武汉邮电科学研究院硕士学位论文

开题报告

专 业： 信号与信息处理

研究方向： 人工智能

论文题目：基于改进的DBN网络在视频文本

识别中的研究

研 究 生： 肖青秀 学号： 20160082

导 师： 汤鲲 彭艳兵

报告日期： 2017年11月20日

|  |  |
| --- | --- |
| 1. 课题的来源、目的及意义 | |
| **来源：**  目前，有不少针对视频中文本检测和识别的相关竞赛，比如著名的文档分析与识别国际会议（International Conference on Document Analysis and Recognition，ICDAR）在每一年都会举办有关视频中文本识别的各类竞赛，本文的选题也是受竞赛试题的启发而来。  **目的：**  当前移动互联网的发展正在改变着人们的行为习惯。以微信，Facebook等为主的社交网络，以优酷、爱奇艺、YouTube等为主的视频网站得益于移动互联网的飞速发展，积累了大量的音频，视频，图片等数据。根据2012 年的互联网统计数据[1]，每天上传到社交网络Facebook的照片数量达到3亿张，每月增加的照片内容达到了7PB，每月视频分享网站YouTube的视频播放超过40亿小时。图像、视频已经成为互联网上主流的信息传播方法，以播客、视频网站和社交网络等为代表的图像、视频类服务已经成为网民生活的一个重要部分。  面对互联网上如此巨大的视频、图片数据量，如何正确的理解这些图像、视频的内容变得非常重要。由于图像、视频是以像素形式存储被拍摄对象的颜色、亮度、形状等低层信息，通过这些低层信息很难获得图像内容的描述，给图像内容理解和检索带来了极大的困难。与这些低层信息相比，图像、视频中的文字信息属于高层语义信息，通常作为图像、视频内容的辅助说明，与图像内容紧密相关，是理解图像、视频内容的重要线索，相对于图像、视频中的其它语义内容，文字信息可以被提取出来直接利用。  因此，本课题的研究就是对如何快速的检测并识别视频中的文本信息，获得基于高层语义的精准视频内容，提高视频中文本识别的精度。  **意义：**  近些年，随着科学技术和经济的发展，越来越多的人用上了更加智能的设备，国家的基础设施也在向着智能化方向前进，这也使得越来越多的信息以视频的形式出现。比如，机场、高速、公路等地的安防摄像头采集到的视频信息；社交网络上积累的大量短视频和场景图片；新闻媒体中实时播放的最新事件报道等等。视频和图像，作为更形象的内容展现形式，从直接观看的角度来说，是比较的容易理解的。但是如何在海量的视频图像中准确的找出我们想要的信息，这就必须要用到图像、视频的文本检测和识别了。  目前，图像、视频的文本检测和识别技术的应用范围是非常广泛的，具体有变现在智能交通系统、视觉识别系统、机器人视觉、工业自动化以及智能导航等各行各业。因此本文将改进的DBN网络用于视频中文本的检测和识别，将会加快视频中信息的提取速度，同时提高检测精度，降低实际使用中的成本。 | |
| 1. 本课题的国内外研究现状及发展趋势分析   （在文献调研的基础上完成此部分内容，并列出至少10篇相关文献。） | |
| **国内外研究现状：**  文字识别的研究起源于20 世纪50 年代，经过半个多世纪的发展，很多文字识别方法被提出，Yasuda 等使用直接模式匹配法，将二值化输入字符直接与标准原型集合进行比较。Nakagawa 等使用弹性匹配，将位置符号和每一个原型所有可能的弹性拉伸和压缩进行最优的匹配。Iijima 等使用基于模糊技术的方法，将字符看作一个笔划的集合，使用模糊相似性度量和应用模式进行比较。对于印刷体扫描文档的识别，现有方法的识别率已经达到了99%以上。  但是目前视频的传播越来越广泛，互联网上已经积累了大量的视频资源，如何解读这些资源的含义，挖掘出它们的内在价值，是一项值得关注的工作。为了准确地提取出视频中的具体信息，针对视频中的文本进行定位和识别得到了越多越多的研究。  1999年，T.Sato,T.Kannde和E.K.Kughes开发了一个针对新闻视频静态标题的识别系统。该系统主要工作是提取新闻视频中的标题，将标题4倍放大达到图像增强的效果，然后将标题文本块集成并与最小像素搜索方法多帧融合，提出了新的文本信息特征提取和识别方法。该方法对于新闻节目的字幕识别取得了比较好的识别结果，但是对于动态字幕的识别还有待改进。  2002年，X.A.Tang和B.Luo根据相邻帧之间直方图的差值特征，采用模糊聚类神经网络的方法对字幕出现和小时的帧进行检测。该算法使用相邻帧自动检测视频中的文本信息，表现出比较优异的抗干扰能力，能检测到字幕文本和场景文本。但是此算法针对的是特定的图像、视频模型，一旦输入的文件有变化，神经网络必须要重新训练，适应性较差。  **国内研究现状：**  我国对文字识别的研究工作起步较晚，20 世纪70 年代才开始对数字、字母及符号等识别技术进行研究，70 年代末才开始进行汉字字符识别相关的研究工作。为了推动我国对汉字识别技术的研究，从1986年到1999年，先后进行了汉字识别系统的OCR研究、印刷体汉字识别、手写体汉字识别以及各种OCR产品，，促进了我国中文汉字识别技术的发展，使汉字识别软件走入了产业化。  目前国内的研究机构对于视频文本的识别研究有突破的有：香港中文大学信息工程系、浙江大学、微软中国研究院、中国科学院计算技术研究所、诺基亚研究中心视觉通讯实验室、清华大学、中国科学院自动化研究所、南京大学等。  浙江大学的张引、潘云鹤等提出了一个检测视频帧中文字的彩色边缘算子进行文本提取，其算子针对彩色视频帧的检测定位效果明显。中科院计算所的王伟强等人研究了一个主要基于颜色信息检测和分割的新闻视频中重叠文本检测系统。赵峰等[8]提出了一种可以提取工程图纸中文字的方法，首先计算每个像素点及其邻域内的分形特征，根据直方图统计结果给定阈值形成分割区域图，最后检测分割区域图中的长方形区域作为文字区域。延昊等[9]通过小波变换压缩图像，在压缩图像使用CIP 算子进行处理，通过BP神经网络识别文字区域。  **发展趋势分析：**  现在的生活当中，人们获得图像视频数据的途径越来越多，数字图像数据也越来越多[。图像中包含着海量的数据信息，如果让人去处理这些数据那么无疑是费时费力的，再加上随着现代社会的科技进步速度越来越快，对实时性的要求也就越来越高了。随着人工智能、计算机视觉、模式识别等技术的发展，让计算机自动“理解”图像越来越成为实际应用中的迫切需求。  目前很多学者这对图片、视频中的文本识别和提取问题进行了研究，每年还有ICDAR等场景文本识别的相关竞赛，由此见得的对视频中文本识别的研究已经引起了普遍的重视。但是现有的研究存在以下问题：多应用于简单背景的情况下的小规模数据分析，常见于识别字幕、路标等较为清晰、容易辨别的文字，识别难度并不高；应用机器学习算法或者基于受限玻尔兹曼级模型的DBN网络处理视频中文本的纹理信息，但是模型的学习速度不够迅速。  针对以上问题，本文提出了基于改进的受限玻尔兹曼机模型来接收纹理的文本信息，同时设计了两种改进的DBN网络结构用以提升深度神经网络在视频文本提取中的训练速度。  **参考文献：**  [1] Yasuda M., Fujisawa H. An Improved Correlation Method for Character  Recognition. Systems, Computers, and Controls, 1979, 10(2): 29-38.  [2] Nakagawa M., Akiyama K. A Linear-time Elastic Matching for Stroke Number Free Recognition of on-line Handwritten Characters. Proceedings of the WorkshopFrontiers Handwriting Recognition. Taiwan, R.O.C., 1994: 48-56.  [3] 黄剑华. 自然场景中文本信息提取方法. 哈尔滨工业大学博士论文, 2007.  [4] T.Sato,T.Kallade and E.K.Kughes.Video OCR: indexing digital news libraries by recognition of superimposed captions. ACM Multimedia Syst.(Special Issue on Video Libraries). Japan, 1999,7(5):385-395  [5] X.A.Tang and B.Luo, Pissaloux Using emporal feature vectors.Processing’s Multinedia and ExPo,Hongkong,2002:85-88  [6] 颜建强.图像视频复杂场景中文字检测识别方法研究[D].西安电子科技大学.2014  [7] 张引,潘云鹤.复杂背景下文本提取的彩色边缘检测算子设计[J].软件学报.2001,12(8):1129-1235  [8] Melissa L, Rueden C T, Chris A, et al. Metadata matters: access to image data in the realworld.[J]. Journal of Cell Biology, 2010, 189(5):777-82.  [9] 陈宇.基于深度置信网络的中文信息抽取方法[D].哈尔滨工业大学.2014  [10] 杨玲玲,叶东毅.一种基于图像矩和纹理特征的自然场景文本检测算法[J].小型微型计算机系统.2016,6(6):1313-1317  [11] 蒋人杰,戚飞虎.基于学习的自然背景中文本提取[D].上海交通大学,2007  [12] 彭浩,陈继锋,刘琼等.一种面向聚焦爬虫的自然场景文本定位技术[J].小型微型计算机系统.2014,35(9):2014-2018  [13] 何飞.基于Gabor滤波器的虹膜多特征提取及融合识别方法研究[D].吉林大学.2015  [14] 孙志军,薛磊,许阳明,等.深度学习研究综述[J].计算机应用研究,2012, 29(8): 1001-3695.  [15] 吕刚,郝平,盛建荣.一种改进的深度神经网络在小图像分类中的应用研究[J].计算机应用与软件.2014,31(4):182-185  [16] 吴丹敏.复杂背景下视频文本提取的研究[D].西安交通大学.2015  [17] 毛勇华,代兆胜,桂小林.一种改进的5层深度学习结构与优化方法研究[J].计算机工程.2017.(优先出版)  [18]王忠民,李卓,范琳.基于滑动窗特征融合的深信度网络驾驶行为识别[J].计算机应用研究. 2017.(优先出版)  [19]刘璐,贾彩燕.基于文本扩展模型的网络视频聚类方法[D].智能系统学报.2017(录用搞) | |
| 三、课题中待解决的关键问题 | |
| 对于本课题，如何利用DBN网络实现快速而精确地识别视频中出现的文本的纹理信息是核心问题，具体包括以下两个方面：  **1、DBN网络识别精度的问题**  DBN网络是由受限玻尔兹曼机RBM组成，但是RBM只能接受二值输入，这会Gabor滤波器输出的纹理信息不能完全输入到DBN网络中，从而造成信息丢失问题，对最终文本检测和识别的精度造成影响。  **2、DBN网络的结构设计问题**  对于提取到的文本纹理信息，在提高或不影响视频中文本识别精度的前提下，如何设计DBN网络的结构来加快DBN的训练过程是本课题的研究重点，同时选择合适的网络层数以保证识别的精度也是待解决的问题。 | |
| 四、课题的研究内容 | |
| 近年来，很多学者在视频文字检测和识别方面做了大量的研究，提出了不少性能较好的算法，但是已有的视频中文本识别算法多应用于简单背景的情况下的小规模数据分析，常见于识别字幕、路标等较为清晰、容易辨别的文字，识别难度并不高。随着大量丰富的视频文件出现在网络上，如何快速准确地把其中的信息整理出来是制约信息快速传播的一个重要环节。  本课题对深度学习网络用于视频中文本信息的识别和提取做了相关研究，主要内容有：   1. 在处理视频中文本的纹理信息时，针对受限玻尔兹曼机只能接受二值输入，文本提出将第一层RBM的可视节点设计成符合高斯分布，隐层符合伯努利分布的状态，用以改善接受因二值输入带来的信息丢失问题。 2. 为了加快深度置信网络的学习速率，本文提出两种对称结构的DBN网络。第一种对称结构为：首先保证最后一层的节点数等于输入维度数，其次中间隐层的节点数逐渐减少，随后又逐渐增加，类似于两个塔尖相对的金字塔结构。第二种对称结构为：首先保证最后一层的节点数等于输入维度数，中间隐层的节点数设计与第一种结构相反，先逐渐增加，随后又逐渐减少，类似于两个底座相对的金字塔结构。 3. 对比两种对称结构DBN的学习速率和文本识别精度，同时测试发现结果精度最高的网络层数。 | |
| 五、拟采取的研究方法、实验方案、技术路线 | |
| **1、纹理信息的获取**  纹理图像分类是视觉研究中的基本问题, 也是一个非常具有挑战性的问题. 纹理图像分类是指基于待分类图像或图像区域的内容为其指定一个预先定义的纹理类别.纹理特征描述和分类器是图像分类的两个关键环节. 对图像进行纹理特征描述是纹理图像分类的主要研究内容, 因为如果提取的纹理特征不好, 即便好的分类器也无法完成识别任务。  基于纹理特征的文本检测方法认为文本是一种特殊的纹理，这种纹理特征是由于字符特定的排列方向以及字符颜色与背景颜色周期性变化而产生的。通常采用Gabor 变换、小波变换和傅里叶变换等方法检测文本的纹理特征，根据纹理特征得到文本区域。  为了获取视频中的文本纹理信息，本文采用Gabor 滤波器提取视频中文本的纹理信息，为下一步的机器学习算法或者深度学习网络的分类学习做准备。这里，我们简要的介绍一下Gabor 滤波器。二维Gabor 函数定义如下：    其中，*K*表示高斯核函数的幅度，(*a*,*b*)是高斯核函数两个轴方向的尺度，(*x*0, *y* 0)是高斯核函数的峰值，(*u*0,*v*0 )表示在直角坐标空间频率的正弦载波，*P*是正弦载波相位，下标*r* 表示旋转操作，定义如下操作：      在Gabor 滤波器中，*u*0 和*v*0定义了直角坐标空间频率的正弦载波，空间频率在笛卡尔坐标系中可以通过幅度和方向角度 表示，因此，在笛卡尔坐标系中，Gabor 滤波器变换成如下的表示公式：    为了得到图像的局部特征，需要将上面的公式在空间域和频率域上离散化，通过二维Gabor 的Fourier 变换得到：  在笛卡尔坐标系红Gabor滤波器表示为：  一般来说，高斯包络函数的峰值位置 (*x0* , *y0* )为 0，Gabor 滤波器主要通过滤波参数*a* 和*b*调整。因此，在选择Gabor滤波器时，我们只需要计算滤波参数*a*和*b*就可以选择不同的Gabor 滤波器。由于文字区域的纹理具有丰富的中频信息和高频信息，因此，我们用*Ul* 和*Uh* 表示感兴趣区域的低频率中心和高频率中心，*K*表示方向数，*S*表示尺度数。滤波器的滤波参数*a* 和 *b*通过下面的公式计算。  **2、RBM网络的设计**  深度学习算法采用深度置信网络，通过一系列受限波尔兹曼机的概率模型组成。Hinton 和Sejnowski 设计了一种能够用在机器学习中的随机循环网络-受限波尔兹曼机(RBM)，它是一种包括一层可视化单元(v)和一层隐藏单元(h)，并且限制单元之间的连接只存在于不同层的单元之间，相同层的单元之间不存在连接。  深度置信网络对于图像和视频来说，他们均为实值数据，采用二值分布对其建模效果不明显。针对受限玻尔兹曼机只能接受二值输入，文本提出将第一层RBM的可视节点设计成符合高斯分布，隐层符合伯努利分布的状态，用以改善接受因二值输入带来的信息丢失问题。其能量函数表示为：  **3、DBN网络的结构**  因为DBN网络对高维特征向量具有很强的特征提取和特征分类能力，其深层结构更能帮助它提取出更抽象、更具可分类性的特征。因此，采用DBN对文字区域和背景区域进行分离处理，能够用来区分文字区域和背景区域。用DBN算法处理三种滤波器提取到的纹理信息，得到3种组合的文本检测系统，将结果输出。实验表明，深度学习网络具有较高的分类准确性，效果要好于传统的神经网络、SVM 等方法。  深度学习算法采用深度置信网络，通过一系列受限波尔兹曼机的概率模型组成。Hinton 和Sejnowski 设计了一种能够用在机器学习中的随机循环网络-受限波尔兹曼机(RBM)，它是一种如图包括一层可视化单元(v)和一层隐藏单元(h)，并且限制单元之间的连接只存在于不同层的单元之间，相同层的单元之间不存在连接。  **3.1 传统的DBN网络结构**  通过自底向上组合多个RBM 可以构建一个DBN 网络，如图5.2所示，可用隐单元的输出作为训练上层RBM 的输入，第二层RBM 的输出作为第三层RBM的输入等等形成DBN 网络。    图1 典型的DBN网络结构  **3.2 对称设计的DBN网络结构**  现有的DBN网络处理视频中文本的纹理信息，识别的正确率基本在80%左右，可以看出对视频中文本的检测和识别还有较大的提升空间，所以本文参照图像数据压缩的方法，以5层深度置信网络为例，设计了两种对称的DBN网络结构，尝试提高DBN网络的学习速度。  **3.2.1 塔尖相对的DBN网络结构**  具体的网络结构为：第一层节点数等于样本维度数，本课题采用Gabor滤波器输出的8x8共64个维度的纹理信息为输入信号；第二隐层节点数量为第一层的三分之二；第三隐层节点数为第一层的二分之一；第四隐层节点数与第三隐层相同，第五隐层节点数与第二隐层相同。其中，总体来看就是两个塔尖相对的金字塔结构。由于网络结构的对称性，在对DBN进行训练时，只需要训练前两层的权值和即可，可以通过和这样就可以降低模型的训练速度了。    图2 塔尖相对的DBN网络结构  **2.2.1 底座相对的DBN网络结构**  具体的网络结构为：第一层节点数等于样本维度数，本课题采用Gabor滤波器输出的8x8共64个维度的纹理信息为输入信号；第二隐层节点数量为第一层的两倍；第三隐层节点数为第一层的三倍；第四隐层节点数与第三隐层相同，第五隐层节点数与第二隐层相同。其中，总体来看就是两个底座相对的金字塔结构。    图3：底座相对的DBN网络结构  **3、DBN网络的训练**  DBN 网络的训练分两步完成，先完成RBM的预训练，再使用BP算法，将DBN网络构建阶段得到的参数作为初始权值进行训练。在这里，BP 网络和RBM 网络采用同样的网络结构，它们包含相同的输入层和隐藏层，每层的结点数目也是相同的，只是BP 网络最后还有一个输出层，该层的结点数就是所有的类别总数。总体的训练流程图为：    图4：DBN网络的训练过程  DBN网络采用无标注的训练样本构建，以H0和H1之间的训练为例，H0和H1层构成一个RBM，可见层的节点数和H0一样，隐藏层的节点数和H1一样，进行交替的吉布斯抽样调整权值，直到RBM收敛于某个特定分布。RBM预训练过程如下：  1) 设定迭代次数step ，随机初始化参数集合；  2) 将输入特征向量赋值给v0，计算h0，v1和h1 ；  3) 计算RBM网络初始状态与更新状态下的联合概率分布，利用如下公式修正参数，得；  4) 如果t >= step，程序结束；如果t<Step，则将 赋值于，并转步骤(2)。  BP网络的训练过程：   1. 将DBN网络的参数作为BP网络初始值，训练步长为*N*； 2. 进行前向计算，对第*l*层的*j*单元，其值为，当神经元*j*为输出层时，则令，误差; 3. 计算并反向传递修正网络的权值参数，输出单元：；隐含单元：; 4. 修改权值，其中是学习速率； 5. *n* = *n+*1，当*n<N* 时，转步骤(2)继续执行，否则结束训练。   **4、形态学处理**  训练结束后，利用形态学方法对文本检测系统的输出结果进行处理，首先剔除图像中一些孤立的噪声点，其次对结果进行膨胀、腐蚀等操作，突出待检测区域的轮廓信息，然后根据图像块的八邻域对文字区域中的空洞进行填充，最终形成的文字区域。 | |
| 六、论文的创新点或实用性 |
| **论文创新点：**  1、针对当前视频中文本信息提取的现状，根据文本区域的纹理特征，本文提出了具有高斯分布的实数节点替代DBN的二进制可视节点，改善受限玻尔兹曼机只能接收二值输入的缺陷，尽可能多的接受Gabor滤波器输出的纹理特征。  2、在高斯噪声的可视节点RBM的基础上，为了加快深度神经网络的计算速度，本文提出了两种对称的DBN网络结构，将前面RBM层级的权值求得后，转置，赋给后面的RBM层级，以期减少更新的权值个数，从而达到加快网络训练速度的目的。  **论文实用性：**  随着大数据时代的到来，数据以指数形式快速的增长，给视频数据的检索和管理带来了很大的挑战，本文对图像视频复杂场景中文字的检测和识别进行了研究，并计划设计和实现一个简单的视频文本检测识别系统，同时基于视频中提取到的文本信息，改善当前视频聚类算法中的稀疏性问题，实现基于视频内容的聚类，提高聚类的准确性。  文本提出视频中文本信息提取方法可以用于复杂场景下的信息发现；基于内容的视频聚类方法可以用于实际业务场景中发现不法视频，也可以用来设计实现基于内容的视频推荐系统。总体而言，本文的研究有着非常强的现实可用性。 |
| 七、研究计划进度和预期成果 |
| **计划进度：**  2017年10月-2017年11月：阅读文本提取相关文献和书籍，了解相关理论知识，并确定论文题目；  2017年12月-2018年1月：熟悉数据集的特点，并收集资料，了解论文题目相关法；  2018年2月-2018年4月：熟悉常用的机器学习算法，初步完成文本纹理信息提取和DBN网络建模；  2018年4月-2018年6月：利用测试集测试模型，调整参数，完善实验结果；  2018年6月-2018年7月：完成论文写作、修改定稿。  **预期成果：**  利用公开数据集（如ICDAR2013），训练测试本文提出的基于Gabor滤波器+改进DBN网络的视频文本提取算法模型，对比传统的基于SVM的文本提取模型，提高结果的查准率和查全率。同时基于内容的视频聚类模型将提高推荐的准确度。 |
| 八、导师意见  （请对项目的意义、具体内容、创新点和取得预期成果的可能性等进行评价） |
| 本论文对如何从视频中快速检测并识别文本信息进行了相关研究，通过改进受限玻尔兹曼机提高输出文本区域纹理特征的数量，同时提出两种对称结构DBN来提升模型训练速度和文本识别进度。相比已有研究工作，论文的研究成果能有更好的检测精度和训练速度提升。论文选题合理，工作量适中，论文研究工作具有较强的实用价值，同意开题。  导师签字： |
| 九、研究生部审核意见 |
| 审核老师签字： |