硕 士 专 业 学 位 论 文

中插广告自动识别系统的设计与实现

Design and implementation of an automatic recognization system for interpolation advertisements

作者：张莹

导师：郭宇春

北京交通大学

2019 年 5 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月

学校代码：10004 密级：公开

北京交通大学

硕士专业学位论文

中插广告自动识别系统的设计与实现

Design and implementation of an automatic recognization system for interpolation advertisements

作者姓名：张莹 学 号：17125082

导师姓名：郭宇春 职 称：教授

工程硕士专业领域：电子与通信工程 学位级别：硕士

北京交通大学

2019 年 5 月

致谢

本论文的研究工作是在我的导师郭宇春教授的悉心指导下完成的。郭宇春教授科学的工作方法、开阔的视野以及渊博的知识给了我极大的帮助。郭宇春教授严谨的治学态度、精益求精的学术风范、恪尽职守的工作作风以及豁达的人生理念，深深的感染和激励着我不断进取，对我以后的工作和学习有着很大的影响。在此衷心感谢多年来郭宇春教授对我的悉心指导和关怀。

同样感谢陈一帅老师对本次毕设的帮助和指导，在读研期间，多次和陈老师一起修改代码、讨论论文，每次交流都能产生新的想法。陈老师在工作和生活中饱满的热情以及负责的态度将会一直是我学习的榜样。

感谢实验室里的所有老师。衷心感谢感谢孙强老师、张立军老师、赵永祥老师、李纯喜老师、郑宏云老师在我研究生阶段对我的帮助和关怀，我所有的科研成果都凝结着各位老师的辛勤汗水。在此向各位老师表示诚挚的谢意。

另外，在实验室工作和撰写论文期间，王一师姐、李勇宏师兄、陈滨师兄、唐伟康师兄、魏中锐、尹姜谊、苏健、艾方哲、于滋灏、冯梦菲等同学对我的研究工作给予了热心帮助，在此向他们表示我的感谢之意，还要诚挚的感谢国家自然科学基金(No. 61572071，61271199，61301082)的资助。。

最后，特别感谢一直无微不至的关心、支持我的父母和朋友，正是他们热情的鼓励和默默的奉献，才使得我顺利的完成学业，成为社会的有用之才。

摘要

随着互联网广告行业的快速发展，视频广告在广告投放中占据着越来越重要的地位，视频广告的自动识别可以帮助广告商判断网站是否按约定投放广告，同时也可以帮助咨询机构根据广告投放量分析广告主的经营状况。随着广告转化率提升需求的不断增长，近几年出现了一种在电视剧中插播的新型视频广告形式，广告方依托电视剧的故事背景和人物关系构思广告创意，将广告做成“番外篇短视频”穿插在剧集中，称为“中插广告”。这种广告形式模糊了广告与剧情的界限，使得传统广告的自动识别方法不再适用。

本文针对传统广告自动识别系统在镜头切分、镜头分类和广告内容识别模块上不适应中插广告识别的具体问题，利用深度学习技术与传统的计算机视觉处理技术，结合图像特征、音频特征和文本特征，从时间和空间多角度出发，设计了一种中插广告的自动识别系统。

本文的主要贡献有如下三点：

(1) 针对视频中镜头间渐变情况提出一种新的镜头切分方法。由于中插广告和剧情的场景相似，因而在剧情镜头与广告镜头之间较多采用镜头渐变切换，相比传统广告采用的镜头突变切换，镜头切分更为困难。基于对中插广告的观察，发现渐变过程中会出现黑镜头，本文跳出计算帧间距离的常规思路，提出一种简单有效的利用颜色变化趋势切分渐变镜头的解决方案。

(2) 针对广告镜头与剧情镜头的视频特征高度相似难以区分的问题，本文提出利用 LSTM 网络和 Attention 网络组合，获取音视频时序高维特征并强化显著特征，改善视频镜头分类性能。传统的广告识别系统采用 CNN 网络进行音视频特征深度表达，未利用视频帧序列的时序相关性，本文利用 LSTM 来获取前后帧的时序关系；进一步，本文使用了 Attention 网络获得不同维度特征在镜头分类结果中的占比，强化性能影响显著的特征向量。实验表明本文提出的方法分类准确率可以达到 88%，相比传统的机器学习方法提高了 4%。

(3) 针对广告中 Logo 不显著甚至没有 Logo 导致基于 Logo 的广告识别方法无效的问题，本文提出了结合文字识别和音频特征匹配的广告内容识别方法。基于对中插广告的观察，大量广告以文字替代 Logo 标识广告商品，也有一些广告甚至没有文字仅以声音标识广告商品。因此本文采用 OCR 文字识别技术结合音频色谱图特征匹配的方法进行广告内容识别，准确率可以达到 98%。

本文设计实现的中插广告自动识别系统不仅可以检测出视频中是否存在中插广告，同时还可以识别具体的广告内容，具有很强的现实意义和实际价值。

关键词：中插广告；镜头切分；镜头分类；广告内容识别

ABSTRACT

With the rapid development of the Internet advertising industry, video advertising is playing an increasingly important role in the advertising. The automatic identification of video advertising can help the advertiser to judge whether the advertisements are released according to the agreement, and also help consulting agency to analyze the business status of advertisers according to the amount of advertising. With the increasing demand for increasing advertising conversion rate, a new type of advertisements which is interpolated in the video appears in recent years, advertisers conceive the creative advertisements based on video story background and role setting, and make the advertisement into "video clips" interwoven in the video, which is called "Interpolation advertisements". The appearance of this kind of advertisements blurs the boundary between the advertisement and the video, making the traditional methods of advertising automatic identification is no longer feasible.

This paper aims to slove the problem where the traditional methods of advertising automatic identification does not adapt to the interpolation advertisements in shot segmentation, shot classify and content recognition, designing an automatic recognization system for interpolation advertisements, which applies deep learning technology and traditional computer vision processing technology, integrates image features, audio features and text features and takes information in time and space domain into consideration.

The main contributions of the thesis are listed as follows：

(1) A new method of shot boundary detection is proposed for video intershot gradual transition. Because the scene of advertisement and video is similar, so there are more gradual transition shot between the video and the advertising, compared with the hot abrupt transition used in traditional advertising, it is more difficult to segment the shots. Based on the observation of the interpolation advertisements, we found that black shot would appear in the process of shot gradual transition, this paper breaks the conventional idea of calculating the inter-frame distance, and proposes a simple and effective solution to detect boundary for intershot gradual transition, which uses the color change trend.

(2) This paper aims to slove the problem that the high similarity between video shots and advertising shots and difficult to distinguish, presenting a combination of LSTM model and Attention network to obtain timing sequence features existing in higher

dimensions of audio and video and strengthen significant features to improve the classification performance. The traditional system uses CNN network to extract deep expression of audio and video features without using the timing correlation of video frame sequence, LSTM is used to get the correlation between the current frame and the next frame in this paper. Furthermore, the Attention network is used to get the weights of features in different dimensions in the results of shot classification in this paper, and the feature vectors with significant influence on performance are enhanced. Experiments show that the final classification accuracy can reach 88%, 4% higher than the traditional machine learning method.

(3) This paper aims to slove the problem that the indistintion of advertising logo leads to the invalidation in identifying the advertising content in which case the logo is extremely relied on, presenting a method of text recognition combined with audio feature matching to identify the advertising content. Based on the observation of the endogenous advertisements derived from video, a large number of the text replaces the logo to identify advertising products, there are also advertisements that don't even have text but sound to identify the product. Therefore, this paper adopts the method of combining OCR text recognition with audio chromaticity diagram matching to identify advertising content, and the accuracy can reach 98%.

The recognition system of interpolation advertisements proposed in this paper can not only detect the existence of interpolation advertisements in video, but also identify the advertising content, which has a strong practical significance and practical value.

KEYWORDS： Interpolation advertisements; Shot segmentation; Shot classification; Advertising content recognition

目录

摘要 III
ABSTRACT IV
1 引言 1
1.1 研究背景及意义 1
1.2 国内外研究现状 3
1.2.1 镜头边界检测 3
1.2.2 视频分类 5
1.2.3 多特征视频分类 6
1.2.4 文字区域检测 6
1.2.5 文字内容识别 8
1.3 本论文的主要研究内容 9
1.4 本论文的主要贡献 9
1.5 本论文的组织结构 10
2 技术背景 11
2.1 开发平台 11
2.1.1 视频处理平台 11
2.1.2 Anaconda 简介 11
2.1.3 Pycharm 简介 12
2.1.4 Scikit-Learn 库 13
2.1.5 Tensorflow 开源软件库 13
2.1.6 Keras 开源软件库 13
2.2 镜头边界检测方法 14
2.2.1 颜色直方图 14
2.2.2 边缘检测 15
2.2.3 曼哈顿差分距离 16
2.3 机器学习算法 16
2.3.1 支持向量机 17
2.3.2 随机森林算法 19
2.3.3 梯度提升树 20
2.3.4 极端梯度提升 20
2.4 深度学习算法 21
2.4.1 卷积神经网络 21
2.4.2 LSTM 网络 22
2.4.3 Attention 网络 23
2.5 文字区域检测 24
2.5.1 R-CNN 简介 24
2.5.2 Faster R-CNN 算法简介 26
2.6 文字内容识别 26
2.6.1 CTC 损失函数简介 26

2.6.2 CRNN 算法简介 27
2.7 本章小结 27
3 系统设计与数据集 28
3.1 系统设计 28
3.2 镜头边界检测 29
3.2.1 突变与渐变 29
3.2.2 镜头边界检测 31
3.3 数据集介绍 34
3.3.1 镜头图像数据集 35
3.3.2 镜头音频数据集 36
3.4 本章小结 36
4 镜头分类 37
4.1 问题描述 37
4.2 提取镜头的深度卷积特征 37
4.2.1 镜头图像特征 39
4.2.3 镜头音频特征 42
4.3 特征融合 42
4.3.1 特征并联 43
4.3.2 特征拼接 44
4.4 训练镜头分类器 45
4.4.1 机器学习 45
4.4.2 深度学习 46
4.5 本章小结 48
5 广告内容识别 49
5.1 问题描述 49
5.2 利用文字进行内容识别 49
5.2.1 文字数据库 50
5.2.2 文字区域检测 50
5.2.3 区域内容识别 52
5.3 利用音频特征进行内容识别 52
5.3.1 音频数据库 53
5.3.2 色度特征匹配 53
5.4 综合文字与音频方法进行文字识别 54
5.5 本章小结 55
6 总结及展望 56
6.1 本文工作总结 56
6.2 未来工作展望 57
参考文献 58
作者简历及攻读硕士/博士学位期间取得的研究成果 61
独创性声明 62

学位论文数据集..... 63

缩略词表

英文缩写	英文全称	中文全称
SVM	Support Vector Machine	支持向量机
RF	Random Forests	随机森林
DT	Decision Trees	决策树
GBDT	Gradient Boosting Decison Tree	梯度提升树
CNN	Convolutional Neural Network	卷积神经网络
RNN	Recurrent Neural Network	循环神经网络
Xgboost	eXtreme Gradient Boosting	极端梯度提升
CTC	Connectionist Temporal Classfication	时序分类
LSTM	Long Short-Term Memory	长短期记忆
RPN	Region proposal network	候选区域网络
DBNs	Deep Belief Nets	深度置信网
NN	Neural Networks	神经网络
M	Manhattan Distance	曼哈顿距离
GPU	Graphics Processing Unit	图形处理器
PCA	Principal Component Analysis	主成分分析

1 引言

1.1 研究背景及意义

随着互联网广告行业的快速发展，视频广告在广告投放中占据着越来越重要的地位，它以丰富的视听感受给人留下深刻的印象和记忆。视频广告的自动识别一方面可以帮助咨询机构根据广告投放量分析广告主的经营状况；另一方面通过对视频中广告的识别获得广告内容，可以帮助广告商判断网站是否按约定投放该广告。因此识别视频中的广告内容、挖掘出大量有价值的广告信息，可以更好的满足未来市场信息的需求。

对于商业分析机构而言，通过观察视频流中不同品牌广告的播放时间及播放时长，有助于机构分析企业的运营状况、市场细分情况等。

对于广告商来说，广告是一种营销工具，通过广告吸引顾客对产品的关注来增加产品的销量。他们需要通过广告识别来验证广告是以合同形式播出的，或许公司也在关注他们的竞争者在做什么。

传统的广告识别系统基本都是利用广告与非广告之间的视频、音频特征差异作为区分。但随着广告转化率提升需求的不断增长，近几年出现了一种在电视剧中插播的新型视频广告形式，该广告沿用剧中主创和人物关系打造主线剧情外的番外小剧场，广告演员是剧中的人物，广告的内容也和剧情有一定的关联，称为“中插广告”或者“创意中插广告”[1]。这种广告形式模糊了广告与剧情的界限，使得识别视频中广告内容的任务变的更加困难。

中插广告作为一种新的广告形式，依托电视剧的故事背景和人物关系打造“番外篇短视频”被穿插在剧集中。该广告 2016 年开始在视频网站播出的剧中遍地开花，《楚乔传》、《白夜追凶》、《大军师司马懿之军师联盟》、《那年花开月正圆》等热播剧中都会插入中插广告。如图 1-1 所示为两组剧情镜头与广告镜头，左边为剧情镜头中一帧，右边是广告镜头中一帧，从中可以看出除了广告镜头中会出现广告商品，我们几乎看不到二者的差异。广告中的演员是剧中的人物，广告的内容也和剧情有一定的关联。艺恩数据显示[2]，2017 年视频中插广告的市场规模突破了 30 亿大关。如图 1-2 所示[3]中插广告价格正在逐年递增，在优酷 APP 的电视剧的广告营收中中插广告已经达到了四分之一，未来能占到广告收入的三分之一。中插广告正在呈爆发式增长，从几年前的无人问津到现在 300 万一条抢破头；中插广告大有取代传统广告位置之势，成为市场监测的新挑战。因此本文设计的中插广告识别系统，能够更好地满足未来市场信息的需求。

(a) 剧情镜头 (b) 广告镜头

(a) Video shots (b) Advertising shots

图 1-1 剧情镜头和广告镜头

Figure 1-1 The shot of video and advertisement

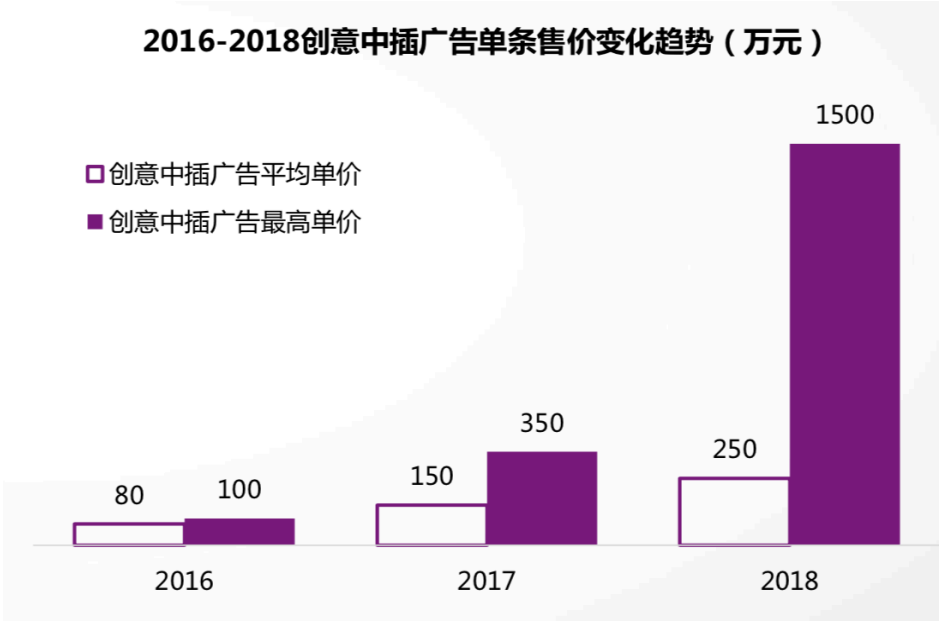


图 1-2 广告价格变化

Figure 1-2 The price change in advertisement

传统的视频中广告的识别主要基于图像和声音，由于广告中的场景、人物以及声音和剧情中的都是不一样的，有较明显的区分。而中插广告中的演员是剧中角色，场景也是由剧情内生而来，因此单靠传统的广告识别的方法不能获得很好的区分。目前市场上还没有针对这种广告的识别方法，识别中插广告所面临的挑战主要有如下三部分：

(1) 中插广告与剧情的界限比较模糊，寻找剧情镜头和广告镜头的边界存在一

定的困难。

(2) 广告镜头与剧情镜头相似度太高，仅仅利用机器学习方法通过简单的颜色特征已经不能区分二者的差异，需要寻找更高维度的特征进行区分。

(3) 中插广告的 Logo 已经变得不再显著，传统广告 Logo 匹配的方法不能获得很好的识别，需要寻找新的广告特征进行识别。

为了实现对中插广告的识别，我们需要解决以上三点困难。首先针对中插广告与剧情的界限比较模糊的问题，我们需要找到一个边界画面能够将剧情镜头和广告镜头进行区分；对于镜头分类的困难，我们需要找到新的算法能够提取到图像的高维特征，并且通过训练可以获得剧情镜头和广告镜头的差异；最后针对本文视频中广告 Logo 不显著的问题，我们需要通过对广告的观察，找到更有辨识度的特征进行识别。

1.2 国内外研究现状

随着经济的发展，视频中的广告已经成为社会生活中越来越重要的一部分，而随着广告转化率提升需求的不断增长，一种新的中插广告随之而来，国内中插广告于 2016 年开始在视频网站播出的剧中遍地开花，2017 年市场规模突破了 30 亿大关。在优酷 APP 的电视剧的广告营收中，中插广告收入已经达到了四分之一，未来能占到广告收入的三分之一，视频中的广告包含着大量有价值的信息，识别视频中的广告内容对社会的发展具有重要的意义。

目前针对这种中插广告，更多的是有关其市场发展的趋势分析[2]，由于该类广告近年才出现，但是其发展速度在持续的增长，继各大网剧后《偶像练习生》、《创造 101》、《热血街舞团》、《这!就是街舞》等越来越多的综艺节目中也植入了此种类型的广告，未来电影、动漫等视频内容中都将开放中插广告资源位。目前并没有对这种中插广告的内容识别系统，因此本文设计了一种中插广告识别系统，主要包括三部分：镜头边界检测、镜头分类、广告内容识别，本节将对这三部分的国内外研究现状进行介绍。

1.2.1 镜头边界检测

视频拍摄是由单个相机连续拍摄的一系列相互关联的画面。摄像头从开机到关机所记录下来的一段没有间隔的画面或者前、后光学转换之间的一个完整片段，称之为一个镜头。场景是由单个或多个镜头组成，若干个场景组成电视剧，因此镜头也是电视剧组成的基本单位。在长片段视频中往往都是以镜头为单位进行拍摄

的，而我们关心的是如何找到视频中带有广告的镜头，我们并不是很关心电视剧的内容是什么，因此我们需要把电视剧分解成镜头，只要找到广告镜头即可。切分广告镜头的前提是先找到镜头的变化边界，镜头边界是基于画面过渡引起的视觉差异，前后两帧之间的差异通常是在切换镜头时产生的，这种差异存在两种形式，一种是突变，可以明显看出前后帧的变化即前一帧属于一个镜头，后一帧属于另外一个镜头；另一种是渐变，渐变镜头中亮度是缓慢变化的，一般前后帧的变化并不是很明显，渐变镜头主要有两种展示形式，一种是前一镜头最后的一些即将消失的帧会被下一个镜头开始帧所替代，本文中称为镜头内渐变；一种是随时间从前一个镜头淡出或从后一个镜头淡入，本文称为镜头间渐变。

文献[4]介绍了基于颜色直方图对比的检测方法。该方法需要确定该图像中具有多少特征值即颜色的种类，接着计算每帧图像中的具有某一特征值的像素的个数，再计算该特征出现在概率来生成该图像的颜色直方图。通过对比相邻帧的颜色直方图可以得知，如果变化较大则就可能在该帧前后发生了镜头突变或者镜头渐变，由于颜色直方图的方法只是考虑了颜色的组成以及出现的概率，对于图像中人物的变化以及平移、缩放等变换是不敏感的，因此这种方法只适合在那些不需要考虑目标空间位置变换的场景下使用。文献[5]提出了一种边缘检测的方法，该方法检测依据是边缘的变化率，边缘的变化分为新出现的边缘和消失的边缘。对于新出现的边缘与前一帧的边缘像素进行比较，如果前后帧的边缘像素距离大于某一阈值，则认为是新出现的边缘像素；对于消失的边缘与后一帧的边缘像素进行比较，如果前后帧的边缘像素距离大于某一阈值，则认为是消失的边缘像素。该方法可以对突变镜头或者渐变镜头进行检测，其缺点在于计算边缘的复杂度很高，对于处理大量的图片十分耗时。文献[6]中介绍了一种新的镜头边界检测的方法。首先需要计算图像在三个通道上的像素之和作为特征，计算出相邻帧的曼哈顿距离。由于渐变镜头的边界帧曼哈顿距离位于内部帧和突变边界帧之间，如果阈值选择不当就会将渐变镜头进行切分，因此需要计算曼哈顿差分距离。曼哈顿差分距离相对于曼哈顿距离能够增加突变镜头和渐变镜头的差距防止将渐变镜头进行误切分，由于传统广告中视频与广告的界限非常明显，切分镜头的时候只检测镜头突变即可，因此此种方法在传统广告的识别上可以获得较好的应用，而本文中要识别的广告和剧情镜头的分界有一部分是渐变的，镜头边界情况更为复杂。

已有方法利用帧间距离也可以将渐变镜头与突变镜头一起检测出来，但是本文中存在两种渐变方式，一种是由剧情渐入到广告的渐变称之镜头间渐变，另一种是剧情内部存在的一种渐变称之为镜头内渐变。为了保证视频内容主题的完整性，本文跳出计算帧间距离的常规思路，提出一种简单有效的利用颜色变化趋势切分渐变镜头的解决方案，只对镜头间的渐变进行切分而保留镜头内的渐变，本文相对

已有的识别方法不仅简单，而且计算效率更高。

1.2.2 视频分类

视频分类首先要做的是提取特征，根据不同的特征将视频进行分类。特征提取可以分为两类，一类是局部特征提取，它是指视频中的局部区域，比如某一变化比较剧烈的局部时空区域；另一类是全局特征提取，它是指整个画面中的整体特征。局部特征比全局特征对视频中的一些背景变化、人为抖动等鲁棒性更强。

传统的视频分类方法基本上都是基于手工设计的特征和一些常见的机器学习方法。比如：基于局部时空域的运动和表现信息，利用词袋模型等方式生成视频编码，然后通过训练视频编码来得到分类器。目前，基于轨迹的方法主要代表是 DT（Dense Trajectories）和 IDT[7]（Improved Dense Trajectorie），这两种方法是人工设计特征算法发展的基础，在此基础上许多研究者又进行了深入探究，如对描述符使用不同的池化策略如 FV（Fisher Vector）[8]和 Rank-Pooling[9]等在 HMDB51 等数据集上取得很好的效果。

然而随着深度学习的发展，特别是 CNN（Convolutional Neural Network）、LSTM（Long Short-Term Memory）、Attention 等深度学习网络在视频分类中的大量应用，其分类性能逐渐超越了基于 DT 和 IDT 的传统方法，使得这些基于人工设计特征的方法慢慢淡出了人们的视野。

在视频分类中有两类信息至关重要：单帧的静态表观信息以及多帧之间的时序关系。针对这一点很多视频分类相关的文章都进行了研究，文献[10]介绍了一种 Two-stream CNNs 的方法。Two-stream CNNs 通过两个分支来捕获外观和移动的信息，这对于视频分类是比较有效的，但是它要训练两个网络比较费时且需要提前提取光流。文献[11,12]提出了一种 3D CNNs 方法，3D CNNs 利用 3 维卷积直接从 RGB 的堆叠序列中学习到时空域的特征，但是 3D CNNs 的效果还是比 Two-stream 要稍差一些，这可能说明了 3D 的结构可能并不能够有效地同时对表观信息和时序关系进行建模。文献[13,14]提出了一种 2D CNNs + 时序模型（如 LSTM，时域卷积，稀疏采样和聚合[15]，Attention 等）。2D CNNs + 时序模型利用 CNN 网络去捕捉图像特征，利用 LSTM 网络去捕获多帧之间时序关系，因此获得更好的分类准确率。文献[16-18]介绍了一种 Attention 机制，Attention 机制是一种注意力机制，能够更好的捕获图像表达的重要信息。应用于图像识别[19]或分类时，模仿人看图像目光的焦点在不同的物体上移动；应用于语言识别时，每次集中于部分特征上识别更加准确。

目前应用于广告识别系统中的镜头分类方法主要采用机器学习方法，由于视

频的前后帧之间存在时序关系，所以本文将利用 LSTM 网络获取前后帧之间的关联信息。

1.2.3 多特征视频分类

电视剧中除了图像信息还存在与图像同步的音频信息，由于每个人的声音以及说话语速都是不一样的，所以通过音频也可以获得大量有价值的信息，因此对视频进行分析的时候，可以将音频信息与图像信息融合获得更多的信息，从而提高分类准确率。

文献[20]中采用了图像特征、音频特征以及文字特征，图像特征包括 RGB 和光流。它的拼接方式有两种，一种是在输入分类网络之前直接进行拼接即将所有特征进行相连；一种是在特征输入分类网络之后在输入全连接层之前将特征进行拼接。由于有些特征可能没有对应的标签，文献[21]介绍了一种半监督的分类方法，即利用已有标签与未知标签的相似度进行训练分类。文献[22]创建了一个潜在空间，将不同维度的特征按照权重映射到这个潜在空间作为融合后的新特征进行训练分类。文献[23,24]分别提取视频和音频的特征，然后利用支持向量机进行训练分类，这种方法对于区分度比较大的镜头来说比较有效，但对于本文的中插广告，由于该广告与电视剧的区分度较小，因此此种方法并不能获得很好的分类准确度。

上述方法都是将不同维度的特征进行了简单的融合，但是并未区分不同维度的特征对视频分类的影响力的差异。因此本文针对融合的特征使用了 Attention 网络，Attention 网络可以获得不同镜头中图像和音频对分类结果的权重，进而区分不同镜头中多维特征的差异，从而提高分类准确率。

1.2.4 文字区域检测

文字识别算法分为两部分，首先是文字区域检测，也叫目标检测。

传统的文字区域检测方法有基于纹理分析的方法[25-28]，该方法将文本作为一种特殊的纹理类型利用文本的局部强度、滤波响应、和小波系数等纹理特性来区分图像中的文本区域和非文本区域，这种方法的缺点是计算量大，因为所有的位置和比例尺都应该扫描，适用于比较简单的场景。还有基于区域的方法[29-33]，首先通过颜色聚类或极值区域提取的方法提取候选区域，然后使用手动设计的规则或自动训练的分类器过滤出非文本区域，因为要处理的区域数量相对较少，因此这类方法效率相比更高，但是这种方法对旋转、缩放和字体变化不敏感。文献[34,35]提出了一种混合方法，即将基于纹理的方法和基于区域的方法进行了结合，分别利用了这

两种方法的优点，首先提取所有可能是文本区域的边缘像素，接着利用区域轮廓的几何特性和梯度生成候选文本区域，然后利用纹理分析技术区分真正的文本区域和非文本区域。

随着深度学习技术的发展，Ross Girshick 等人提出了一种基于 R-CNN（Region-CNN）的目标检测方法[36]，对输入的一张图片用选择性搜索算法选出目标候选框输入到 CNN，然后 CNN 对每个候选框输出这个框的类别得分以及这个框图片对应的坐标位置。一般 CNN 后接全连接层或者分类器都需要固定的输入尺寸，因此不得不对输入数据进行修剪或者变形，这些预处理会造成数据的丢失或几何的失真，而 SPP（Spatial Pyramid Pooling）[37] 网络使输出尺度始终是固定的。Fast R-CNN[38]便在 R-CNN 中加入了 SppNET，但是 Fast R-CNN 在选择性搜索的时候需要对每个候选框进行一次卷积比较耗时。而 Faster R-CNN[39]将 RPN(region proposal network)网络[37]代替了选择性搜索网络[40],对输入的整张图进行特征提取，再把候选框映射到卷积层上进行分类回归。Faster R-CNN 做目标检测的一个缺点就是没有考虑带文本自身的特点，文本行一般以水平长矩形的形式存在，而且文本行中每个字都有间隔。针对这个特点，算法 CTPN （Detecting Text in Natural Image with Connectionist Text Proposal Network）[41-43] 提出一个新奇的想法，它把文本检测的任务拆分，第一步检测文本框中的一部分，判断它是不是一个文本的一部分，当对一幅图里所有小文本框都检测之后，将属于同一个文本框的小文本框合并就可以得到一个完整的、大的文本框，也就完成文本的检测任务。CTPN 中还加入 RNN 来进一步提升效果。把文本检测切割成多个阶段来进行无疑增大了文本检测精度的损失和时间的消耗，对于文本检测任务上中间处理越多可能效果越差，所以 EAST （Efficient and Accurate Scene Text Detector）[44] 算法实现了优雅且简洁地完成多角度文本检测，首先 EAST 采取了 FCN（Fully Convolutional Networks）的思路，借助 FCN 的架构做特征提取和学习，最终还是一个回归问题，在 EAST 最后预测出相应的文本行参数。

EAST 网络分为特征提取层、特征融合层、输出层三大部分。特征提取层是将提取的特征送入卷积层，抽取不同层级的特征得到不同尺度的特征图，目的是解决文本行尺度变换剧烈的问题，尺度大的层可用于预测小的文本行，尺寸小的层可用于预测大的文本，特征合并层将抽取的特征进行合并。合并规则是从特征提取网络的顶部特征按照相应的规则向下进行合并，网络输出层最终输出有 5 大部分，包括旋转角度、位置坐标、参数等。

1.2.5 文字内容识别

文字识别的第二步是对已检测出的区域进行文字内容识别。

由于传统文字与自然场景下文字有很大的不同，直接将传统的文字识别运用于自然场景下文字识别将会产生大量的错误，为了解决这些问题，Sawaki 等人提出了一种根据自然图像特征自动创建字符模板的方法[45]，Zhou 等人采用表面拟合分类器和特殊设计的字符识别算法对网络图像中的字符进行识别[46,47]，然而这些算法并没有在复杂的自然场景图片上进行评估，因此这些方法的适应性还没有得到充分的验证。在文献[48]中，de Campos 等人对当前计算机视觉和模式识别中常用的特征和分类算法进行了测试、比较和分析，并发布了一个 Chars74K 的图像数据集用于评估字符识别算法。Chars74K 在自然图像字符识别领域得到了广泛的应用，然而与主流的汉字识别方法不同的是主流汉字识别方法是以文字为基本单元，Campos 的方法只考虑单个字符识别的问题。Mishra 等人提出一种自底向上和自顶向下的方法进行场景文本识别[49]，由于自然场景中存在复杂的背景，很难直接将人物从局部背景中分割出来，该方法利用滑动窗口检测可能的字符，并将检测结果作为自底向上的信息处理。自顶向下的信息来自于大型词典的统计数据，通过条件随机场将自底向上和自顶向下的信息集成到一个统一的模型中，该方法的优点之一是可以允许字符检测中出现错误。文献[50]提出一种基于字符序列的编码模型(CHAR)，该模型假设所有图像都具有相同的大小并且存在最大可识别字符数量 k，但对于较长的单词，单词中只有 k 个字符能够被识别出。

现今基于深度学习框架下文字内容识别算法有两大主流技术 CRNN（CNN+RNN）算法[51]和 Attention 算法[52]。这两大方法主要区别在于最后的输出层，即怎么将网络学习到的序列特征信息转化为最终的识别结果，这两大主流技术在其特征学习阶段都采用了 CNN+RNN 的网络结构，CRNN OCR 在对齐时采取的方式是 CTC 算法，而 Attention OCR 采取的方式则是 Attention 机制。

CTC[53]（Connectionist Temporal Classficatio）模型使用 seq2seq 的 CTC 损失函数用于模型的训练并输出一系列字符。 CTC 采用一个端到端的网络，无需对训练数据进行预先分隔或者对输出数据进行处理，而是直接对输入序列进行建模直接得到输出序列。 CTC 网络最基本的思路是首先将输出字符集中添加 blank 占位符；然后对输入图片进行多尺度地分割和识别得到包含 blank 占位符的中间结果及其概率；再将包含 blank 占位符的中间结果映射到不包含 blank 占位符的所有可能的输出序列，并以可能的输出序列为单位求和；最后按照可能的输出序列所对应的概率进行排序，概率最大的即为输出序列。

1.3 本论文的主要研究内容

通过调研国内外现状我们知道，虽然中插广告的发展迅猛，未来很有可能取代传统广告，但是目前还没有能识别该广告的系统，所以本文主要设计一种中插广告的识别系统。本文主要针对以下三点挑战进行研究。

(1) 如何进行视频的镜头切分。切分镜头的前提是检测镜头边界，传统视频中的剧情镜头与广告镜头间的差异较大，仅仅靠颜色差异即可区分；而本文中剧情镜头是以渐变方式连接广告镜头，因此两个镜头的界限比较模糊，传统镜头切分的方法不再完全适用，所以本文系统在镜头边界检测上存在很大的挑战。通过对大量剧情与广告镜头边界的观察，我们找到了一种新的颜色变化特征，我们利用颜色变化趋势切分渐变镜头。

(2) 如何分类广告镜头和剧情镜头。由于广告中的演员是剧中的角色，广告背景和剧中场景也十分相似，因此广告镜头和剧情镜头存在很高的相似度，仅仅简单的表像特征已经不能够将二者进行区分，基于机器学习的模型也不能很好的提取到有用的差异特征。本文采用深度学习模型组合的方法来提取高维度特征，通过不同模型进行互补从而使模型能够学习到中插广告镜头和剧情镜头的差异特征，并强化性能影响显著的特征向量。

(3) 如何识别广告镜头中的广告内容。传统广告内容识别可以通过 Logo 匹配的方法进行识别，但是中插广告中的 Logo 并不明显。基于对中插广告的观察，大量广告以文字替代 Logo 标识广告商品，但是自然场景下的文字识别存在很大的背景干扰难题，而且有些广告中并没有出现广告文字，仅仅通过声音和视觉的变化对产品进行宣传，因此本文将文字识别算法进行改进，并结合音频特征匹配进行广告内容识别。

1.4 本论文的主要贡献

为了实现对中插广告的识别，本文做出了以下贡献。

(1) 针对视频中镜头间渐变情况提出一种新的镜头切分方法。由于中插广告和剧情的场景相似，因而在剧情镜头与广告镜头之间较多采用镜头渐变切换，相比传统广告采用的镜头突变切换，镜头切分更为困难。基于对中插广告的观察，发现渐变过程中会出现黑镜头，本文跳出计算帧间距离的常规思路，提出一种简单有效的利用颜色变化趋势切分渐变镜头的解决方案。

(2) 针对广告镜头与剧情镜头的视频特征高度相似难以区分的问题，本文提出利用 LSTM 网络和 Attention 网络组合，获取音视频时序高维特征并强化显著特征，

改善视频镜头分类性能。传统的广告识别系统采用 CNN 网络进行音视频特征深度表达，未利用视频帧序列的时序相关性，本文利用 LSTM 来获取前后帧的时序关系；进一步，本文使用了 Attention 网络获得不同维度特征在镜头分类结果中的占比，强化性能影响显著的特征向量。实验表明本文提出的方法分类准确率可以达到 88%，相比传统的机器学习方法提高了 4%。

(3) 针对广告中 Logo 不显著甚至没有 Logo 导致基于 Logo 的广告识别方法无效的问题，本文提出了结合文字识别和音频特征匹配的广告内容识别方法。基于对中插广告的观察，大量广告以文字替代 Logo 标识广告商品，也有一些广告甚至没有文字仅以声音标识广告商品。因此本文采用 OCR 文字识别技术结合音频色谱图特征匹配的方法进行广告内容识别，准确率可以达到 98%。

1.5 本论文的组织结构

本文下面章节的组织结构如下。

第二章将详细介绍本文的技术背景及相关知识。包括搭建开发平台，安装开源软件库，镜头边界检测以及本文使用的一些重要的机器学习算法和深度学习算法。

第三章主要介绍系统构成以及实验数据集。首先是设计该系统的原理和目的以及该系统每一个模块的作用和实现方法，接着是实验数据集的准备，重点讲述了如何将视频切分成镜头即镜头边界检测的方法，然后对镜头数据集进行标注，将镜头分为广告镜头和非广告镜头。

第四章介绍了镜头特征的提取和融合以及镜头分类的方法。首先描述了本章节要解决的问题即将镜头分类为广告镜头和非广告镜头，然后描述了如何使用卷积神经网络对镜头提取特征以及音频特征与图像特征融合的几种方法，最后是镜头分类过程，本文使用了机器学习的四种算法和深度学习算法进行分类，实验结果表明深度学习的方法取得了更高的准确率。

第五章介绍了对视频中广告内容的识别方法。由于广告 Logo 已经变得不再显著，文字特征比 Logo 更容易识别，有一些广告甚至没有文字仅以声音标识广告商品，因此本文采用 OCR（Optical Character Recognition）文字识别技术结合音频色谱图特征匹配的方法进行广告内容识别。

第六章是对整篇论文进行一个概括。阐述本文的工作成果，说明本文主要贡献，同时结合现有的一些技术发展说明本文工作的局限性与不足，然后阐述未来可以改进的地方以及对未来工作的展望。

2 技术背景

本章首先介绍本文系统应用到的开发平台，包括视频处理平台、机器学习平台和深度学习平台；由于本文系统是在镜头级别上设计完成，而镜头切分的前提是检测镜头边界，因此接着介绍几种传统镜头边界检测方法；最后是对机器学习四种算法：支持向量机（SVM）、随机森林（RF）、极端梯度提升（Xgboost）、梯度提升树算法（GBDT）以及本文用到的深度学习算法和文字识别算法分别进行介绍。

2.1 开发平台

本文实验所需要的开发平台主要有三个，一个是视频处理平台，需要将视频进行分帧以及读取图片信息，其余两个是机器学习和深度学习实验平台 Anaconda 和 Pycharm，本章还会介绍一些工具的安装，如：深度学习框架 Tensorflow 和 Keras 等。本文使用的电脑系统为 Ubuntu16.04，所有平台均搭建在 Ubuntu16.04 系统下，下面将详细介绍平台的搭建过程。

2.1.1 视频处理平台

帧是组成视频的一系列连续的图像，本文镜头边界检测也是检测前后帧之间的变化，因此需要先将视频切分成连续的帧图像。本文采用的是开源视频处理软件 FFmpeg，FFmpeg 可以传输或者转化音频和视频，不仅使用简单，其安装也非常方便。首先需要去 FFmpeg 的官网下载源码包，然后解压到本地电脑，在终端进入到安装路径下，执行“make install”即可完成安装。

2.1.2 Anaconda 简介

Anaconda 是 Python 的一个发型版本，其集成了涉及数据处理、机器学习、深度学习等多方面的 180 多个开源包，只需安装 Anaconda，便可通过指令直接调用集成在 Anaconda 里的 180 多个开源包，其操作简单功能丰富性使得很多人都愿意使用。

Ubuntu 下安装 Anaconda 的步骤如下：

（1）从 https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/ 下载 Anaconda3。

（2）在命令行执行下面两条语句，加清华镜像

conda config --add channelshttps://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/

conda config --set show_channel_urls yes

（3）创建 conda 环境，

比如：conda create -n ds35 python=3.5

（4）激活 conda 环境：source activate ds35

（5）在该环境下安装各种 conda 官方支持软件

conda install jupyter

2.1.3 Pycharm 简介

Pycharm 也是一种 Python 集成开发软件，它带有许多在使用 Python 进行开发时可以方便使用的工具，比如 debug、不同语言不同颜色、项目整体规划等。除此以外，Pycharm 还提供了一些更有效的功能，支持数据库框架下的网页的开发，如图 2-1 所示可以分为菜单工具栏、项目结构区、代码区、信息显示区 4 个区。相比于 Anaconda，Pycharm 可以建立一个项目，将相关的文件直接放在统一一个文件夹下即可通过语句互相调用，而且 Pycharm 的视觉效果会比 Anaconda 要好，但是 Anaconda 中的类库相比 Pycharm 比较全，因此本文使用时先安装的 Anaconda，然后将 Anaconda 配置到 Pycharm。

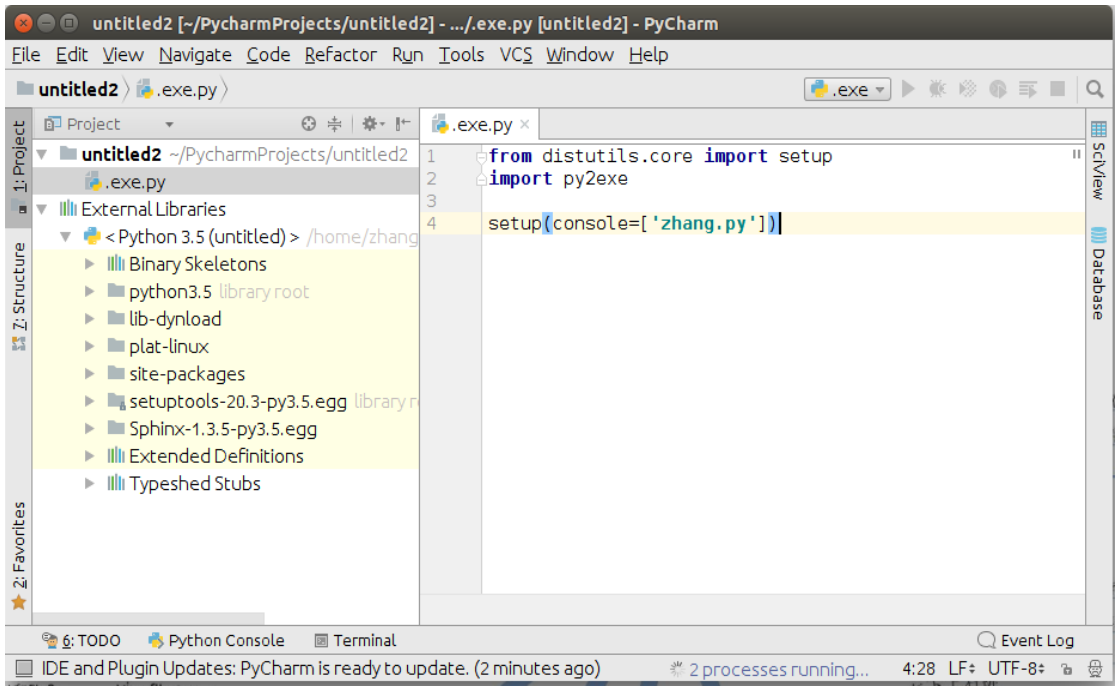


图 2-1 Pycharm 界面

Figure 2-1 The interface of Pycharm

2.1.4 Scikit-Learn 库

在机器学习库中，最常用的就是 Scikit-Learn 库，它集成了机器学习中许多经典的算法，并且将这些算法封装成类，用户只需要简单的调整一些参数，便可以将算法应用于自己的数据集中，比如本文用到 SVM、RF 等算法均可以通过 Scikit-Learn 库调用，在 Scikit-Learn 的 SVM 算法中，用户根据自己的数据可以选择内核类型，通过改变核函数参数，惩罚参数等找到该算法的最优解。对于随机森林算法，可以调整森林数数目和树的深度等参数。在数据集上，用户可以通过 Scikit-Learn 库中的数据分隔函数自行分配训练集和测试集的比例，除此以外，Scikit-Learn 还集成了数据处理、特征选择等功能，因此，这是一个高度集成，易于操作的 Python 库，本文中所使用的机器学习方法大多来源于 Scikit-Learn 库。Scikit-Learn 库安装时需要 NumPy 和 SciPy 等其他包的支持，需要先装 Numpy\Scipy\Matlotlib 包，再安装 Scikit-Learn，在终端输入“pip install Scikit-Learn”即可安装成功。

2.1.5 Tensorflow 开源软件库

Tensorflow 是 Google 的第二代机器学习系统，是由谷歌研究员和算法工程师开发，用于机器学习领域和深度神经网络的研究，它是采用了数据流图，可以加将复杂的数据结构传送到神经网络结构中进行分析处理，目前主要应用于图像识别、文字识别、语音处理等领域。

在编程时，我们都是编写一步运行后可以得到一个结果，使用 TensorFlow 时，首先需要构建一个计算图，按照计算图执行计算得到最终计算结果。TensorFlow 的计算图主要分为两个部分，一个是构造部分，该部分包括计算流程图，另一个是执行部分。Tensorflow 的安装比较简单，在安装 Anaconda 基础上，终端输入“conda install -c conda-forge Tensorflow”即可安装成功。

2.1.6 Keras 开源软件库

Keras 是基于 Tensorflow 的深度学习框架，它是用 Python 语言编写，是一个高层神经网络 API，可以极大减轻用户的工作量，Keras 主要包括 14 个模块包，而且这些模块都是独立分割的，用户可以使用它们自己来构建模型。Keras 只需仿照现有模块编写自己的类和函数，便可增加一个新模块，许多先进的研究工作均使用 Keras 创建新模块。

Keras 的核心结构是 model，它有两种不同的模型，一种是 Sequential 顺序模

型，一种是函数式 API。常用的模型是函数式 API 模型，而 Sequential 顺序模型是函数式 API 的一种特例，这里以 Sequential 顺序模型为例，讲述该模型下代码的构成过程：

（1） model = Sequential()

（2） model.add(Dense(32, activation='relu', input_dim=100))

（3） model.add(Dense(1, activation='softmax'))

（4） model.compile(optimizer=' adam',loss='squared_error')

Keras 和 Tersorflow 的使用方法基本相同，即首先需要构建一个计算图，然后把需要计算的输入值放入后，便可在整个模型中形成数据流，进而形成输出结果。Keras 相比 Tensorflow 更容易上手，因为它在 Tensorflow 基础上进行了封装，因此只需要直接调用相关函数即可，但有时候我们需要对某种算法进行一些修改，此时我们就需要用 Tersorflow，本文中即使用了 Tersorflow 也使用了 Keras。Keras 的安装也十分简单，在已经安装了 Anaconda 基础上，只需要在终端输入“conda install -c conda-forge keras”命令即可安装成功。

2.2 镜头边界检测方法

本文的广告镜头分类与内容识别是在镜头级别上进行的，因此需要先将视频切分成镜头，切分镜头时需要先将镜头边界检测出来，因此本节主要介绍几种传统镜头边界检测方法。

2.2.1 颜色直方图

颜色直方图是一种基于 RGB 的颜色特征，它描述了不同种类颜色在一幅图像中所占比例，这种方法并不在乎每种颜色所处的空间位置，所以不能用于图像中的目标检测。颜色直方图可以描述一幅图像中的颜色数量特征，可以反映图像的基本色调和颜色的分布情况，直方图只包含某一种颜色出现的频率，但是丢失了像素的空间位置信息，任何一幅图像都有其自己唯一与之对应的直方图，但是两张图片可能会有相同的颜色直方图，因此颜色直方图对于图片的旋转、变化等并不敏感，颜色直方图比较适合那些不需要考虑空间位置关系的图像。

彩色直方图的有两种表达形式，一种是单一直方图，另外一种是三维直方图，三维直方图比较简单，即三个维度分别对应红蓝绿三种颜色，定义三个直方图的某个像素点的 RGB 频率为 HR, HG，HB，某像素点 P 的 RGB 为(4, 231,129)，则直方图的计算为：HR[4]+=1，HG[231]+=1，HB[129]+=1，遍历图像上所有的像

素点后，三维彩色直方图就生成了。

对于单一直方图，由于三个通道各有 256 个颜色，如果直接生成颜色向量，向量长达到了 1600 万，计算比较困难，因此我们首先需要设定一个分割，即每个通道分割成等比例的几部分，在该范围内的像素属于同一颜色特征，即假设分为 n 份时，直方图索引值(index)分别为: （0， 4，16）， 根据 2-1 索引公式计算索引值。

$$\text{index} = R + G*256/n + B*256/n*256/n \tag{2-1}$$

对应的直方图 index = 0 + 4*16 + 13*16*16 即SH[3392]+=1，遍历图像上所有的像素点后，单一直方图就生成了。SH[3392]+=1，接下来就是求取两个图片之间的相似度或者距离。一般采用巴氏距离，计算公式如 2-2：

$$\rho(p, p') = \sum_{i=1}^{N} \sqrt{p(i)p'(i)} \tag{2-2}$$

其中ρ, ρ'分别代表两张图片的直方图数据，遍历相同i的数据点乘积开平方以后再相加，计算的结果即为图像的相似度值，范围为 0 到 1 之间。

2.2.2 边缘检测

边缘一般是指图像的某一发生剧烈变化的局部区域，边缘检测是图像处理中常见的一种方法，其目的是找到图像中变换比较明显的点，这些变换包括：表面方向的不连续、深度上的不连续和场景上的照明变化。图像边缘检测大大减少了计算数据量，保留了图像比较重要的结构特征，剔除了不相关的信息。

边缘检测的方法大致可分为两类，一种是基于搜索，这种方法首先需要计算边缘强度，然后再估计边缘的局部方向，一般采用梯度方向，并在此方向上找到局部梯度模的最大值，另一种是基于零交叉，该方法是找到图像的二阶导数的零交叉点定位边缘，一般用非线性微分方程或拉普拉斯算子的零交叉点。

这里介绍一种 Sobel 边缘检测算法，该方法是基于搜索的一种边缘检测算法，首先需要使用索贝尔算子（Sobel operator）产生图像上某点对应的灰度矢量g_x和g_y，接着计算横向及纵向边缘检测的图像灰度值G_x及G_y：

$$G_x = g_x * A \tag{2-3}$$

$$G_y = g_y * A \tag{2-4}$$

A 为原图像，接着通过以下公式结合，来计算该点灰度值：

$$P(p,p')=\sum_{i=1}^{N}\sqrt{p(i)p'(i)} \tag{2-5}$$

如果梯度 G 大于某一指定的阀值，则认为该点是边缘点，然后可用公式 2-6 计算梯度方向：

$$\theta = \arctan\left(\frac{G_y}{G_x}\right) \tag{2-6}$$

2.2.3 曼哈顿差分距离

曼哈顿距离是在像素特征上计算相邻帧的变化的，首先，按照如下公式 2-7 计算 R、G、B 三个颜色空间的像素平均值：

$$P = \frac{1}{m\times n}\sum_{j=0}^{n-1}\sum_{i=0}^{m-1}p(i,j) \tag{2-7}$$

其中 P 表示某一通道的所有像素之和，$p(i,j)$表示在该通道(i,j)位置的像素值，m,n 表示图片尺寸。接下来按照公式 2-8 计算曼哈顿距离：

$$M(k) = \sum_{s\in\{r,g,b\}}\left|P_{s,k-1} - P_{s,k}\right| \quad k = 2,3,\dots,n \tag{2-8}$$

其中，$M(k)$表示第 k-1 帧与 k 帧的曼哈顿距离。使用计算曼哈顿距离的方法检测突变边界帧中的元素很有效，但对于渐变边界，由于边界帧是渐变的不容易区分。所以还需要计算曼哈顿距离的差分值的绝对值 DM，如公式 2-9 所示：

$$DM(k) = |M(k-1) - M(k)| \quad k = 3,4,\dots n \tag{2-9}$$

2.3 机器学习算法

机器学习，是一门多领域交叉的学科，它涉及学科甚广，包括概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。它是通过大量的已有数据来模拟或实现人类的学习能力，以此可以达到重新组织已有的结构并不断的改善提高其性能，机器学习中通过对大量数据中挖掘的规律进行建模并预测未来的方法就是机器学习算法。

机器学习的学习方式总体可以分为四大类，即“监督学习”、“半监督学习”“无监督学习”、“强化学习”。监督学习是指输入的数据都有明确的结果或者标签的，比如我们对一张图片进行分类，我们是知道这张图片的标签是猫还是狗的，监督学习的算法模型是通过不断的将自己的预测结果与真正的标签进行比对，不

断调整模型的参数，从而达到预测结果最大概率的接近真正的标签。半监督学习是指输入的数据有一部分有标签，而有一部分数据没有标签，算法模型通过学习有标签和无标签数据的差异进行预测。无监督学习的所有的数据集完全没有标签，无监督学习通过数据的内在规律，从而对数据进行预测。强化学习是通过“试错”的方式进行学习，通过与环境交互获得奖赏从而指导其行为，强化学习的目标是通过不断试错从而获得最大的奖赏。

2.3.1 支持向量机

支持向量机（Support Vector Machine, SVM）是监督学习的一种二分类算法，它通过寻找一个超平面 wx+b=0 将不同类别的样本分开，对于非线性的情况，可以通过核函数将低维度空间映射到高维，来达到线性可分的目的。支持向量机可分为三类：线性可分支持向量机、线性支持向量机以及非线性支持向量机。

对于线性可分的数据，通过不断优化来寻找一个超平面 wx+b=0，将正负样本分开，即对于正样本（实点）使其符合 wx+b>0，对于负样本使其 wx+b<0 ，有时候会存在着无数个这样的超平面，因此我们在保证正确分类的前提下，会在两侧寻找两个极限位置，即越过极限位置的时候就会出现错误分类的情况，因此两个极限位置之间的距离就是分类间隔，如图 2-2，我们需要找到是这个最大间隔存在的决策。

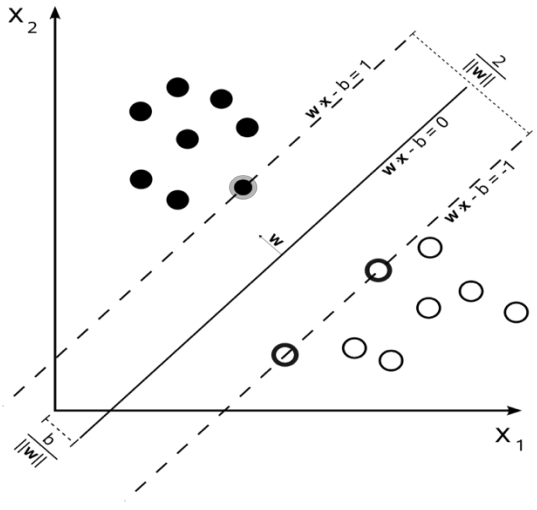


图 2-2 支持向量机模型

Figure 2-2 Support vector machine model

对于正样本 y=+1，wx+b>0，负样本 y=−1，wx+b<0，所以对于正确分类的样本，满足：

$$y(wx+b)=|wx+b| \tag{2-10}$$

因此函数间隔：

$$\gamma=y(wx+b) \tag{2-11}$$

根据点到平面的距离公式可得距离：

$$\frac{|\omega^{T}x^{(i)}+b|}{||\omega||} \tag{2-12}$$

则几何间隔表示为：

$$\frac{\gamma}{||\omega||}=\frac{y^{(i)}(\omega^{T}x^{(i)}+b)}{||\omega||} \tag{2-13}$$

因此求解几何间隔的最大值即求解如下：

$$max_{\gamma,w,b}\ \ \frac{\gamma}{||w||} \tag{2-14}$$

$$s.t.\ \ y^{(i)}\left(w^{T}x^{(i)}+b\right)\geq\gamma,\ \ i=1,2,\dots,m \tag{2-15}$$

令 γ=1，则几何间隔为：

$$\min\ \ 2/||w|| \tag{2-16}$$

$$s.t.\ \ y^{(i)}\left(w^{T}x^{(i)}+b\right)\geq\gamma,\ \ i=1,2,\dots,m \tag{2-17}$$

目标函数变为平方，则目标函数变为：

$$\min_{w,b}\ \ \frac{1}{2}||w||^{2} \tag{2-18}$$

$$s.t.\ \ y^{(i)}\left(w^{T}x^{(i)}+b\right)\geq 1,\ \ i=1,2,\dots,m \tag{2-19}$$

之所以目标函数变成 1/2 平方的形式，是因为 1/2 可以在求导数的时候消除平方，$||\omega||^{2}$的函数特性更好。以上所述为硬间隔，硬间隔 SVM 要求所有的数据点都要分类正确，但有时候可以接受错误分类几个数据点，来得到几何间距最大，此时函数变为：

$$\min_{\gamma,\omega,b}\ \ \frac{1}{2}||\omega||^{2}+C\sum_{i=1}^{m}\varepsilon_{i} \tag{2-20}$$

$$s.t.\ \ y^{(i)}\left(\omega^{T}x^{(i)}+b\right)\geq 1-\varepsilon_{i,}\ \ \ i=1,\dots,m \tag{2-21}$$

$$\varepsilon_{i}\geq 0,\ \ i=1,\dots,m. \tag{2-22}$$

其中 C 为惩罚因子，ε 代表松弛变量，松弛变量越大，表示样本点离超平面越近。

2.3.2 随机森林算法

随机森林（Random Forest，RF）是由多棵决策树组成，通过对样本进行训练进而预测的一种算法。它的输出也是由组成决策树以投票的形式决定的，如图 2-3 所示，随机森林主要应用于分类，但是也可以对数据进行回归预测。决策树主要由三部分构成：特征选择、生成决策树、剪枝，特征选择是根据某一量化评估指标从训练数据集中选择一个特征作为当前节点分列标准，评估指标有 ID3 算法，它是根据信息增益评估和选择特征，C4.5 算法是使用信息增益率来进行选择，CART 算法采用的是 Gini 指数作为分裂标准，由于选取的评估指标不同，所以生成的决策树也会有很多种。决策树生成是根据选取的特征评估标准，从上往下依次生成子节点，直到数据集不能再分后则停止生长，由于决策树比较容易过拟合，所以需要进行剪枝，一般有预剪枝和后剪枝两种，随机森林的构建过程如下：

(1) 从原始样本随机抽取 m 个样本，构建 m 个决策树；

(2) 假设在样本数据中有 n 个特征，每次分裂时根据特征指标选择最好的特征分裂每棵树都这样分裂，直到该节点的所有样例都被分到同一类别下；

(3) 然后让每颗决策树最大限度生长，不做任何修剪；

(4) 最终生成的多棵分类树组成随机森林分类器，用生成的分类器便可以对新的数据进行分类与回归。对于分类问题，由多颗决策树分类器进行投票决定最后分类结果；对于回归问题，则由多棵决策树预测值的均值代表最终预测的结果。

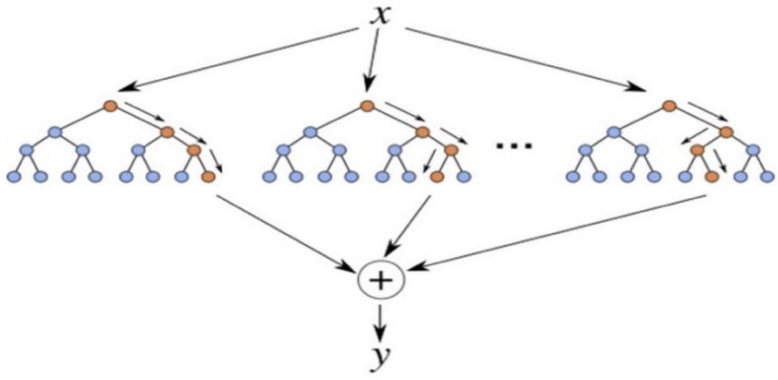


图 2-3 随机森林模型

Figure 2-3 Random forest model

随机森林有几个超参数，也是本文训练模型是时需要用到的。

n_estimators：这个参数表述随机森林中树的数量。 树的数量一般要根据数据集的大小来确定，数据集较大时，一般树的数量越多性能越好，此参数一般默认是 100。

n_jobs：表示允许使用处理器的数量，调节此参数可以加快模型计算速度。 若设置为 1，表示只能使用一个处理器。若设置为-1 表示对处理器数量无限制。

2.3.3 梯度提升树

梯度提升树 (Gradient Boosting Decison Tree,GBDT) 是集成学习中的一种，可以用来做分类和回归任务，由于本文只使用分类算法，因此只对分类原理介绍。

在 GBDT 的迭代中，假设在前一轮迭代得到的强学习器是:$f_{t-1}(x)$, 损失函数是$L(y,f_{t-1}(x))$, 则本轮迭代的目标是找到一个弱学习器$h_t(x)$，让本轮的损失函数$L(y,f_t(x))=L(y,f_{t-1}(x)+h_x(x))$最小，也就是说，通过不断迭代找到决策树，要让样本的损失变得更小。GBDT 的思想可以用生活中的一些例子解释，比如有件物品 100 元，我们首先用 50 元拟合，损失 50 元，接着我们用 30 元拟合剩下的损失，发现差距还有 20 元，接着用 15 元拟合剩下的差距，就只剩下 5 元了，我们还可以继续迭代，每进行一轮迭代，拟合的误差都会减小，GBDT 分类的计算步骤如下：
对于二元 GBDT 常使用对数似然损失函数为：

$$L\big(y,f(x)\big)=\log\left(1+exp(-yf(x))\right) \tag{2-23}$$

第 m 轮的第 i 个样本的负梯度误差为：

$$r_{ti}=-\left[\frac{\partial L(y,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}\ (x)}=y_i/(1+exp(y_i f(x_i))) \tag{2-24}$$

通过以上误差进行不断更新迭代，得到新一轮的决策树各个叶子结点的输出值：

$$c_{tj}=\underbrace{\arg min}_{c}\sum_{x_i\in R_{tj}}\log\left(1+exp(-y_i(f_{t-1}(x_i)+c))\right) \tag{2-25}$$

2.3.4 极端梯度提升

极端梯度提升（eXtreme Gradient Boosting，Xgboost）是 GBDT 的一种改进方法，同样应用于分类或者回归问题，Xgboost 相比 GBDT 算法在代开函数里加了正则项，便于控制模型的复杂程度，防止出现过拟合，xgboost 的损失函数的误差部分采用了二阶泰勒展开，同时用到了一阶和二阶导数，使得损失函数的计算更精准，Xgboost 得每颗子树都增加了一个参数 shrinkage，这样能够使每颗子树的权重得到降低，防止过拟合。

2.4 深度学习算法

深度学习是 2006 年 Hinton 等人基于对人工神经网络的研究提出的一种特殊的机器学习，其目的在于通过模仿人的大脑思维来解释数据，比如图像，声音和文本，深度学习可以通过底层特征的组合形成抽象的高层表示特征，以发现数据更高层次的特征表现。

同机器学习一样，深度机器学习也分为监督学习和无监督学习. 不同的学习框架下建立的学习模型也是不一样的。监督学习下典型的代表就是卷积神经网络（Convolutional neural networks，CNN），而深度置信网（Deep Belief Nets，DBNs）属于一种无监督学习模型。

2.4.1 卷积神经网络

卷积神经网络（Convolutional Neural Network, CNN）是一种特殊结构的神经网络。该结构含有一个卷积核，而神经网络是若干个感知机单元的集合，感知机是一个二元线性分类器。如图 2-4 所示，输入 x1 和 x2 分别和各自的权重 w1 和 w2 相乘相和，则函数 f=x1*w1+x2*w2+b（b 为偏置项）。函数 f 后会连接一个激活函数 g 来实现期望分类，最终输出分类结果 y。

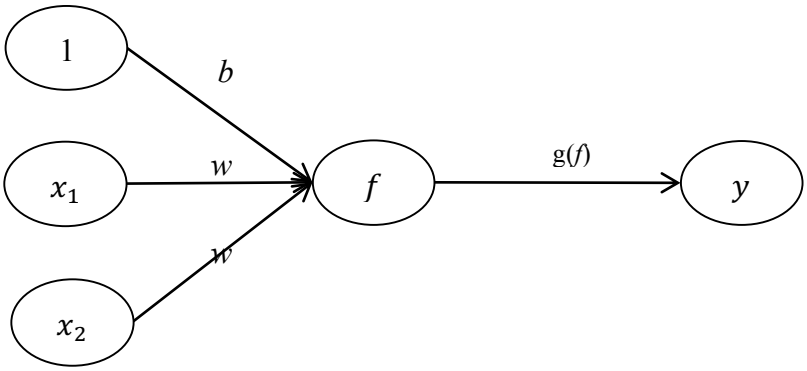


图 2-4 感知机算法

Figure 2-4 Perceptron linear algorithm

将多个输入堆叠在一起，这些单元的输出成为下一个单元的输入，再通过函数 f 和激活函数得到最后的分类。如图 2-5 为一个简单的神经网络。

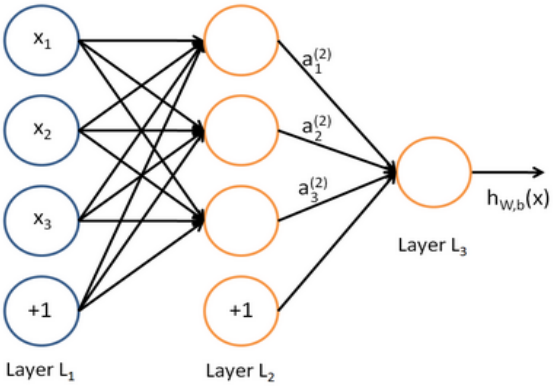


图 2-5 神经网络

Figure 2-5 Neural network

卷积神经网络在传统的神经网络上加入其它功能层，主要包括五层：输入层、卷积层、激励层、池化层、全连接层，图 2-6 为 CNN 的结构图。这里以图 2-6 为例介绍卷积神经网络结构，第一层输入图片，进行卷积（Convolution）得到第二层的特征图，接着进行池化，得到第三层深度为 3 的特征图。循环操作得到深度为 5 的特征矩阵，然后将这 5 个特征矩阵，按行展开连接成向量，输入全连接（Fully Connected）层，全连接层就是一个神经网络，最后得到最终结果。

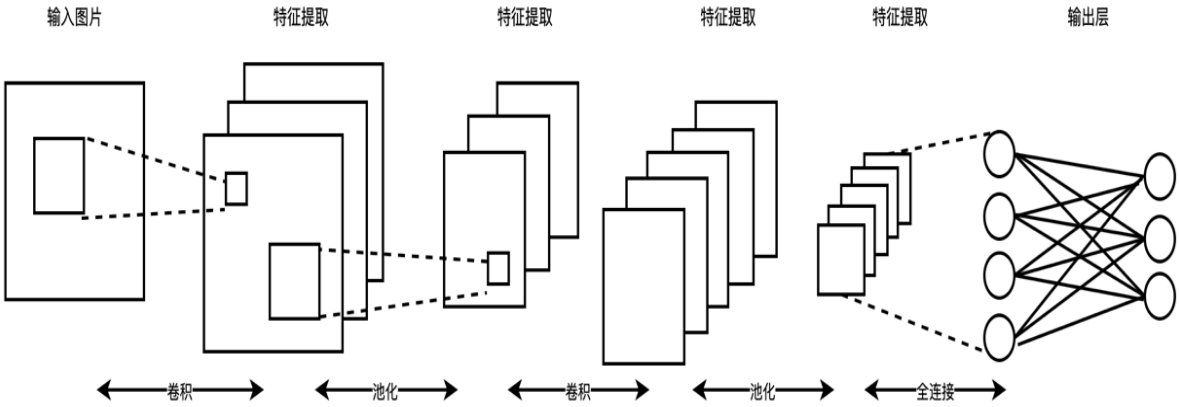


图 2-6 卷积神经网络结构图

Figure 2-6 Architecture of convolutional neural network

2.4.2 LSTM 网络

LSTM（Long Short-Term Memory）是在循环神经网络（Recurrent Neural Network,RNN）基础上的改进，解决了 RNN 不能长期依赖的难题。LSTM 的核心

在于细胞状态和门机制，接下来我们详细讲述。

首先是遗忘门f_t，遗忘门用来决定我们该忘记什么信息，它把上一次的状态 ht-1 和这一次的x_t相比较，通过 gate 输出一个 0 到 1 的值，1 代表记住，0 代表忘记。

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{2-26}$$

接着是记忆门，即哪些信息该记住，分为两步，第一步是用 sigmoid 函数决定什么信息需要被我们更新，忘记哪些旧的信息，第二步是用 tanh 构建一个新的细胞状态，更新后的细胞状态。

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{2-27}$$

$$\widetilde{C_t}=tanh(W_C\cdot[h_{t-1},x_t]+b_C \tag{2-28}$$

然后是更新门，把老的细胞状态更新为新的细胞状态，用 XOR 异或门和 AND 与门来更新细胞状态。

$$C_t=f_t*C_{t-1}+i_t*\widetilde{C}_t \tag{2-29}$$

最后是输出门，由记忆来决定输出什么值，这里的细胞状态已经被更新，于是通过这个记忆纽带，来决定输出值。

$$o_t=\sigma(W_o[h_{t-1},x_t]+b_o) \tag{2-30}$$

$$h_t=o_t*\tanh(C_t) \tag{2-31}$$

2.4.3 Attention 网络

Attention 网络是一种注意力机制，可以将注意力集中在那些对当前任务更重要的特征向量上。Attention 的应用非常广泛，可以应用到文本翻译上，在翻译过程，Attention 机制会专注输入序列中的一些可能会比较重要的词；还可以应用在图片描述上，可以帮助卷积神经网络重点关注图片的一些重要信息来生成序列，Attention 机制应用在语音识别上，比如输入一个英文的语音片段，输出对应的音素序列；应用于文本摘要，输入一篇英文文章，输出其摘要序列等。

本节以翻译“机器学习”为例，讲述 Attention 网络的实现原理。Attention 在翻译“machine”的时候，我们更加希望模型关注的是“机器”而不是“学习”，Attention 网络通过训练便可以帮我们把注意力集中到“机器”。

如图 2-7(a)所示，Attention 其实就是对当前的输入与输出的匹配度。h^0为当前时刻 RNN 的隐藏层输出向量, z^0为初始化向量， match 为输出向量和初始化向量进行匹配的模块，∂_0^1为 match 算出来的相似度。

如图 2-7(b)所示，得到匹配度之后，我们需要计算所有输入和当前的输出的相似度，我们使用 softmax 进行归一化。利用所有输入和当前输出的相似度，然后计算其加权向量和，作为下一次的输入。

如图 2-7(c)所示，算出c^0之后，把这个向量作为 RNN 的输入。然后第一时间点的输出z^1由c^0和z^0共同决定。计算z^1之后，替换之前的z^0，再和每个输入的 encoder 的向量计算匹配度，然后 softmax 归一化，接着计算向量加权，作为第二时刻的输入……如此循环直至结束。

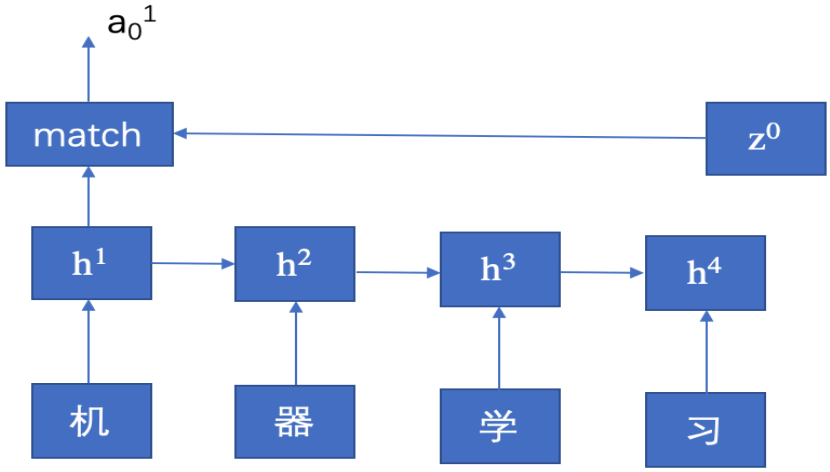
2.5 文字区域检测

文字识别是指将图片中的文字进行识别并输出，传统意义上的文字识别是指背景比较简单的文字，比如书本印刷的文字，背景比较干净只有文字，而本文的文字是指自然场景下的文字识别，即一张图片不仅仅有文字，还存在其它复杂的背景物体，自然场景下的文字识别主要分为两步，第一步是将文字区域进行检测，即本章节介绍的主要内容。

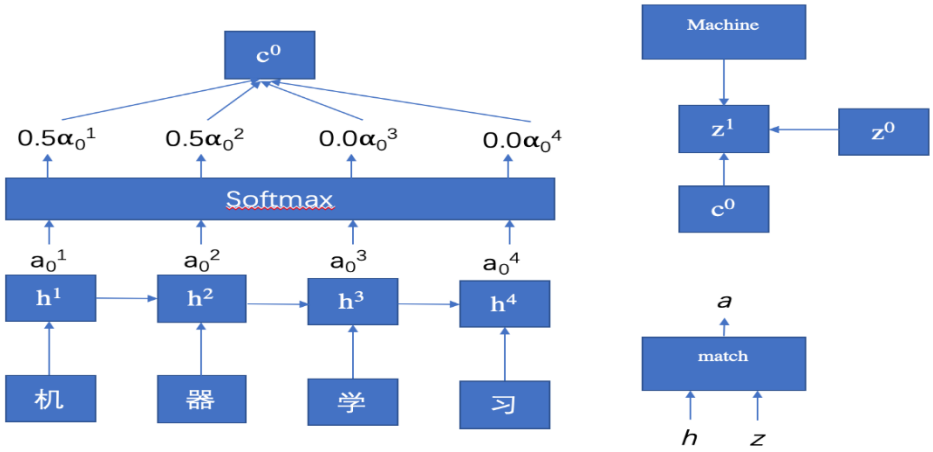
2.5.1 R-CNN 简介

R-CNN（Region-Convolutional Neural Network）是一种将候选区域和卷积神经网络结合起来的目标检测的算法，该算法主要由三个模块构成，第一个是 region proposal，即有可能是检测目标的候选区域，第二个是卷积神经网络，用于提取图片的特征向量，第三个是分类器，在本文中是一个二分类问题，即候选区域是检测目标和不是目标区域。

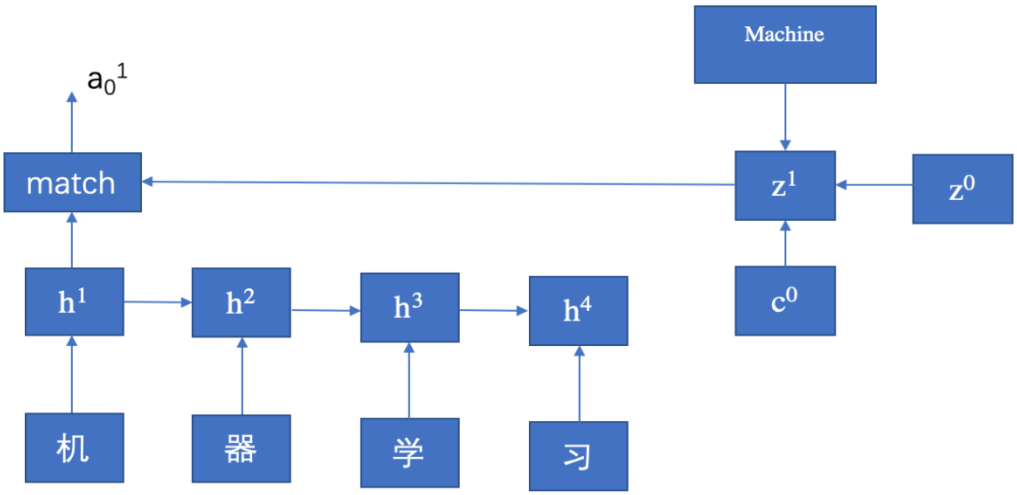
首先是候选区域的生成，R-CNN 使用了（选择性搜索）Selective Search 的方法，先将图像分割成小区域，接着合并两个最可能是一个区域的两个小区域，直到整张图像合并成一个区域位置，此时输出分割的 2000-3000 个小区域即为候选区域，接下来需要对候选区域利用卷积神经网络进行特征提取， R-CNN 采用的是利用 ImageNet 数据训练的卷积神经网络，将提取的特征输入分类器，输出该区域是否是目标区域。



(a)



(b)



(c)

图 2-7 Attention 网络应用实例

Figure 2-7 Application example in Attention network

2.5.2 Faster R-CNN 算法简介

Faster R-CNN 是在 R-CNN 基础上的改进，由于 R-CNN 的候选区域输入卷积神经网络之前需要对数据进行修建，使得一些数据变形，而且计算数据量比较大，而 Faster R-CNN 是先对一张图片进行整体特征提取，在整体特征上利用 RPN（Region Proposal Networks）网络划分候选区域，最后对候选区域进行分类和边框回归（Bounding box regression），具体步骤如图 2-8 所示：

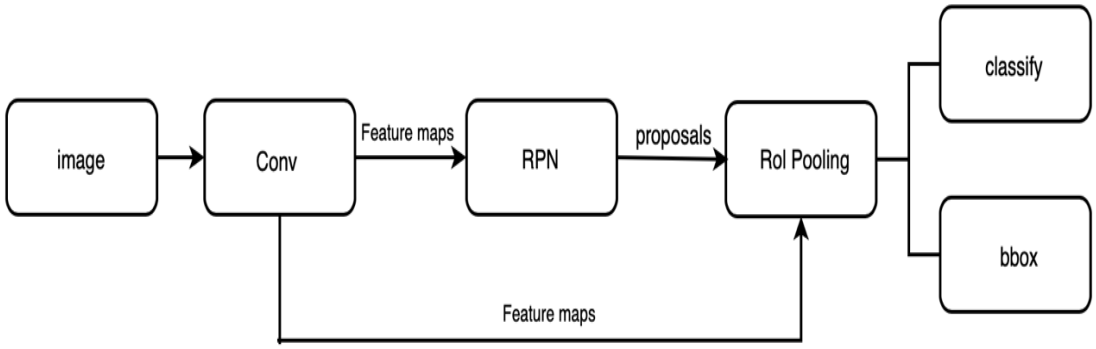


图 2-8 Faster R-CNN 算法流程图

Figure 2-8 Flow chart of faster region-convolutional neural network

2.6 文字内容识别

文字内容识别是文字识别的第二步，即对上一小节中得到的目标区域中的文字进行内容识别。

2.6.1 CTC 损失函数简介

CTC（Connectionist Temporal Classfication）主要用于解决输入数据与给定标签的对齐问题，即如何将识别到序列正确的输出。CTC 采用一个端到端的网络，无需对训练数据进行预先分隔，或者对输出数据进行处理，而是直接对输入序列进行建模，并且直接得到输出序列，CTC 网络最基本的思路是，首先将输出字符集中添加 blank 占位符；然后对输入图片进行多尺度地分割和识别，从而得到包含 blank 占位符的中间结果及其概率；再将包含 blank 占位符的中间结果映射到不包含 blank 占位符的所有可能的输出序列，并以可能的输出序列为单位求和；最后按照可能的输出序列所对应的概率进行排序，概率最大的即为输出序列。

2.6.2 CRNN 算法简介

CRNN（DCNN+RNN）是一种卷积循环神经网络结构，是由深度卷积神经网络（DCNN）和循环神经网络（RNN）组合的一种结构，主要用于解决图像的序列识别问题，特别是自然场景下文字识别的任务。

CRNN 包含三部分，从上往下依次是卷积层、循环层、转录层。卷积层主要利用 DCNN 网络进行特征提取，循环层由双向 LSTM 循环神经网络构成，可以预测特征序列中每一个特征向量的标签分布，从而得到所有可能结果序列的概率。转录层使用的是上小节中介绍的 CTC 模型，转录层主要作用是将上一部分 LSTM 网络输出的特征序列的预测结果进行整合，得到最终的文字识别结果序列。

2.7 本章小结

本章主要介绍了本论文使用的开发平台以及关键技术背景和原理，主要有：(1)开发平台和工具的安装和使用；(2)介绍了机器学习、深度学习的概念及其算法，包括 SVM、RF、Xgboost、GBDT 和几种神经网络；(3)介绍了文字识别的两步：文字区域检测算法和文字内容识别算法的原理。这些技术是本文进行研究的理论支持，也是本文研究中会用到的一些算法，这对理解本文的研究工作具有重要的作用。

3 系统设计与数据集

本章首先介绍本文提出的中插广告识别系统，包括系统设计思路以及每一部分的功能和作用，由于本文的系统是在镜头级别进行设计，所以接着介绍本文使用的镜头边界检测的方法，即如何将视频切分成镜头，最后是对带有中插广告的视频的分析以及本文镜头分类与识别的数据集，包括镜头图像数据集和镜头音频数据集。

3.1 系统设计

本文的视频时长 50 分钟左右，而广告时长只有 90s，在长时间的视频中，我们往往只关心很短的那几十秒的镜头内容，如果对整个视频进行处理识别，这样不仅会使结果存在误差，使结果偏向数据量大的一类，而且也会大大增加工作量。镜头是视频的基本组成单位，单个镜头不会破坏视频的主题完整性，所以本系统是在镜头级别上进行设计。已有传统系统[54]也存在镜头级别上进行广告识别，但只是检测突变镜头即可，而本文中除了突变镜头，还存在着渐变镜头需要切分，将视频切分镜头后还需对切分的镜头分类，最后对分类的广告镜头进行内容识别并输出广告内容，因此本文系统分为三部分：镜头切分，镜头分类，广告内容识别。系统结构图如图 3-1 所示，系统第一步是对视频切分镜头，镜头切分的前提是寻找到镜头的边界，根据检测到的边界切分视频，便可以得到镜头数据集，由于中插广告与剧情的界限非常模糊，因此本文基于对中插广告的观察，设计了一种新的镜头边界检测流程与方法，可以将模糊的边界进行切分。系统第二部分就是对镜头数据集进行分类，将镜头分为广告镜头和非广告镜头，由于剧情镜头和广告镜头的内容有一定的相关性，而且广告演员也是剧中的角色，因此仅仅依靠简单的特征不能完成该分类任务，需要提取高维度特征将二者进行区分。最后一步是对广告镜头的内容进行识别，本文采用了文字识别与音频匹配相结合的方法进行内容识别，由于一个镜头中包含着几百张图片，而且很多图片几乎是一模一样的，因此在内容识别时只需要选取其中的几张或者几个镜头进行识别即可，并将识别内容输出。

通过本文视频中插广告的识别系统，不仅可以检测出该视频中是否存在广告，同时还可以精确识别到具体广告内容，挖掘出广告中包含的丰富商业信息，进而帮助咨询机构分析广告主的经营状况。

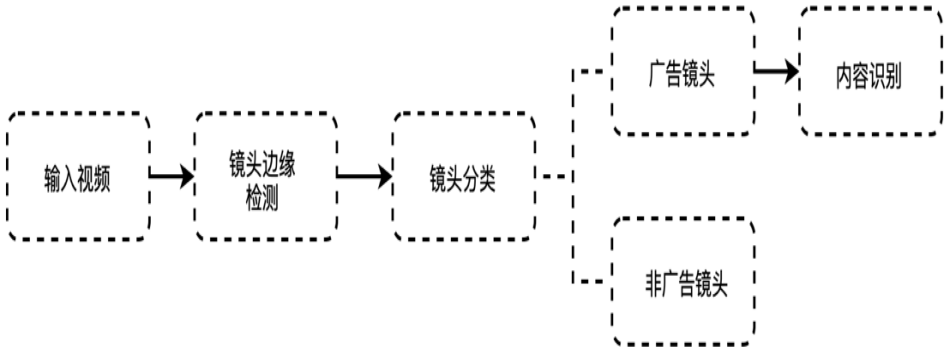


图 3-1 系统流程图

Figure 3-1 Flow chart of system

3.2 镜头边界检测

镜头切分的前提是找到镜头的边界，所以本节主要介绍如何进行镜头边界检测。镜头是指摄像头从开机到关机所记录下来的一段没有间隔的画面，镜头是组成一个视频的基本单位，对于本文一个视频样本，一集网剧的时常在 50 分钟左右，而广告时常只有 90s，我们只关心那几十秒的广告内容，并不需要对整个视频的内容进行分析，所以本文将视频进行切分镜头，只需要找到广告和非广告在镜头级别上的差异，并将其区分开即可，这样可以大大减少工作量。传统的视频中广告与剧情的分界线非常明显，只需要检测突变镜头即可，而本文中广告与剧情的界限变得模糊，二者的分界有一部分是渐变的，镜头边界情况更为复杂，而且剧情镜头和广告镜头中都存在渐变情况，传统的方法不能将这两种情况区分，因此需要寻找新的方法进行镜头边界检测，将视频进行镜头切分。

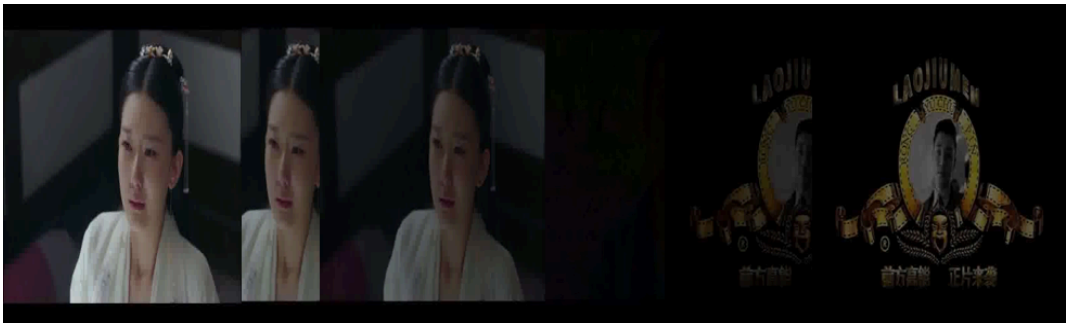
3.2.1 突变与渐变

镜头是视频的基本单位，视频是由拍摄者根据拍摄内容、场景、对话等镜头组合而成，由于不同镜头之间存在着不同的差异，因此在切分镜头的时候我们只需要找到连接镜头的差异帧，便可以找到镜头的边界。镜头之间的差异分为两种，一种是突变，一种是渐变，突变是指后一个镜头的头帧与前一个镜头的尾帧直接相连，中间没有过渡帧，连接突变镜头的边界帧之间无论是颜色还是亮度上均存在着明显的差异，如图 3-2 所示。而渐变是指后一个镜头的头帧与前一个镜头的尾帧之间存在着过渡帧，即边界帧，本文中的渐变有两种，一种是镜头间的渐变，一种是镜

头内的渐变，如图 3-3 (a)所示，这种渐变属于镜头间的渐变，主要是为了衔接两个不同主题内容的镜头，这些过渡帧之间的差异并不明显，旨在能够让观看者在最小视觉差异下连接两个镜头。镜头内部的渐变是指两个画面的内容主题是一样的，由于人物的转换或者关注点的不同，导致两个画面之间存在着渐变的情况，但是这两个画面共同构成一段主题内容，所以本文中将这样的画面称为镜头内渐变，如图 3-3 (b)所示。

图 3-2 突变镜头

Figure 3-2 Shot abrupt transition



(a) 镜头间渐变

(a) Between the shots gradual transition

(b) 镜头内渐变

(b) Inside the shots gradual transition

图 3-3 渐变镜头

Figure 3-3 Shot gradual transition

3.2.2 镜头边界检测

本文针对突变镜头和渐变镜头，提出了一种新的镜头切分方法，切分流程如图3-4 所示，首先对视频中的突变镜头进行切分，接着再对渐变镜头进行切分。

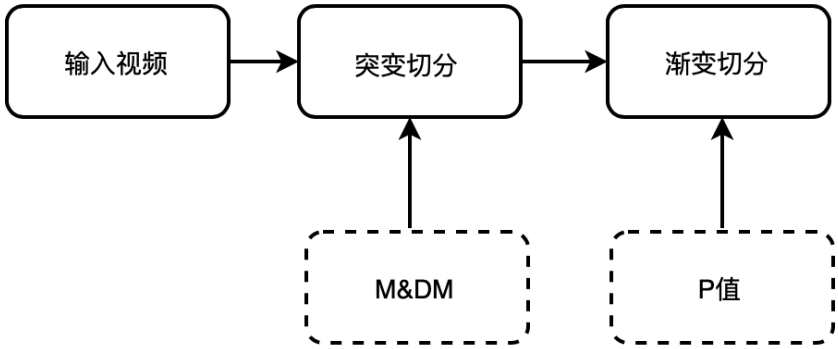


图 3-4 镜头切分步骤

Figure 3-4 The process of shot segmentation

首先我们利用第二章介绍的曼哈顿差分距离对输入的视频的突变镜头进行切分，切分步骤如算法 3.1 所示。

算法 3. 1 突变镜头边界检测

Algorithm 3.1 Shot abrupt transition boundary detection

算法 3.1 突变镜头边界检测

1：将视频分解成帧

2：计算每一帧的颜色特征

3：计算相邻帧的曼哈顿距离

4：计算每一帧的曼哈顿距离差分

5：求取阈值，切分镜头

如图 3-5 所示，该图横坐标为曼哈顿差分距离，即上述第 5 步待求取的阈值，纵坐标为在不同的阈值下切分镜头后的所有镜头内部帧的曼哈顿差分距离均值的平均值，由图可知，T=14 为最佳阈值，因此本文中选取了阈值大小为 14，即当相邻帧的曼哈顿差分距离大于 14 时，便将此帧认为是边界帧，从而得到突变镜头的边界。

由于本文视频中广告与剧情的界限非常模糊，以一种渐变的连接方式进行连接，所以接下来我们要针对本文的渐变镜头进行切分，本文存在两种渐变，一种是镜头间渐变，一种是镜头内部渐变，但是为了更好的进行镜头分类，需要保留镜头主题的完整性，所以本文并不希望把镜头内部的渐变帧进行切分。如图 3-6 所示为

镜头内部渐变（剧情）与镜头间渐变（由剧情进入广告）的 M(曼哈顿距离)和 DM(曼哈顿差分距离)，由图可知二者的差异不大，因此上述突变镜头的检测方法已经不能适用。

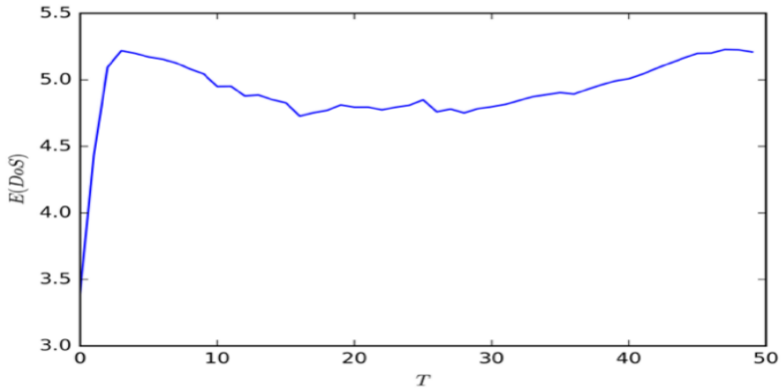


图 3-5 E(DoS)-T 关系图

Figure 3-5 Relation diagram of E(DoS)-T

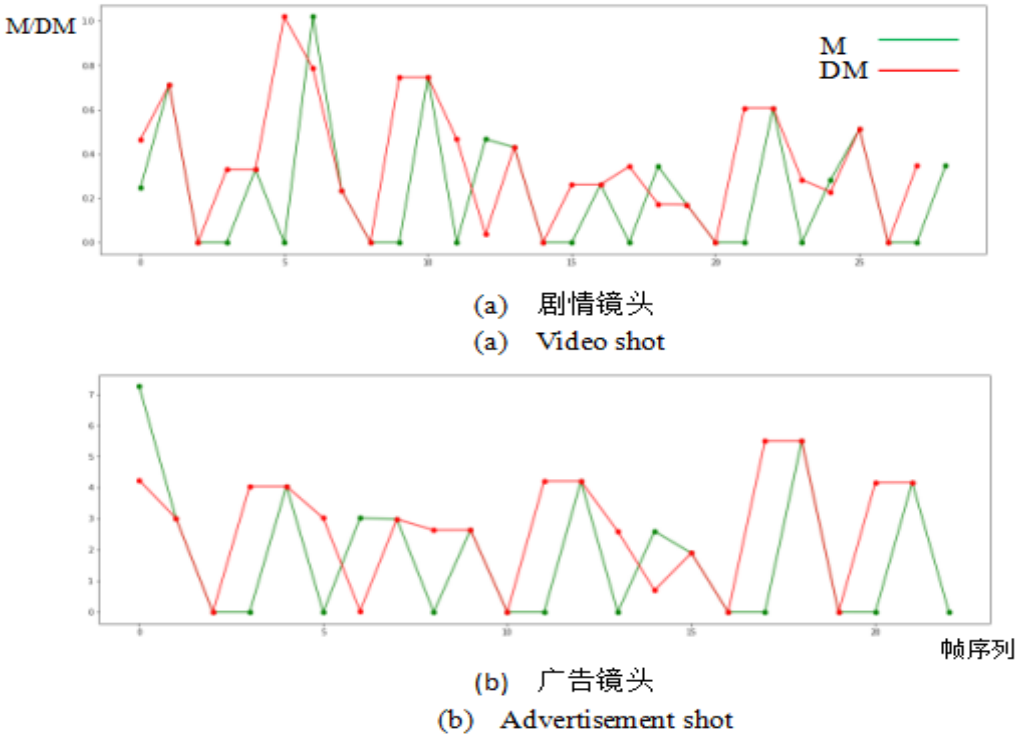


图 3-6 剧情和广告的 M 与 DM 值

Figure 3-6 The values of video and advertisement for M and DM

本文基于对镜头间的渐变的观察，发现了由剧情进入广告的渐变颜色变化是从亮变暗再到亮的过程，如图 3-7 (a)，即从黑色淡出前一个镜头，从黑色渐入下一个镜头，本文称为渐变黑镜头。而视频内部的渐变是两个画面在时空上的叠加，通常是同一场景同一内容下由于人物的转换或者关注点的不同从一个画面逐渐转移

到另一个画面的过程，如图 3-7 (b)所示。

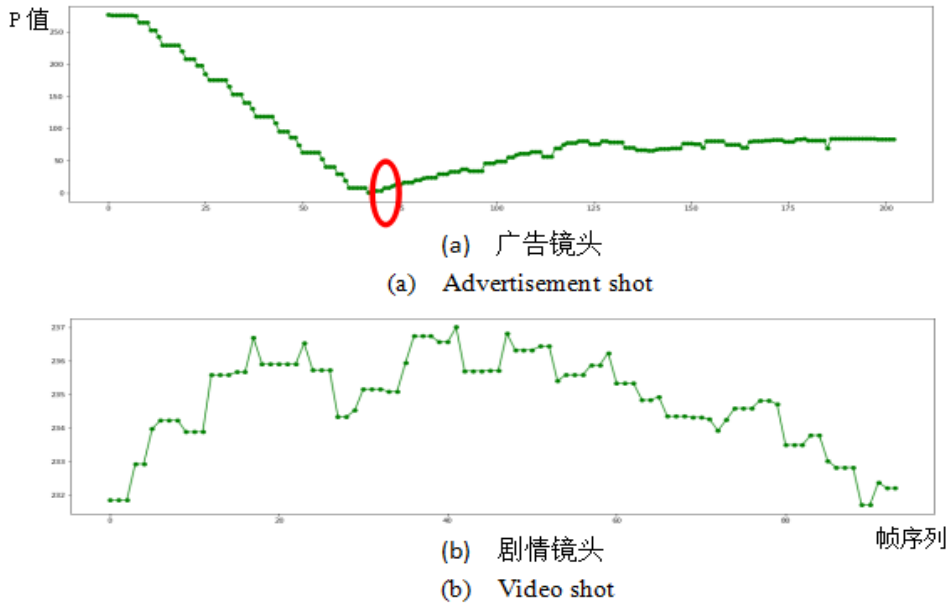


图 3-7 剧情与广告镜头的 RGB 变化

Figure 3-7 The RGB change of video and advertising shots

基于以上观察，针对中插广告特性我们提出了一种更简单的区分两种渐变的方法，首先对已经切分的每一个突变镜头的内部帧求颜色特征值 P，将镜头中不存在 P 值小于 50 的镜头剔除，因为此类镜头中不存在渐变黑镜头。我们对每个镜头的帧数进行了统计，如图 3-8 所示，发现大部分镜头包含帧数在 50-400 之间，因此我们对剩余镜头每隔 5 帧进行取帧，计算相邻帧的颜色特征值的变化斜率，在此基础上，本文选取了 6 个斜率值为一个信任度，如果前 3 个变化斜率为负，后 3 个为正，并且转换点的颜色特征值小于 30，我们便认为该帧为广告镜头间的渐变边界帧，切分步骤如算法 3.2 所示。

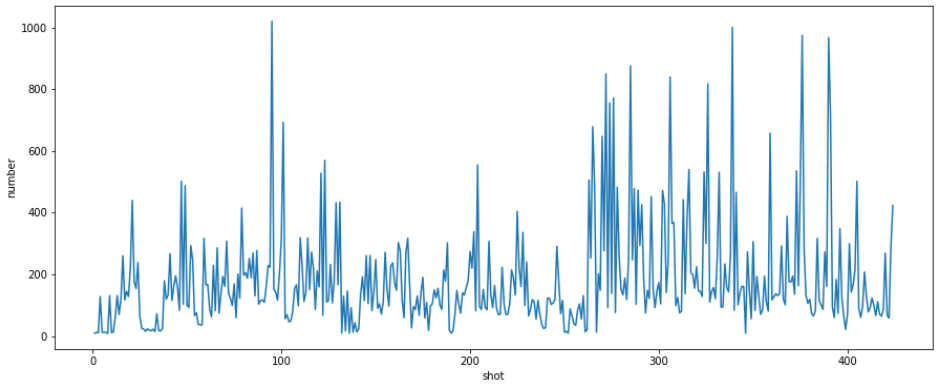


图 3-8 镜头内帧数

Figure 3-8 The number of images in a shot

算法 3.2 渐变镜头边界检测
Algorithm 3.2 Shot gradual transition boundary detection

算法 3.2 渐变镜头边界检测

1：求镜头内每帧图像颜色特征值 P

2：剔除镜头内不存在 P 值小于 50 的镜头

3：每隔 5 帧进行取帧，计算相邻帧的颜色特征值的变化斜率

4：由图 3-8 选取 6 个斜率值为一个判断间隔

5：判断前 3 个变化斜率为负，后 3 个为正

6：上述条件下判断转换点的颜色特征值小于 30，则为镜头边界帧

基于上述两步，先将突变镜头切分，再切分由剧情进入广告的渐变镜头的方法，可以解决本文中中插广告与剧情界限模糊的问题，能够正确的将广告与剧情切分开，同时保留了视频内部镜头的完整性，为镜头分类提供了更完整的数据集。

3.3 数据集介绍

本文使用的视频是带有中插广告的网剧，我们从腾讯视频、优酷视频以及爱奇艺视频等各大视屏网站下载了《老九门》、《白夜追凶》、《大军师司马懿之军师联盟》、《春风十里不如你》、《那年花开月正圆》、《虎啸龙吟》、《扶摇》七部电视剧，共 344 集电视剧，挑选出带有中插广告的剧集共 262 集，带有广告的这 262 集电视剧中还存在有些剧集里面广告内容一致的情况，我们将这种情况也进行了挑选，只保留了不同广告内容的电视剧集，因此经过筛选后满足条件的数据集共 142 集，广告视频集数占每个剧场总集数的比例如图 3-9 所示。本文广告播放的形式和传统的插播广告也有所不同，传统广告的播放形式是外插，即广告和电视剧是分割开的，而且通过视频 VIP 我们可以直接跳过广告，而本文中的广告是内插到电视剧里面组成电视剧的一部分，与电视剧连接成一体，即使 VIP 会员也没有跳过的权利，每集剧长 50 分钟左右，广告时常 90s 左右，广告时长分布如图 3-10 所示。本部分主要介绍视频切分后的镜头如何构建的镜头分类数据集，从而使分类器能够学习到广告镜头与非广告镜头的差异性，提高分类结果的准确性。

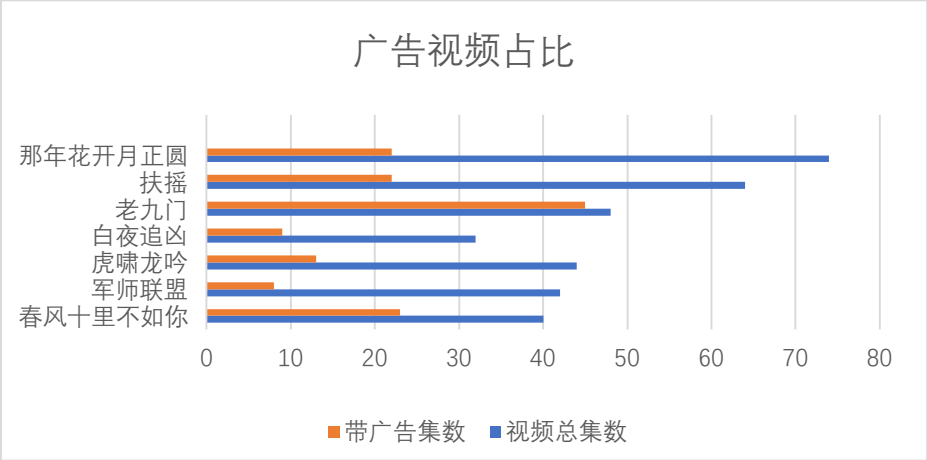


图 3-9 广告视频占比

Figure 3-9 The ratio of advertising video

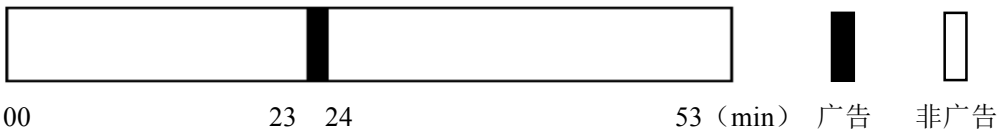


图 3-10 广告时长分布

Figure 3-10 Distribution of advertisng time

3.3.1 镜头图像数据集

本文采用的视频数据中的广告播放时间一般都在整集剧情的 2/5 到 4/5 时间段，并且一集电视剧中只播放一个广告，电视剧的开头和结尾都是和内容无关的开场曲和结尾曲或者一些情景回顾，为了减少镜头切分的工作量，再进行镜头切分之前，我们先将视频进行剪辑，我们将每集电视剧的前 5 分钟与后 5 分钟切除，只保留电视剧的剧情内容与广告内容，接着通过上一小节介绍的镜头边界检测的方法进行镜头切分，通过对突变镜头切分，我们将视频切分成 37845 个镜头，再通过渐变镜头的切分，最终视频镜头为 38256 个镜头，其中包括广告镜头 4620 个，非广告镜头 33636 个。由于在每一集中广告镜头是连续的，因此我们只需人工找到每一集广告镜头的开始镜头和结尾镜头然后通过文件夹命名的方式对广告镜头进行打标“ad”，同时将广告内容一并标注，将剩余的非广告镜头标注“not”。由上可知，广告镜头和非广告镜头的数据集极度不平衡，几乎相差 10 倍，这样在训练的时候很可能会产生模型会偏向样本数目比较多的类别，而“轻视”样本数量少的类别，模型在测试集上也会所有偏向，泛化能力就会减弱，因此本文结合两种方法解决样本不平衡的问题，一个是上采样，一个是下采样，上采样是指对较少样本进行

复制，从而达到扩充稀有样本的目的，但是这种方法容易导致过拟合，因此本文在上采样的时候均匀的对每集电视剧中的广告镜头进行上采样，并且控制复制样本比例为 10:1，经过上采样后的广告样本数量最终为 5482 个，但是和非广告样本进行比较广告样本还是十分稀有，所以本文又进行了下采样，即剔除一些非广告镜头，为了避免将集中在一集的非广告镜头剔除，本文对每一集的镜头都按照广告镜头与非广告镜头 1:1 的比例随机将非广告镜头进行剔除，最终用于镜头分类的数据集总共 11261 个，包括广告镜头 5482 个，非广告镜头 5779 个。

3.3.2 镜头音频数据集

本文镜头分类时所使用的特征包括两个：视频特征和音频特征，镜头的音频文件和视频文件是一一对应的，但是本文是先将视频切分成连续的帧，通过上节所述镜头切分方法，找到镜头边界帧将视频切分成镜头，因此构建音频数据的关键是找到镜头边界帧，本文利用 FFmpeg 将视频切分成连续帧，生成的帧按照视频时间顺序依次从 0 进行编号，因此我们可以利用图片的编号找到对应的视频，然后抽取其中的音频。本文采用的 FFmpeg 分帧的帧率为 20，即每秒切分成 20 帧，即相邻帧之间的时间间隔为 0.05 秒，因此将图片编号与 0.05 相乘即可以得到该镜头的开始和结束时间，这样便可以得到本文需要的音频数据集，与视频数据集一样，音频数据集包括广告镜头 5482 个，非广告镜头 5779 个。

3.4 本章小结

本章主要介绍了中插广告的识别系统设计与实现，系统主要包括三部分：镜头切分、镜头分类、广告内容识别，本章主要介绍第一部分镜头切分，镜头边界检测是切分镜头的前提，镜头切分有两大类，一是突变镜头，另一个是渐变镜头，而本文中存在两种渐变情况，一种是镜头间的渐变，一种是镜头内的渐变，本文需要将镜头间的渐变进行切分，因此根据现有数据集的特点：镜头间的渐变颜色变化由亮变暗再变亮，本文称为渐变黑镜头，设计了一种新的镜头切分方法，即先切分突变镜头，再利用渐变黑镜头切分渐变镜头，利用此方法得到镜头数据集。由于数据集中的广告时长相对剧情比较短，切分后的广告镜头和非广告镜头的数据存在不平衡的现象，本文利用上采样和下采样结合方法将数据集进行平衡，同时利用帧频和图像编号找到镜头的音频文件，最终得到镜头分类的数据集，包括视频镜头和音频镜头，其中广告镜头 5482 个，非广告镜头 5779 个。

4 镜头分类

本章主要对镜头分类的方法进行介绍，首先介绍镜头特征，包括视频特征、音频特征以及提取特征的过程，接下来介绍如何将两个特征进行融合，以达到让每一个特征信息发挥其最大作用的效果，最后是镜头分类的训练方法，包括机器学习的调参和深度学习的模型构建。

4.1 问题描述

传统广告的播放形式是插播，是将一段和剧情没有任何关系的广告插放在电视剧的某一时间段，不仅内容毫无关联，广告演员和剧场人物也不相同，尤其是在古装剧中，传统广告一般都是在现代场景下拍摄的，和古装剧集的场景有着很大的差异。而本文中的广告属于内插，即广告属于剧情内容的一部分，无法像传统广告一样通过 VIP 特权直接跳过，而且本文广告中的演员是剧中的人物，广告的内容也和剧情有一定的关联，无论剧情是古装剧还是现代剧，其中插广告的表现形式均与其对应，因此在分类中插广告镜头与剧情镜头上还存在着巨大的挑战。

目前，大多数文献都是对传统广告进行分类识别，对中插广告的研究大都停留在观测状态，目前还没有文献对该广告进行分类，传统广告的镜头分类大都是通过简单的机器学习方法就可以解决，因为传统广告和剧情的差异性比较大，现有的机器学习方法通过训练便可以提取到二者差异，并将其区分开，而本文的中插广告和剧情的差异微乎其微，简单的机器学习的方法已经不能适用，因此需要寻找高维抽象的特征来进行识别。所以本文便是通过将多个深度学习模型进行组合，通过多层卷积策略和非线性的计算方法，再通过反向传播的参数调节，使得模型能够识别广告镜头与非广告镜头的深层差异，并将其正确分类。

4.2 提取镜头的深度卷积特征

由于中插广告与剧情十分相似，简单的颜色特征已经不能将其进行区分，因此需要提取更高维度的特征。深度卷积特征是指通过一种高维的向量来表示该图片，像素是构成图片的基本单位，图片由一组数组构成，黑色图片由一组二维数组构成，彩色图片是由 RGB 三维数组所构成，以黑色图片为例，如图 4-1 所示为一张数字为 8 的图片通过卷积神经网络提取特征并对图片进行识别过程，该图是由(28，28)

的二维数组构成，将数组横向展开后变成一个(1，784)的数组，该数组通过卷积网络进行特征提取，最终通过 softmax 函数输出卷积神经网络的识别结果概率。

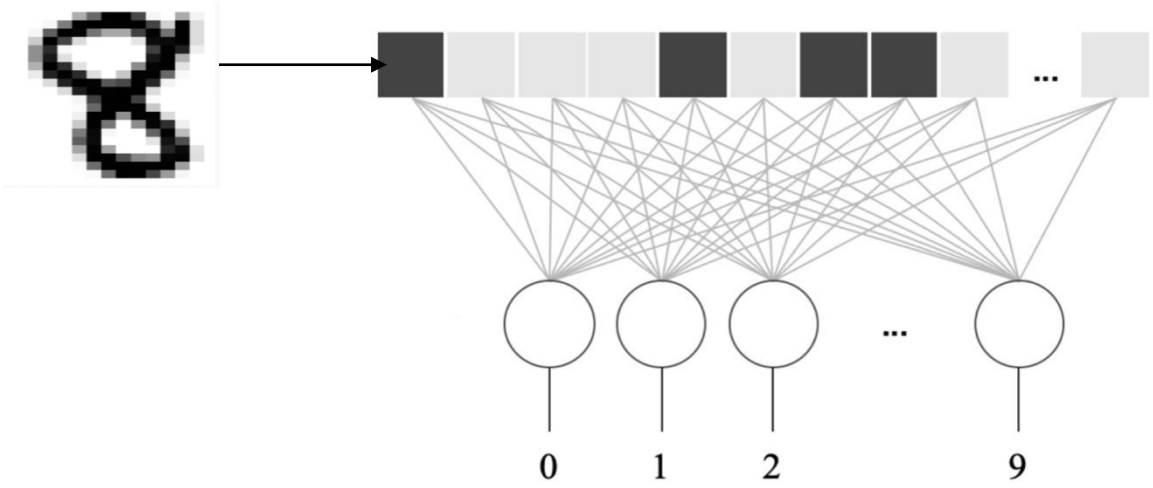


图 4-1 图片识别过程

Figure 4-1 The process of image recognition

如图 4-2 所示为提取该图像的卷积神经网络，该网络为 5 层，每层的单元数不同，下一层单元均可由上一层通过卷积计算得到。针对整张图片，每一层的关注点不一样，有的层提取的是背景，有的提取的是轮廓，有的提取色差，因此网络的层数也是该网络提取特征的关键参数。

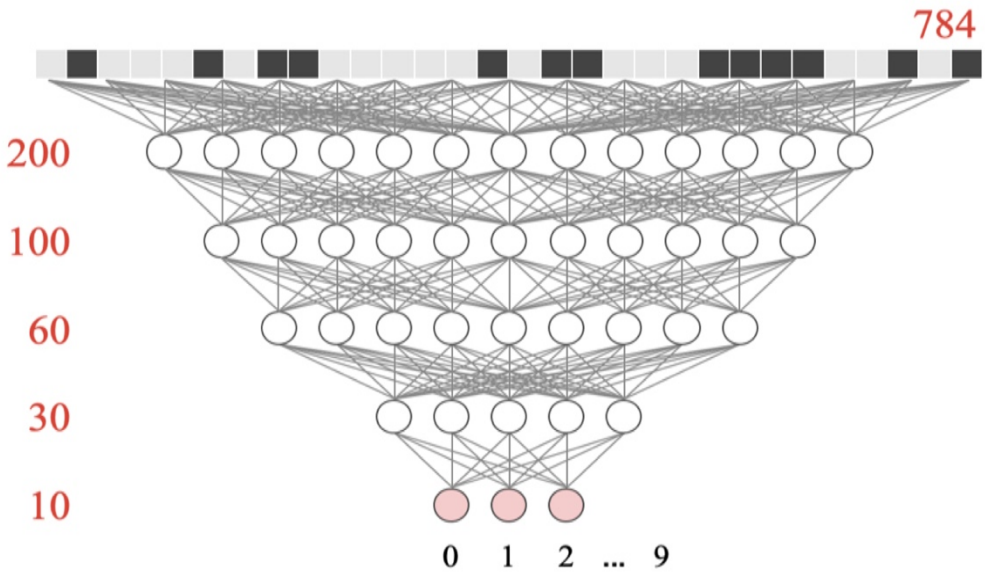


图 4-2 卷积神经网络

Figure 4-2 Convolutional neural networks

卷积在数字信号处理领域有广泛的应用，在图像处理时也时常用到，图 4-3 展

示了卷积计算过程，输入的数据可以视为一张图片，卷积核的大小和权值不需要提前设计，我们只需要给定一个初始值，卷积神经网络通过梯度下降算法可以进行不断优化从而得到该卷积核。卷积计算时，卷积核在输入数据上根据滑窗的大小左右上下移动依次和输入数据进行卷机，即卷积核和输入数据对应的元素相乘相加计算得到一个新的像素值。

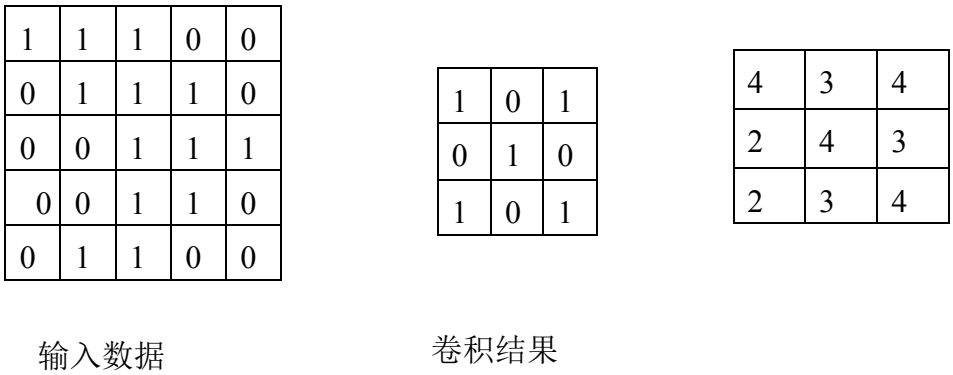


图 4-3 卷积运算

Figure 4-3 Convolution operation

以上便为卷积神经网络对图像进行特征提取的过程，我们可以通过构建更大的网络，或者改变每一层的参数等方法提高提取的特征的有效性，使其能够更准确的表示该图像。本文需要提取的特征有两种，一种是图像特征，一种是音频特征。

4.2.1 镜头图像特征

本文广告分类是在镜头级别上进行的，因此图像特征也是在镜头级别上进行提取的，由于一个镜头中包含着几百张图片，我们不能将所有的图片都进行提取特征，这样提取太耗费时间，而且有些相邻图片几乎是相同的，因此我们只需要将一个镜头中有代表性的几帧进行特征提取即可，为了保证数据的一致性，我们使用等间隔抽样的方法，从每个镜头中抽取 2、4、8、16、32 帧，并在同一 SVM 分类器上进行实验，从分类结果的准确度可以看出，抽取 16 帧的结果最好，抽取 4 帧的稍微差一些，但是抽取 16 帧的耗时几乎是抽取 4 帧的两倍，因此本文最终选择了每个镜头抽选 4 帧代表一个镜头特征。

由于本文的广告数据只有 5000 多个，数据量不足以训练一个可靠的特征提取器，因此，本文使用已经训练好的卷积神经网络对图片提取特征。

随着深度学习的不断发展，基于卷积神经网络的网络架构也在不断的发展，LeNet 是 1998 年 Yann LeCun 教授提出的卷积神经网络，也是第一次有人对 CNN

网络结构进行演化，ALexNet 被应用到手写体识别的任务，并取得了很好的识别率。ALexNet 的出现掀起了卷积神经网络的发展高潮，后来陆续出现了 ZFNet、VGG、GoogleNet 等网络架构，GoogleNet 的核心网络架构是 Inception 模块，Inception 网络对卷积神经网络的发展做出了重要的贡献，在 Inception 网络出现之前，大部分卷积神经网络的网络架构都是通过堆叠卷基层来获得更好的性能，但是并不是所有的网络层数越多越好，有时候层数越多反而会出现过拟合的现象，导致最终结果变得更差，Inception 网络的设计思想是由深度转为广度，设计了一种局部拓扑结构的网络，对输入的数据并行执行卷机运算的任务，然后将所有的计算结果进行拼接，得到特征图，Inception 网络相比之前的网络在速度和准确率上均得到了提升。因此本文使用了经过大量数据集训练好的 Inception-V3 网络，Inception-V3 网络是在 GoogLeNet 的核心网络架构 Inception 模块上进行的改进，把 googlenet 里的 7*7 的卷积拆分成了 1*7 和 7*1 的两层进行串联，这样不仅可以提高计算速度，而且增加了网络的非线性，降低过拟合的概率。Inception-V3 模块基本构成包括卷积层、池化层、全连接层以及 Inception 层，这也是 Inception-V3 模块的一个重大改变。表 4-1 是 Inception-V3 模型的网络结构框架，图 4-4 中 a)、b)、c)三张图分别表示 Inception-V3 模型。

表 4-1 Inception-V3 的网络结构

Table 4-1 Architecture of Inception-V3

type	patch size/stride or remarks	input size
conv	3×3/2	299×299×3
conv	3×3/1	149×149×32
conv padded	3×3/1	147×147×32
pool	3×3/2	147×147×64
conv	3×3/1	73×74×64
conv	3×3/2	71×71×80
conv	3×3/1	35×35×192
Inception	图 4-4 a)	35×35×288
Inception	图 4-4 b)	17×17×768
Inception	图 4-4 c)	8×8×1280
pool	8 × 8	8 × 8 × 2048
linear	logits	1 × 1 × 2048
softmax	classifier	1 × 1 × 1000

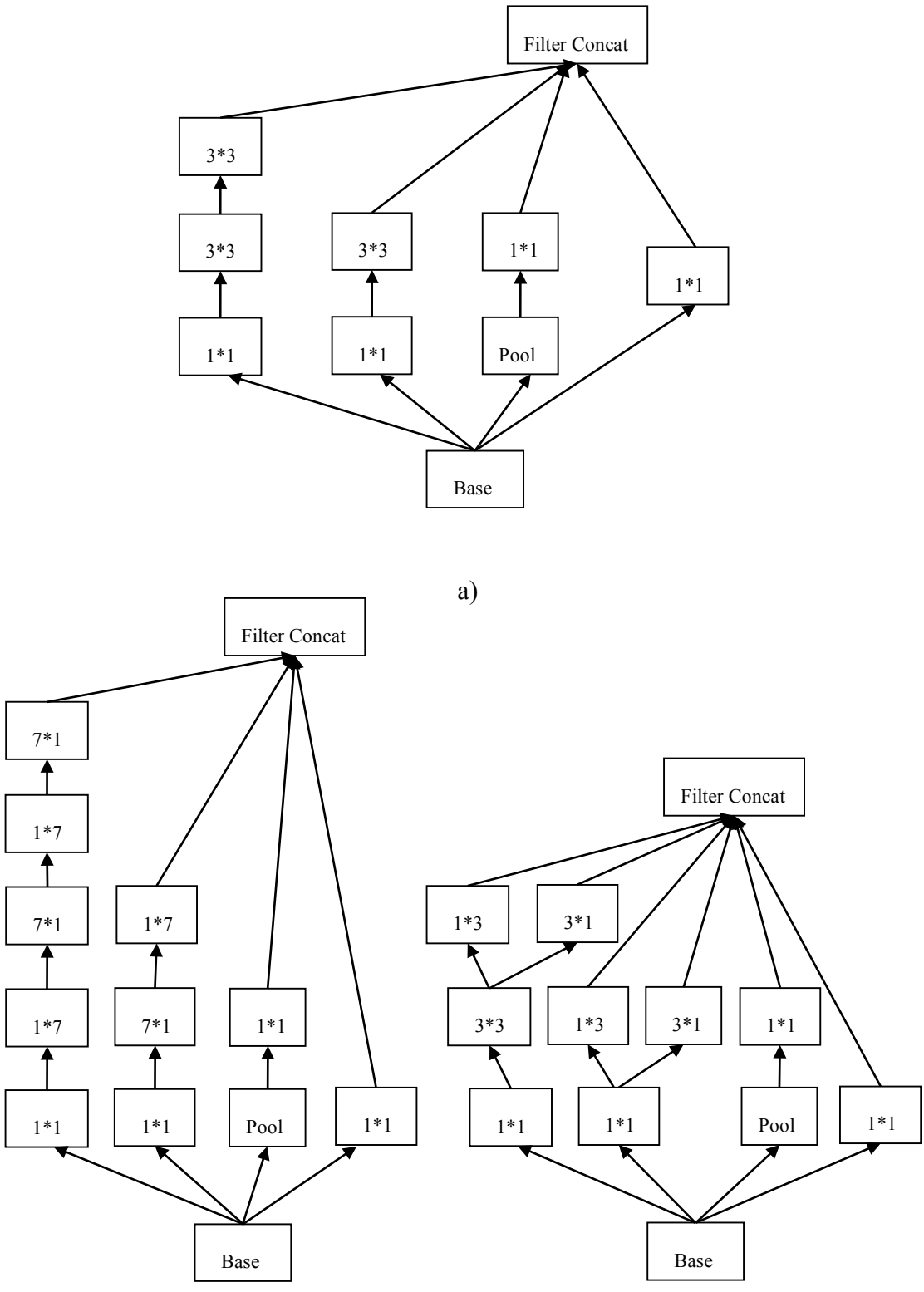


图 4-4 Inception-V3 中的 Inception 层

Figure 4-4 Inception layer of Inception-V3

经过 Inception-V3 网络提取的所有镜头的特征均在一个数组下，该数组内包

含了 11261 个列表，即 11261 个镜头，每个列表下是 4 个(1,2048)的向量，即一个镜头下我们抽取了 4 张图片，每一个向量代表的就是利用 Inception-V3 网络对一张图片提取的特征向量。

4.2.3 镜头音频特征

声音是一个人独有的特性，每个人的声音各不相同，有的人声音尖锐洪亮，有的人声音深沉浑厚，不同场景下的背景声音也不同，比如马场会有骑马、马叫声，洗澡会有流水声等，因此通过声音可以识别一个人，通过声音我们可以识别一个场景。声音特征就是提取音频信号中具有辨识性的成分，能够将不同人或者背景区分，常见的声音特征有过零率，即信号过零点的次数；短时能量，能够反映信号在不同时刻的强弱程；短时平均幅度差，音频也是具有周期性的，短时平均幅度差可以发现一段音频的周期特性。由于本文研究的中插广告中的演员是剧中的人物，广告的内容也和剧情有一定的关联。仅仅依赖上述传统的音频特征不能进行区分，我们同样希望能够挖掘广告与剧情在语音上深层次的差异，因此本文将音频转换为了二维色谱图。并通过 Inception-V3 网络对色谱图提取特征，由于一个镜头只有一段音频，所以经过 Inception-V3 网络提取的特征数组里包含了 11261 个列表，每个列表下是 1 个(1,2048)的向量，每一个向量代表的就是利用 Inception-V3 网络对一个镜头音频提取的特征向量。

4.3 特征融合

为了充分利用不同维度特征的信息，本文将图像特征和音频特征进行了融合。利用 Inception-V3 网络得到的特征是一个 2048 维的向量，由于 Inception-V3 模型是对 1000 种类别进行分类，高维度的向量能够表示更多的信息，但是本文是一个二分类任务，即分类为广告和非广告两类，过高维度的特征表示在训练的时候可能会出现过拟合的现象，而且维度过高在数据训练过程也是非常耗时，因此在进行特征融合之前需要先通过 PCA 对特征向量进行降维。但是我们并不知道降到多少维才是最合适的，维度太低会损失一些信息，维度太高还是会产生过拟合的状态。PCA 是通过参数 p 进行调节降维的比例，原理如公式 4-1 所示：

$$p=\frac{\sum_{j=1}^{k}\lambda_j}{\sum_{j=1}^{n}\lambda_j} \tag{4-1}$$

其中 λ_1，λ_2，…，λ_n 是特征矩阵的特征值从大到小进行排列，k 表示降维数。

p 值代表在不同降维数 k 下的数据能保留的信息，为了确定参数 p 的值，针对同一分类器，我们调节 p 值，观察在不同的 p 值下分类器的分类准确率，选择分类最好时的 P 值，我们选取了支持向量机（SVM）模型进行训练，得到不同 p 值下的分类准确率，如图所示，纵轴是分类准确率，横轴是参数 p 的值。由图 4-5 可知当 p 值等于 0.6 时，支持向量机的分类准确率最高，在 83%左右，此时根据公式 4-1 中 p=0.6 可得到 k 值为 98，因此我们需要将特征向量降成 98 维的向量。

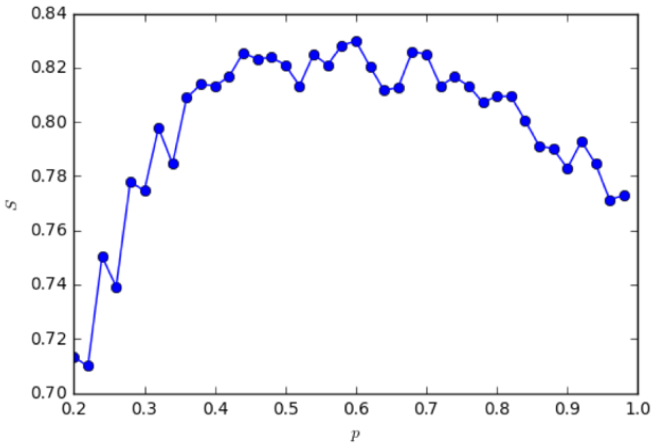


图 4-5 降维后的分类结果

Figure 4-5 Result of classification after dimensionality reduction

因此，经过降维后的图像特征数组内包含了 11261 个列表，每个列表下是 4 个（1，98）的向量，经过降维后的音频特征数组内包含了 11261 个列表，每个列表下是 1 个（1，98）的向量。

有时候音频能够获得一些图像不能获取的信息，比如水流声，声音可以准确的判断这是水，但是图片不一定能捕捉到该信息，因此能够让音频特征和图像特征能够在信息上进行互补，最大程度上利用有效信息进行分类，本文将融合后的音频特征和图像特征输入分类器中，融合方法有两种，一种是并联两个特征，一种是将两种特征直接进行拼接的方法。

4.3.1 特征并联

本文中一个镜头选取了 4 张图片，即一个镜头有 4 个图像特征向量，而音频镜头特征向量只有一个，为了保证两个镜头特征向量的独立性，我们直接将对应镜头的音频特征向量直接组合到图像特征向量的列表中，如图 4-6 所示，组合后的特征数组内包含了 11261 个列表，每个列表下是 5 个（1，98）的向量，包括 4 个图像向量，一个音频向量。

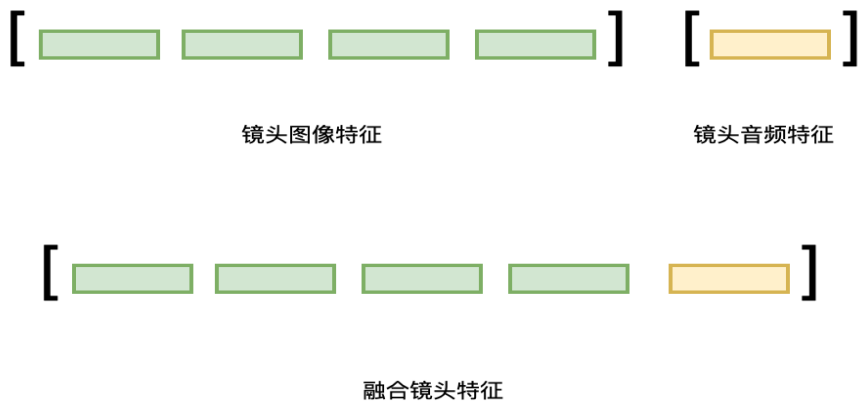


图 4-6 特征并联

Figure 4-6 Feature connection in parallel

4.3.2 特征拼接

直接拼接是指将两个镜头特征直接进行相连，由于本文中一个镜头选取了 4 张图片，即一个图像镜头的特征维度是（4，98），而音频镜头特征维度是（1，98），直接拼接有两种方式，一种是一个镜头的音频特征同时拼接到 4 张图片向量的后面，如图 4-7 （a）所示，即拼接后的特征数组内包含了 11261 个列表，每个列表下是 4 个（1，196）的向量。另外一种是先对图像镜头特征进行取均值，即求取四张图片向量的均值，再将镜头的音频特征直接拼接到求去均值后的图像向量的后面。如图 4-7（b）所示，组合后的特征数组内包含了 11261 个列表，每个列表下是 1 个（1，196）的向量。

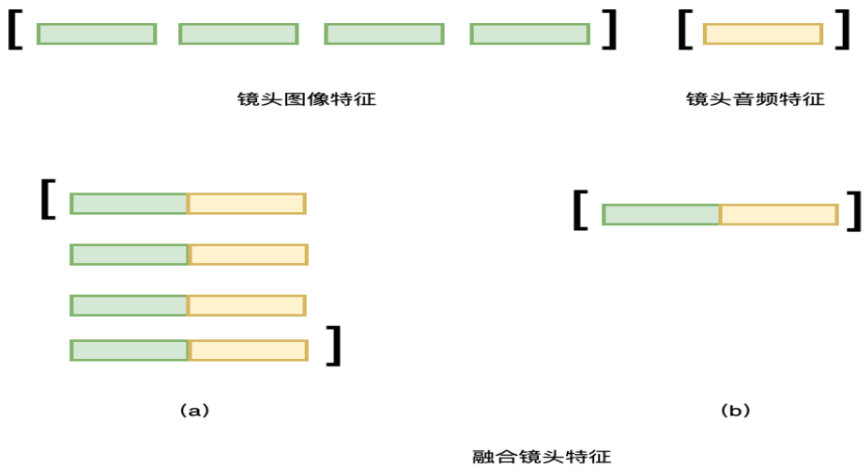


图 4-7 特征拼接

Figure 4-7 Feature concat

4.4 训练镜头分类器

为了使分类器输出最好的分类结果，需要对分类器进行训练，而分类器模型就是分类器算法经过对样本进行训练达到最好效果的时候保存下来的模型。大部分分类器是可以直接通过开源库直接进行调用，但是针对不同的数据集，分类器的参数设置不同，而训练的过程就是调节参数的过程，本文使用了机器学习的算法和深度学习模型组合的方法对中插广告数据集进行训练，最终根据分类准确率选取最优模型对测试样本执行分类预测。

4.4.1 机器学习

本文使用了比较高效的四种分类器对该数据集进行训练，包括随机森林（RF）、提度提升树（GBDT）、极端梯度提升（Xgboost）、支持向量机（SVM）四种分类器，机器学习的 Scikit-Learn 库提供了可以直接调用的调参接口，因此本文直接调用这四种直接分类器对数据集进行训练。

对于支持向量机算法，有两个重要的参数可调节，一个是松弛变量的参数 C，一个是单个样本对整个分类超平面的影响系数γ。对于随机森林算法有三个参数调节，一个是随机森林中树的数量 n_estimators，一个是使用处理器的数量 n_jobs 。对于梯度提升树算法有 n_estimators、learning_rate、subsample、loss 函数等参数进行调节。对于 Xgboost 有 min_ samples_split、min_ samples_leaf、max_ depth、learning_ rate 等参数可以调节。

本文使用总样本的 70%作为训练集，10%作为验证集，20%作为测试集，通过网格搜索的方法对以上四种算法进行调节参数，网格搜索是指遍历所有设置的候选的参数，根据分类结果的准确率来选择参数，并将最优结果的模型进行保存。最终根据试验结果准确率如下表 4-2 所示，由表可知，机器学习方法中 SVM 算法的分类结果最高，通过音频和视频特征结合的方法可以达到 86%的分类准确率。

表 4-2 机器学习算法试验结果
Table 4-2 Experimental results of machine learning algorithm

机器学习	图像特征	音频特征	特征融合
SVM	83%	76%	86%
RF	79%	68%	81%
Xgboost	77%	72%	81%
GBDT	83%	75%	84%

4.4.2 深度学习

本文使用深度学习来获得更高维度的特征。深度学习通过神经网络来模仿人的大脑工作机制，基于更深层次的特征进行结果输出，深度学习网络需要更多的数据支撑，通过大量数据的深度训练，以此来获取在图片上更深度更有区分性的特征表示。

本文使用的基于深度学习的镜头分类模型是多个深度网络搭建而成。由于视频是一个动态过程，前后帧之间存在着一定的关系，因此本文利用 LSTM 网络获取前后帧之间的关联信息，提取时序关系。本文中使用了两个维度的特征，一种是图像特征，一种是音频特征，但是我们并不知道针对某一镜头，是根据图像特征还是音频特征还是二者的组合特征才能够正确的得到分类结果，因此本文使用了 Attention 网络获得图像特征向量和音频特征向量的权重，进而得到在不同镜头中图像和音频对分类结果的贡献比重，并将注意力集中在那些对当前任务更重要的特征向量上，本文中 Attention 有两种，一种是对每个向量取一个权重我们称为向量权重，另一个是对向量中的的每个元素均取一个权重我们称为元素权重。Attention 存在的位置也有两种，一种是对输入向量进行 Attention，叫做前向 Attention，即原始输入向量先输入到 Attention 网络，再输入 LSTM 网络，另一种是对 LSTM 的输出进行 Attention，叫做后向 Attention，即原始向量先输入到 LSTM 中，LSTM 的输出结果再输入到 Attention 网络中，最后输出待分类的镜头是广告还是非广告的分类结果，LSTM 与 Attention 结合的结构原理图如图 4-8 所示。

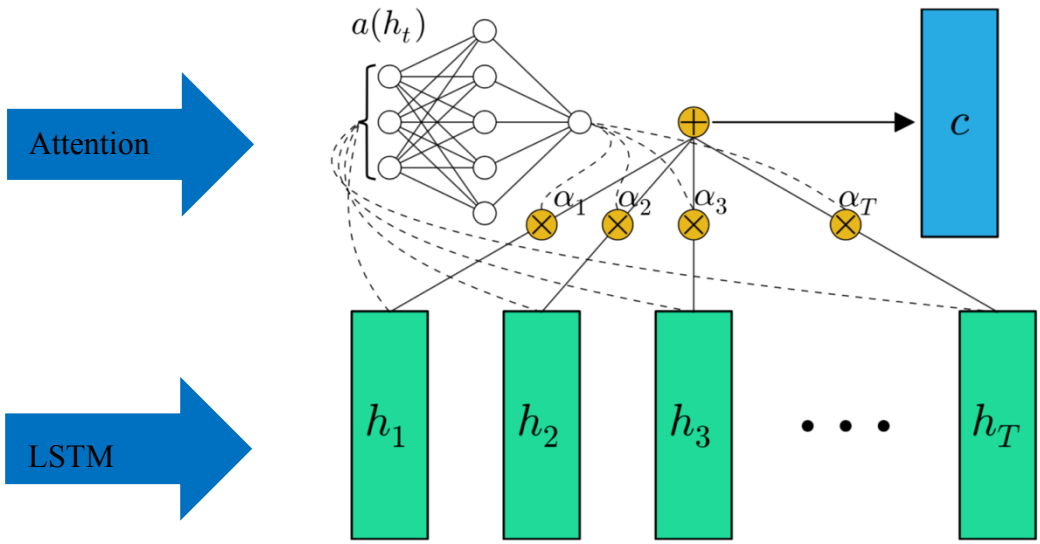


图 4-8 结构原理图

Figure 4-8 Structural schematic diagram

在编码中的每一个时间步长中，LSTM 单元都会输出一个隐状态向量 h，在解码器部分每一个时间步长中的 LSTM 单元都会接受一个输入，这个输入是由编码器部分所有的隐状态相加组成的，由于编码器中的每一个隐状态对解码中当前的时间步长的输出结果的影响并不一定相同，所以对于每一个隐状态都相应的乘上一个权重，再对所有的乘上权重的隐状态进行加和，这就是 Attention 机制在 LSTM 网络中的应用。其中权重又称为 Attension scores，Attension scores 的计算可以利用多种形式，例如利用线性计算的方法，计算解码器在当前的时间步上的隐状态与编码器所有的瘾状态的相似性再进行归一化得到 Attension scores；也可以利用非线性计算的方法，比如将编码器的所有的隐状态送入前向神经网络中进行学习，得到 Attension scores。本文主要调节的参数有 LSTM 的层数、每一层的单元数、激活函数等，同时对上述元素权重和向量权重以及 Attention 的位置变化以及不同输入特征均进行了实验，试验结果如下表 4-3 所示：

直接拼接的方式有两种，这里只显示两种方法中最好的结果。由实验结果可知 Attention 的位置放在 LSTM 后，对向量进行 Attention，拼接方式选择特征并联的方式结果最好，准确率可以达到 88%，高于机器学习的方法。

表 4-3 深度学习算法实验结果

Table 4-3 Experimental results of deep learning algorithm

Attention 位置	权重	特征	精准度
前	元素权重	图像	0.838
		特征拼接	0.853
		特征并联	0.848
	向量权重	图像	0.843
		特征拼接	0.850
		特征并联	0.863
后	元素权重	图像	0.846
		特征拼接	0.874
		特征并联	0.857
	向量权重	图像	0.853
		特征拼接	0.859
		特征并联	0.885

4.5 本章小结

本章主要介绍了卷积神经网络提取特征的过程，由于数据量的限制，本文采用了已经通过大量数据训练好的 Inception-V3 网络提取特征，而音频特征是通过转化为色谱图后再通过神经网络提取特征，接着是对两个维度的特征进行融合，本文使用了两种方法，一种是特征组合，一种是特征拼接的方法，通过特征融合的方法能够让两个维度的特征信息进行互补，提高分类准确率。最后是训练分类器的方法以及分类准确率，本文采用了机器学习中几种典型常用的分类器进行实验，通过调节参数获得最优模型，除此之外还通过 LSTM 和 Attention 这两种深度学习模型组合对数据集进行分类，最终深度学习的方法实验准确度要稍微高于机器学习的方法。由于本文数据集有限，相信增加数据集以后深度学习的方法可以明显的优于机器学习算法。

5 广告内容识别

本章主要对系统的最后一部分广告内容识别进行详细的介绍,由于本文中的视频广告的 Logo 不再显著，所以本文通过观察找到了一种新的识别特征，即文字特征，由于有些视频广告中没有文字出现，仅仅是通过视觉差异和语言介绍结合的方法进行广告宣传，因此仅通过文字识别不能达到很好的识别效果，所以本文采用了两种方法的综合进行广告内容识别，一种是利用广告文字进行识别，一种是基于音频匹配的方法进行识别，最终两种方法结合后的识别准确率可以达到 98%。

5.1 问题描述

广告是以一种传播的方式将商品进行宣传，一个广告片段并不是所有的时间都在介绍该商品，广告商会由用户的兴趣点或者需求分析出发，慢慢引出广告的主体内容，即广告中的商品，本文中的广告商品既有传统广告中的牛奶、洗发水、巧克力、护肤品等商品，与此同时，本文中的广告还增加了手游、APP、理财产品等广告，在传统广告的形式上又增加了新的广告内容，而这些新的广告产品展现形式和传统广告又有很大的不同，传统广告商品内容都有比较明显的商品商标，也就是 Logo，利用图片的特征匹配便可以进行识别，而本文中的手游、APP、理财商品这些广告的 Logo 已经变得非常不明显，因此利用 Logo 匹配的方法已经不能完全适用，基于对广告形式的观察，发现了一种新的可以识别广告内容的特征，即文字特征，文字内容相比 Logo 更容易识别，所以我们用文字识别代替了传统的图像特征匹配的方法。本文中还存在一种广告形式是通过语音介绍的方法，即广告全程屏幕上并没有出现于产品有关的任何文字信息和广告 Logo，这时我们需要用音频匹配的方法进行识别广告内容，因此本文综合了文字识别和音频匹配两种方法进行广告内容识别。

5.2 利用文字进行内容识别

文字是商品的一种表现形式，由于本文大部分广告商品是对手游、APP 等商品进行广告宣传，如图 5-1 所示，该商品在表现形式上文字内容比商品 Logo 更容易识别，甚至有的广告商品已经将文字内容作为 Logo，因此本文采用了文字识别的方法代替传统的图像特征匹配的识别方法，本节主要介绍如何识别文字内容。

图 5-1 广告镜头

Figure 5-1 Advertising shots

5.2.1 文字数据库

广告商为了展示商品的多样性，一张图片中不仅仅会存在商品名称，还会稍带着一些功能性介绍文字，比如介绍一款理财产品的时候，不仅仅会出现该产品的名字，还会附有“投入低、回报高”等吸引用户的文字，但是我们只需要知道该商品的名称，并不需要知道这些商品的其它介绍，这就需要我们对识别出的文字与数据库的文字进行匹配，所以本文数据库是由 103 个广告名称构成的一个列表，包括“PPmoney”、“乳酸菌”、“麦吉丽素颜三部曲”等广告商品，本文共有 142 个广告，其中有 39 个广告名称一样，但是广告的宣传方式并不相同，所以本文将这两种广告视为不同的数据样本。

5.2.2 文字区域检测

文字识别主要分为两步，第一步就是文字区域检测，将带有文字的区域检测出来，本文使用了经典的 Faster R-CNN 算法，该算法流程如下：输入一张完整图片-->CNN 得到 feature map-->卷积特征输入到 RPN，得到候选框的特征信息-->分类+回归，Faster R-CNN 中利用 ResNet 卷积神经网络架构进行特征提取，本文使用旷视科技提出的 ShuffleNet 网络架构代替了 ResNet 网络，在运算效率上得到了大大的提升。

传统网络中，逐点卷积（1×1 卷积，逐一相乘相加）的方法会使通道之间充满约束，而且会浪费大量的计算时间。解决这个问题的方法是采用组卷积(group convolutions)的方法也就是通道稀疏连接。虽然这种方法能够使每个卷积仅在相应

的通道上计算，但是这样会阻碍通道组之间的信息流通，降低了信息表示能力，如图 5-2(a)所示。如果组卷积能够得到对不同组数据进行交换，即 5-2 (b)所示，那么输入和输出的信息就可以实现互通关联。这也是 ShuffleNet 网络架构的核心思想，具体来说就是对于上一层输出的通道，做一个 Shuffle 操作，如图 5-2(c)所示，再分成几个组，输入到下一层。

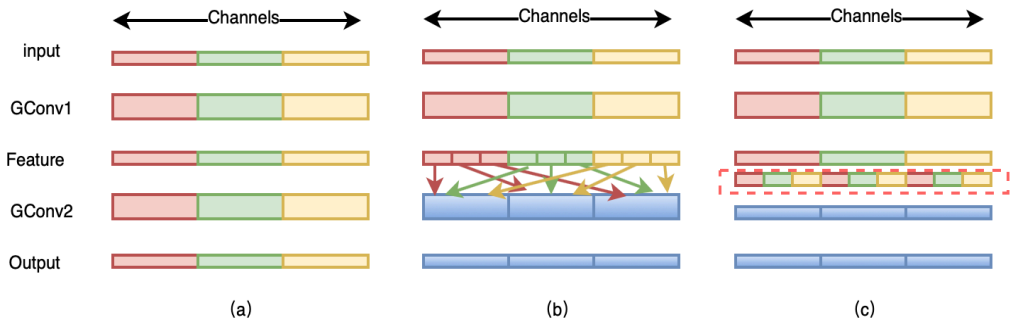


图 5-2 ShuffleNet 原理

Figure 5-2 The ShuffleNet network principle

混洗过程如图 5-3 所示，输入图片有（g，n）组通道，对其重新组合成二维矩阵（g，n），将矩阵进行转置，最后再摊平分组。

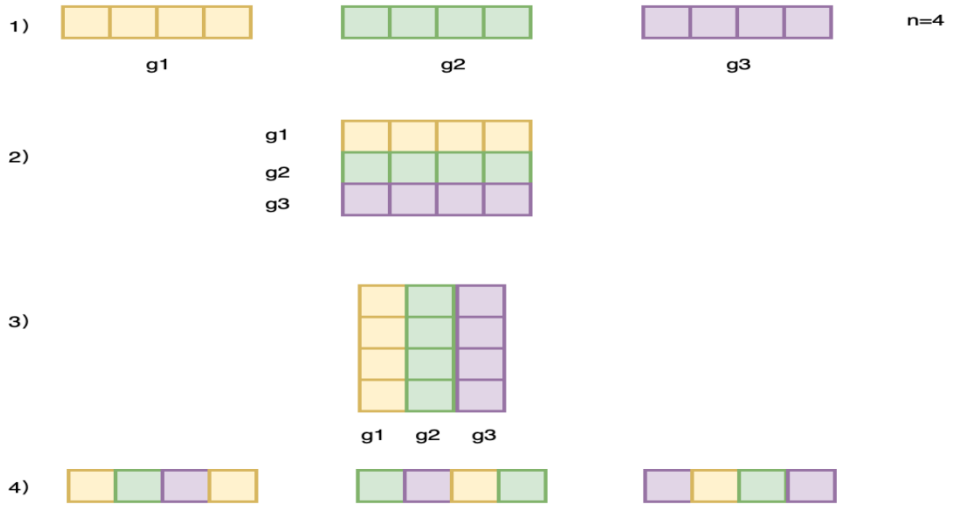


图 5-3 ShuffleNet 过程

Figure 5-3 The process of ShuffleNet

Faster R-CNN 当下已经是非常成熟的一种文本区域检测技术，本文是在已有的代码[55]基础上通过改变了提取特征的网络架构和一些参数（lou 值、loss 函数等）达到本文的最优识别结果。本文通过“print pred_boxes”语句将分类结果为文字的坐标框进行打印，如图 5-4 所示，每一行表示一个文字区域目标框的坐标。

```
[1198.58422852, 1014.32049561, 1291.8581543, 1123.05639648]
Y
[675.29943848, 634.83068848, 766.93762207, 724.48535156]
Y
[1463.50634766, 131.7862854, 1548.50048828, 233.230499277]
Y
[1021.40447998, 367.55706787, 1138.07788086, 479.88537598]
Y
```

图 5-4 候选文本框坐标

Figure 5-4 Coordinates of the region proposals

5.2.3 区域内容识别

区域内容识别是文字识别的第二步，通过上一节我们已经得到了一张图片可能包含文字的坐标，通过坐标我们使用了 opencv 的方法截取图片并保存，由于 CRNN 实现的环境比较简单，而且识别效果也很好，所以本文使用了 CRNN 算法，通过 CTC 损失函数来调整 LSTM 的参数 w，使得输出概率矩阵最大，最终输出得到的文字序列。

5.3 利用音频特征进行内容识别

广告信息除了以一种文字形式播放外，还存在另一种方式，即利用视觉差异和语言介绍结合的方法，如图所示，该广告主要介绍一款面膜，但是整个广告并没有出现任何文字，该广告是通过语音介绍，加上视觉上的变化来把产品信息传递给观看者，对于这种广告，文字识别的方法便不再适用，因此本文在文字识别方法上还增加了音频匹配的方法进行广告识别。

图 5-5 无文本广告镜头

Figure 5-5 Advertising shot without text

5.3.1 音频数据库

一段声音可以代表一个故事、一首音乐，而一个镜头音频则可以表示是一个故事中的某一情节、一首音乐的某几个旋律，我们常常根据听到某一情节或者某几个旋律便可以猜到故事的名字和歌曲的名字，本文中的音频特征匹配也是利用这样的原理，因此本文音频数据库是抽取了完整广告的音频，该音频数据库共包含 142 段完整广告的音频，每段音频以广告名称命名。

5.3.2 色度特征匹配

对于音频信号，色度可以有效的表示其特征。色度特征包含了色度向量和色度图谱，色度向量是指一帧图像的 12 个音级中的能量，来源于不同 8 度的同一音级的能量的累加，因此色度向量是一个含有这 12 个元素的特征向量，而色谱图[56]就是由若干帧图像的色度向量有序组成。

如图 5-6 所示为对音频提取色度特征并生成色谱图的过程。首先是把音频文件通过傅里叶变换把音频从时域变化到频域，对数据做一些降噪等预处理，再微调声音的频率，把声音调到标准频率上，接着把时间转换为帧，并记录每一帧内每一个音高的能量，构成音高图谱；接着在音高图谱上把同一时刻、同一音级下的不同 8 度能量分别叠加到色度向量的对应的同一音级的元素上，这样便构成了色度图谱。

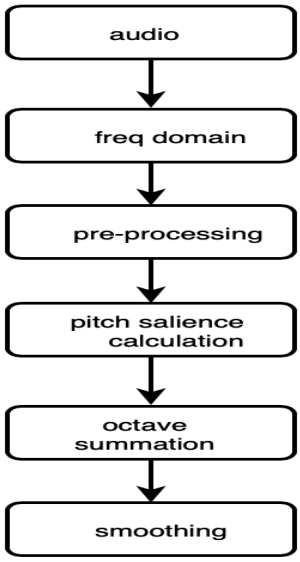


图 5-6 提取色谱图流程图

Figure 5-6 Flow chart of extracting chromatogram

音频的色谱图是由一个（12，n）的二维矩阵构成，其中 n 与音频的长度相关，

其表达形式和图像相同，因此本文中将图像模板匹配的方法应用于色谱图的匹配。如图 5-7 所示，左边为一个完整广告的色谱图，右边为该广告下的一个镜头的色谱图，右图是左图的一部分，进行匹配时，镜头的色度特征会在完整广告的色度向量上从左往右进行平移，计算每一次移动的位置的匹配相似度，假设该完整广告的色度特征矩阵为（12，n），该广告下的一个镜头的色度特征矩阵为（12，m），每次移动一帧，因此会得到 n-m+1 个数，代表了该镜头与完整广告在不同位置的特征匹配相似度，记录这 n-m+1 个数中的最大值，即为该镜头属于此广告的匹配得分，最终我们可以得到该镜头对所有广告的匹配得分，并输出得分最高的广告名称。如果一个镜头属于该广告，通过匹配的后的值应该为 1，但是由于计算过程中难免出现一些误差，因此输出结果基本上不会出现匹配度为 1 的情况，本文中正确匹配的得分一般都在 0.90 以上。

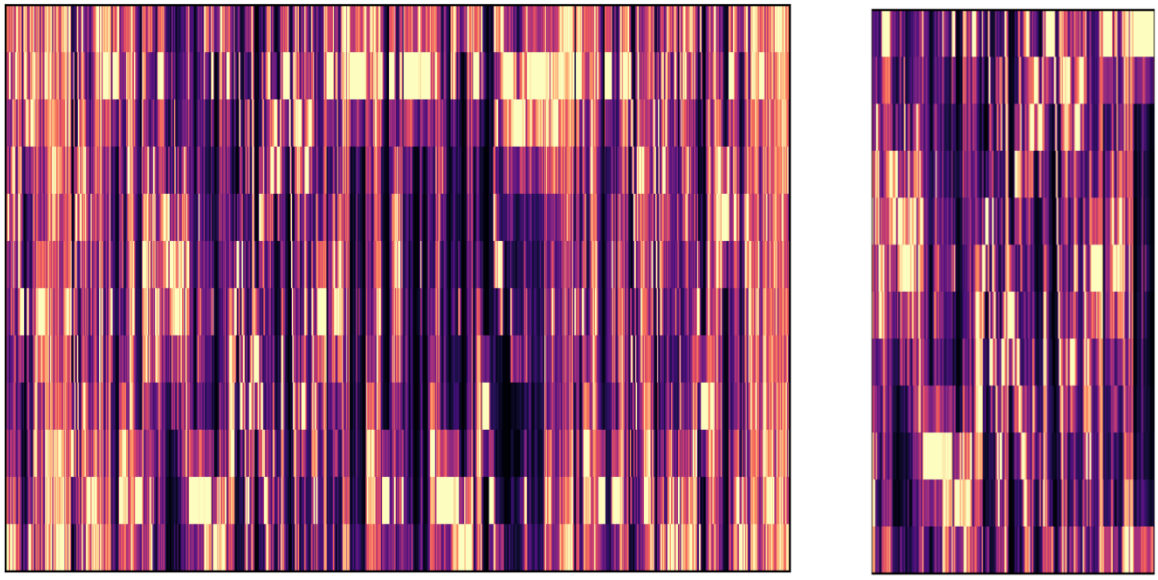


图 5-7 色度图

Figure 5-7 Chromaticity diagram

5.4 综合文字与音频方法进行文字识别

为了避免分类错误对本章广告识别造成影响，本文文字识别的镜头数据集都是分类正确的广告镜头。针对本文数据集中广告的多样性，我们采取了文字识别与音频识别综合的方法进行广告识别，首先进行文字识别，由于每个镜头下有几百张图片，如果随机选择一张图片，该图片带有广告内容的概率是非常小的，因此我们进行等间隔抽样，每个镜头下抽取 10 张图片进行文字提取，依次将文字数据库中 103 个广告名称与该镜头下提取的文字进行匹配，如果存在广告名称则输出，不存

在则跳过该镜头。最后输出的结果是广告内容和镜头名称（即哪个电视剧第几集的镜头编号以及广告内容），并将输出的广告内容与镜头名称中的广告标签进行匹配便可得到文字识别准确率。由于一个广告内容可能会多次出现或者被切分成两个镜头，如果存在同一输出结果属于同一电视剧的同一剧集里面的情况，我们只保留其中一个，将其它的结果直接去除。最终广告的识别准确率为 70%。

对于音频数据，我们没有必要将得到的所有的广告镜头音频与数据库的广告音频进行匹配，只需要一集中的一个镜头即可，但是中插广告间的开场和结尾十分相似，为了避免我们随机抽选到广告的开始和结尾镜头，我们从每一集中随机抽选了 3 个广告镜头，依次与音频数据库的所有广告镜头进行匹配，并输出 3 个镜头匹配得到的广告音频的名称，如果 3 个镜头中有两个及以上镜头的输出结果相同，则认为是该结果，如果三个各不相同，则输出匹配分数最大的结果，最后将输出结果与该镜头中的广告标签进行匹配，得到音频匹配准确率。

本文的音频匹配是在文字识别失败的数据集上进行的，这样可以减少重复识别的工作量，最终在文字识别 70%的准确率上，我们将错误的数据通过音频进行识别，最终识别准确率为 98%，如下表 5-1 所示。

表 5-1 文字识别实验结果

Table 5-1 Experimental results of character recognition

特征	数据	识别准确率
音频	一个镜头	78%
	三个镜头	87%
文字	10 张图片/镜头	70%
音频+文字	三个镜头、10 张图片/镜头	98%

5.5 本章小结

本章主要讲述的是广告内容识别用到的两种方法，由于本文中中插广告的广告商品在传统的商品广告上增加了手游、APP、理财产品等，表现形式也和以往大有不同，文字取代了传统广告的 Logo 成为了更显著的商标，因此本文首先是采用了文字识别的方法进行识别，但是本文也存在整个广告过程中均未出现文字的情况，仅仅是通过声音描述和视频显示来说明广告商品的有效性，此时文字识别的方法不再适用，因此我们在文字识别不能适用的数据集上进行音频匹配的方法输出广告内容，通过两种方法综合后，我们的文字识别准确率可以达到 98%。

6 总结及展望

本章主要是对本文的系统构成以及每一部分的功能进行简要的总结，列举了本文做出的贡献性工作，最后描述了本文工作的一些不足和未来工作的一个展望。

6.1 本文工作总结

本节主要对本文的工作成果进行总结说明。本文针对近两年爆发的中插广告设计了一个识别系统。不同于传统广告，中插广告中的演员是剧中的人物，广告的内容也和剧情有一定的关联，这使得识别中插广告存在一定的挑战性，由于每集视频时长 50 分钟左右，而广告时长只有 90s，如果对整个视频进行处理识别，这样不仅会使结果存在误差，使结果偏向数据量大的一类，而且也会大大增加工作量，因此本文是在镜头级别上进行识别，本系统分为三部分：镜头切分、镜头分类、广告内容识别，下面将会针对本文数据集的特点讲述每一部分的挑战和本文的贡献。

(1) 镜头切分。传统广告与剧情的边界十分明显，因此仅仅靠突变镜头的检测方法就可以将广告与剧情切分开，而本文的中插广告的出现，模糊了广告与剧情的界限，二者的连接已经由突变转化为渐变，渐变镜头又分为镜头间渐变和镜头内部渐变，为了保证剧情主题的完整性，本文只对广告与剧情的渐变镜头进行切分，并不希望将剧情中的渐变镜头也进行切分，基于对镜头间渐变镜头的观察，我们发现由剧情进行广告的时候，镜头颜色变化是先变暗后变亮，即由上一个镜头的黑色淡出，下一个镜头从黑色淡进，本文称这样的渐变为黑镜头，而镜头内部的渐变并不会存在这样的黑镜头，所以本文切分镜头时，首先对突变镜头进行切分，然后再对渐变镜头进行切分，渐变的切分方法主要是根据颜色变化趋势由亮变暗再变亮，转折点存在黑帧的方法进行切分，结果证明该方法能很好的保留镜头间的渐变，而将剧情与广告进行切分。

(2) 镜头分类。传统广告无论是在内容、演员还是背景、声音上都与剧情存在很大的差异，但是中插广告不仅演员是剧中的角色，连内容都是剧情的番外篇，仅仅简单的从背景、声音上很难进行区分，因此需要提取更深层次的特征进行区分。本文使用了深度学习中多模型组合的方法进行分类，由于视频是一个动态过程，前后帧之间存在着一定的关系，因此本文利用LSTM网络获取前后帧之间的关联信息，提取时序关系，并通过 Attention 网络获得图像特征向量和音频特征向量的权重，进而得到在不同镜头中图像和音频对分类结果的贡献比重，并将注意力集中在那些对当前任务更重要的特征向量上，通过镜头分类模型输出待分类的镜头是广告镜头还是非广告镜头。同时也从音频、视频等多个特征维度进行探索，提高分类准

确度。

(3) 广告内容识别。本文的中插广告的宣传方式不再是大大的商品 Logo，文字取代了传统广告的 Logo 成为了更显著的商品标签，但是有些广告中也会出现整个视频无文字的现象，只是通过声音特效和视频演示来展示商品，因此本文用文字识别与音频匹配相结合的方法对广告内容进行识别，为了避免重复计算，本文先使用文字识别算法进行识别，可以达到 70%的识别准确率，接着对未能识别的数据进行音频特征匹配，使用两种方法结合的策略，最终使得识别率达到了 98%。

6.2 未来工作展望

随着经济的发展，视频中广告已经成为社会生活中越来越重要的一部分，广告中包含着大量有价值的信息，识别视频中的广告对市场的发展存在很大意义。而中插广告的出现，改变了传统广告的内容设计和播放形式，而且在播出后呈爆发式增长，大有取代传统广告位置之势，成为市场监测的新挑战。因此本文设计了一种对中插广告进行识别的系统，但是由于本文的数据集有限，深度模型的训练分类结果相比传统的机器学习算法仅仅提升了 2%的准确度，因此，未来工作希望能够增加更多的样本并通过深度学习进行训练，获得更高的分类准确率。另一个工作是本文的广告内容识别必须是对已经离线缓存的视频进行识别，未来我们希望可以在线实时分析并输出此类广告内容。

参考文献

[1] 威思塔.新型广告类型有什么[EB/OL]. https://www.zhihu.com/question/286542738/answer/473397498

[2] 佚名. 中插广告市场白皮书[M]. 2018.

[3] 胡东升. 网剧中中插广告的探析[J]. 参花：下半月, 2018(4): 121-122.

[4] Dey N , Pal G , Rudrapaul D , et al. Video Shot Boundary Detection: A Review[M]// Emergin ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2. Springer International Publishing, 2015.

[5] Liang, L., Liu, Y ., Lu, H., Xue, X., Tan, Y .-P .: A Enhanced Shot Boundary Detection Using Video Text Information. IEEE Transactions on Consumer Electronics51(2), 580–588 (2005)

[6] Mengyue Li, Yuchun Guo, Yishuai Chen, et al. CNN-based Commercial Detection in TV Broadcasting [C]. ACM，2017.

[7] Wang H,Schmid C. Action Recognition with Improved Trajectories[C]//Computer Vision & Pattern Recognition. 2013.

[8] Peng X , Zou C , Qiao Y , et al. Action Recognition with Stacked Fisher Vectors[C]// European Conference on Computer Vision. Springer, Cham, 2014.

[9] Fernando B , Gavves E , Oramas M J , et al. Modeling video evolution for action recognition[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2015.

[10] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, pages 568 – 576, 2014.

[11] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell., 35(1):221 – 231, 2013.

[12] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In ICCV, pages 4489 – 4497, 2015.

[13] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko. Long-term recurrent convolutional networks for visual recognition and description. In CVPR, pages 2625 – 2634, 2015.

[14] J. Y. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In CVPR, pages 4694 – 4702, 2015.

[15] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In ECCV, pages 20 – 36, 2016.

[16] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. Untrimmed-Nets for weakly supervised action recognition and detection.In CVPR, pages 4325 – 4334, 2017.

[17] Long X , Gan C , Melo G D , et al. Attention Clusters: Purely Attention Based Local Feature Integration for Video Classification[J]. 2017.

[18] C. Gan, N.Wang, Y. Yang, D. Yeung, and A. G. Hauptmann. DevNet: A deep event network for multimedia event detection and evidence recounting. In CVPR, pages 2568 – 2577, 2015.

[19] Kelvin Xu，Jimmy Lei Ba , Ryan Kiros ,et al. Show, Attend and Tell: Neural Image Caption

Generation with Visual Attention[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2016.

[20] Hori C, Hori T, Lee T Y, et al. Multimodal Keyless Attention Fusion for Video Classification [C]// IEEE International Conference on Computer Vision. 2017.

[21] Chen J, Song X, Nie L, et al. Micro Tells Macro: Predicting the Popularity of Micro-Videos via a Transductive Model[C]// Acm on Multimedia Conference. 2016.

[22] Jing P , Su Y , Nie L , et al. Low-rank Multi-view Embedding Learning for Micro-video Popularity Prediction[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, PP(99):1-1.

[23] Guha P, Guha P, Guha P, et al. Commercial Block Detection in Broadcast News Videos[C]// Indian Conference on Computer Vision Graphics and Image Processing. ACM, 2014:63.

[24] Hua X S, Lu N L, Zhang H J. Robust learning-based TV commercial detection[C]// IEEE International Conference on Multimedia and Expo. IEEE, 2005:4 pp.

[25] DX.ChenandA.Yuille.Detectingandreadingtextinnaturalscenes.In Proc. of CVPR, 2004.

[26] Y. Zhong, K. Karu, and A. K. Jain. Locating text in complex color images. Pattern Recognition, 28(10):1523âA˘S¸–1535, 1995.

[27] K. I. Kim, K. Jung, and J. H. Kim. Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. IEEE Trans. PAMI, 25(12):1631–1639, 2003.

[28] . Gllavata, R. Ewerth, and B. Freisleben. Text detection in images based on unsupervised classification of high-frequency wavelet coefficients. In Proc. of ICPR, 2004.

[29] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In Proc. of CVPR, 2010.

[30] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu. Detecting texts of arbitrary orientations in natural images. In Proc. of CVPR, 2012.

[31] C. Yi and Y. Tian. Text string detection from natural scenes by structure-based partition and grouping. IEEE Trans. Image Process- ing, 20(9):2594–2605, 2011.

[32] W. Huang, Z. Lin, J. Yang, and J. Wang. Text localization in natural images using stroke feature transform and text covariance descriptors. In Proc. of ICCV, 2013.

[33] A. Jain and B. Yu. Automatic text location in images and video frames. Pattern Recognition, 31(12):2055âA˘S¸–2076, 1998.

[34] Y. Pan, X. Hou, and C. Liu. A hybrid approach to detect and lo- calize texts in natural scene images. IEEE Trans. Image Processing, 20(3):800–813, 2011.

[35] Y. Liu, S. Gotoand, and T. Ikenaga. A contour-based robust algorithm for text detection in color images. IEICE Transactions on Information and Systems, 89(3):1221–1230, 2006.

[36] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 580–587.

[37] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” in Proc. 13th Eur. Conf. Comput. Vis., 2014 pp. 346–361.

[38] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1440–1448.

[39] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. 2015.

[40] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, “Selective search for object recognition,” Int. J. Comput. Vis., vol. 104, no. 2, pp. 154–171, Sep. 2013.

[41] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. De- tecting text in natural image with connectionist text proposal network. In In Proceedings of European Conference on Computer Vision (ECCV), pages 56–72. Springer, 2016.

[42] Dao Wu, Rui Wang, Pengwen Dai, Yueying Zhang, and Xiaochun Cao. Deep strip-based network with cascade learning for scene text localization. In Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on, volume 1, pages 826– 831. IEEE, 2017.

[43] XiangyuZhu,YingyingJiang,ShuliYang,XiaobingWang,WeiLi, Pei Fu, Hua Wang, and Zhenbo Luo. Deep residual text detection network for scene text. In Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on. IEEE, 2017, volume 1, pages 807–812, 2017.

[44] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. EAST: An efficient and accurate scene text detector. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[45] M. Sawaki, H. Murase, and N. Hagita. Automatic acquisition of context-based images templates for degraded character recognition in scene images. In Proc. of ICPR, 2000.

[46] J. Zhou and D. Lopresti. Extracting text from www images. In Proc. of ICDAR, 1997.

[47] J. Zhou, D. P. Lopresti, and Z. Lei. Ocr for world wide web images. InProc. of SPIE, 1997.

[48] T. E. de Campos, B. R. Babu, and M. Varma. Character recognition in natural images. In Proc. of VISAPP, 2009.

[49] A. Mishra, K. Alahari, and C. V. Jawahar. Top-down and bottom-up cues for scene text recognition. In Proc. of CVPR, 2012.

[50] MaxJaderberg,KarenSimonyan,AndreaVedaldi,andAndrewZisserman.2014. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. In NIPS Deep Learning Workshop.

[51] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence, 39(11):2298–2304, 2017.

[52] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2231–2239, 2016.

[53] Alex Graves, Santiago Fernа́ndez, Faustino Gomez, and Ju ̈rgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, pages 369–376. ACM, 2006.

[54] 李梦月. 基于多维特征提取的视频内容识别系统设计与实现[D]. 2018.

[55] YunChen. https://github.com/chenyuntc/simple-faster-rcnn-pytorch, 2017.

[56] SantosRodríguez, Raúl, Ni Y , et al. Automatic chord estimation from audio: A review of the state of the art[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2014, 22(2):556-575

作者简历及攻读硕士学位期间取得的研究成果

张莹，女，1994年2月生。2013年9月至2019年7月就读于大连民族大学信息与通信工程学院通信工程专业。取得工学学士学位。2017年9月入学北京交通大学电子与通信工程专业，攻读工程硕士学位。

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名： 签字日期： 年 月 日

学位论文数据集

表 1.1： 数据集页

关键词*	密级*	中图分类号	UDC	论文资助
中插广告，镜头切分，镜头分类，广告内容识别	公开			国家自然科学基金 No.61572071，61271199，61301082
学位授予单位名称*		学位授予单位代码*	学位类别*	学位级别*
北京交通大学		10004	工学	硕士
论文题名*		并列题名		论文语种*
中插广告自动识别系统的设计与实现				中文
作者姓名*	张莹		学号*	17125082
培养单位名称*		培养单位代码*	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西直门外上园村 3 号	100044
学科专业*		研究方向*	学制*	学位授予年*
电子与通信工程		信息网络	2	2019
论文提交日期*	2019.06.03			
导师姓名*	郭宇春		职称*	教授
评阅人	答辩委员会主席*		答辩委员会成员	
	孙强		赵永祥、李纯喜、郑宏云、张立军	
电子版论文提交格式　文本( )　图像( ) 视频( ) 音频( ) 多媒体( ) 其他( ) 推荐格式：application/msword；application/pdf				
电子版论文出版(发布)者		电子版论文出版(发布)地		权限声明
论文总页数*	60 页			
共 33 项，其中带*为必填数据，为 21 项。				