2014, 51(9): 1919–1928. [doi: 10.7544/issn1000-1239.2014.20140138]

9 Shao H, Chen S, Zhao JY, *et al*. Face recognition based on subset selection via metric learning on manifold. Frontiers of Information Technology & Electronic Engineering, 2015, 16(12): 1046–1058.

10 Yin XC, Yin XW, Huang KZ, *et al*. Robust text detection in natural scene images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(5): 970–983.

11 Xu HL, Xue LK, Su F. Scene text detection based on robust stroke width transform and deep belief network. Computer Vision-ACCV 2014. Cham: Springer, 2014. 195–209.

12 Wang K, Babenko B, Belongie S. End-to-end scene text recognition. Proceedings of International Conference on Computer Vision. Barcelona, Spain. 2012. 1457–1464.

13 Chen HZ, Tsai SS, Schroth G, *et al*. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. Proceedings of IEEE International Conference on Image Processing. Brussels, Belgium. 2011. 2609–2612.

14 Shahab A, Shafait F, Dengel A. ICDAR 2011 robust reading competition challenge 2: Reading text in scene images. Proceedings of International Conference on Document Analysis and Recognition. Beijing, China. 2011. 1491–1496.

15 Minetto R, Thome N, Cord M, *et al*. Text detection and recognition in urban scenes. Proceedings of IEEE International Conference on Computer Vision Workshops. Barcelona, Spain. 2012. 227–234.

16 Yu TS, Wang RS. Scene parsing using graph matching on street-view data. Computer Vision and Image Understanding, 2016, 145: 70–80. [doi: 10.1016/j.cviu.2016.01.004]