基于视频特征和机器学习的视频分割和分类系统

清华大学 计算机系
宋佳铭 章彦恺

2014 年 11 月 30 日

1 研究目的

这次的研究是要解决在给出连续的视频但缺少节目单信息的情况下，对视频中的不同节目进行检出切分，并对节目的类型进行概括，并在原始视频无额外信息的情况下，检测出节目的不同部分。

形式化的，对于一系列的输入序列 x，我们需要给出一个函数 $f(x)$，使得对于正确的分类序列 $\hat{y}$，最小化：

$$L = L(f(x) - \hat{y})$$

从学术的角度来讲，本次研究的目的是希望结合之前的研究成果和学生的自身创新，提出一种新的，具有借鉴意义和继续研究价值的，集分割和分类于一体的视频事件检测和分类方法。

从实际应用的角度来讲，我们希望给出一个监测视频事件和视频事件分类技术在实际应用当中的一个新的应用场景，使得在学术领域的理论发展能够从一个新的角度来服务于现实生活。此外我们将开发一个可供用户直接操作的可视化图形界面，能够方便用户对实验的结果进行一个的评价。

2 研究背景

在当今的互联网上，由于网络可用带宽不断提升，视频能够为用户提供更好的使用体验，因而视频流量占网络总流量的比例正在不断地提高。但与不断增加的视频文件，技术的发展带来的用户自行上传视频的便捷相冲突的，是视频文件的详细信息缺失。例如对于用户录制上传的一天电视节目，往往就缺少了应有的节目单信息。另一方面，每个节目中又存在着观看者想要观看的部分，如武打镜头；或者是观看者不想看的部分，如片头片尾。对于用户自行上传的视频文件来说，这些视频信息的加入无疑将极大地提高观看者的观赏体验。

此外，过去的几年里，借助计算机理论科学领域中机器学习的蓬勃发展，许多原先被认为是不可能由计算机来实现的功能，现在都已经开发出现实的技术。视频事件的检测和不同粒度的事件分类这一工作也不例外地能够借助较为成熟的机器学习技术来实现。

在这样的应用和理论技术背景下，在分析了创新性、可行性和可推广性后，我们提出了本次研究的课题，一个基于视频特征和机器学习的视频分割和分类系统。

3 创新之处

我们认为，这项工作的创新之处在于以下几点：

1. 在视频分类中，同时考虑图像和音频的特征
   目前图像的特征一般属于 CV 领域，而音频特征一般属于 Speech 领域，我们希望将两个领域的长处结合起来，使得它们的研究成果，能够在现实生活中（也就是视频内容分割和筛选这个需求中）得到应用

2. 将视频分类的结果，用于优化视频分割的效果
   我们现有的视频分割片段是基于镜头片段的，它注重视频的图像相似度，却忽略视频的表达含义；而我们的分类结果，可以对视频的意义进行一部分说明，这样我们就可以将零碎的片段整合在一起，使得视频的分割效果更好。

3. 采用深度神经网络和 *GPU* 计算，极大的增强了算法的准确性和效率

   我们利用 DeCAF 的结果，使用已有的训练了数千万张 ImageNet 数据集上图片的神经网络，一方面，可以免去自己训练神经网络的经费和时间，另一方面，这些网络由于已经“见过”许多图片，因此对图片的含义有着比 HOG 和 SIFT 更深的认识，提取其特征分类准确性更高；我们还使用 CUDA 和 Nvidia GPU 来加速我们的特征提取过程，使得在笔记本电脑上的特征提取速度可以和视频播放的帧率相当。

4 算法的理论基础

我们将这个问题分为两个部分：视频的切割和视频的分类。目前，根据镜头片段做出的视频切割已经达到非常成熟的效果，那么如果我们能够将视频分类得出，那么视频分割也可以得到非常良好的结果。

根据我们在计算机视觉上的经验，一种有效的分类器训练方法是通过一种固定的办法视频片段得到特征向量，再利用一个有效的分类器（例如支持向量机和人工神经网络）来训练这些向量；这样，当我们对一个新视频进行分类时，就可以得到其特征向量，然后用分类器得到分类结果。

视频片段的信息一共分为两种：图像信息和语音信息。一种得到特征向量的方法是，我们可以通过现有的成熟方法，分别对这两种信息提取特征向量，然后将其合并，通过支持向量机进行训练，得到分类器。

常见的音频特征有以下两种：LPC 和 MFCC。其中 LPC(Linear Predictive Coding) 通过每秒 30 到 50 次的采样得到音频特征；而 MFCC(Mel Frequence Cepstral Coefficient) 是基于对一个对数能量频谱的线性余弦变换得到的。这两种特征都是广泛使用与声音处理领域的特征。

常见的图像特征有 HOG(Histogram of Gradients)，SIFT(Scale-invariant feature transform) 等，他们也在物体监测，关键点识别等领域得到广泛应用，也常被用作特征提取的手段。最近，UC Berkeley 的 Trevor Darrel 组提出了使用深度神经网路提取特征的方法 (DeCAF)，也取得了很好的效果，其工作也被广泛关注。

我们希望通过利用这些在各自领域中广泛应用的特征的长处，来帮助我们的分类器得到很好的效果。

5 实验流程安排

本次实验的计划开发流程如下表所示：

日期	安排
11 月 8 日	完成课题构思与可行性分析
11 月 15 日	文献调查，整理实验方法
11 月 22 日	完成数据收集工作
11 月 29 日	实现实验方法，完成界面设计
12 月 7 日	完成对不同节目的划分与分类
12 月 15 日	完成对一个节目不同部分的划分与分类
12 月 20 日	进一步完善实验结果，扩大数据规模

6 初步实验结果

6.1 数据集收集

我们收集了来自 CNTV、优酷网、土豆网的不同类型的视频节目，有音乐 MV、动画、自然类、讲座和新闻播报五种类型的节目共约 600MB 的视频资源。这些视频资源通过本地的解码转码和重采样，最终获得分辨率均为 480*320，每秒 15 帧，音频采样率为 44100Hz 的视频片段。

通过 ffmpeg 工具和使用 python 自行实现的小工具，我们能够对特定文件夹下的所有视频划分训练数据集和测试数据集，并实现视频的拼接工作，得到可供程序测试和播放的目标视频文件。

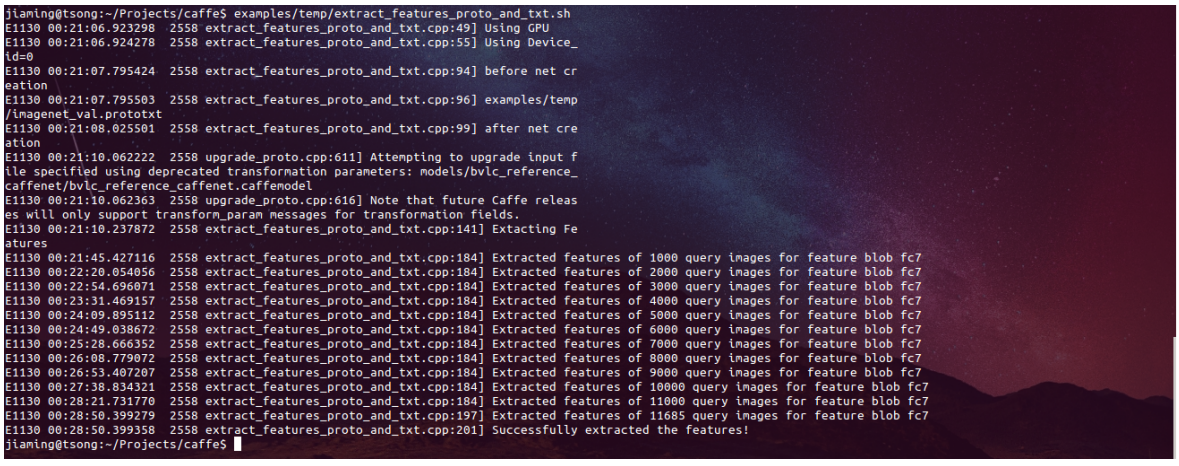
6.2 图像特征提取与预测

我们采用的模型是 Alex 在 2012 年的一片 paper 上提到的 AlexNet 模型，它是一个 7 层神经网络，如下图所示。其中前五层是卷积网络，后两层是全连接网络，用 fc6 和 fc7 表示。后两层网络各 4096 维，作为提取的特征向量非常合适。

我先在 Ubuntu 14.04 环境下安装了开源的深度学习网络框架 Caffe，然后在其基础上对程序进行修改，使得我们能够以字符串形式提取出神经网络中的特征；之后，我编写了生成 SVM 训练集的程序，使得提取出的特征和其分类一一对应；最后，我们使用开源的 LIBLINEAR 库作为我们的大规模线性 SVM 分类器，使用 5 对折交叉验证得到交叉验证的平均准确率。

我们采用的数据是从五类视频：新闻，综艺，讲座，动画，和自然的视频中提取出每一帧得到的图片训练集。

各阶段的实验结果如下：特征提取阶段，由于使用了 GPU，我们提取特征的速率要比在 CPU 上快接近 10 倍，提取 10000 多张图片用时在 10 分钟左右，平均每秒中提取 25 张图片的特征，这已经接近视频播放的帧率了。



在分类阶段，我们的 SVM 训练集为 10000 多张图片，每一个图片用一个 4096 维数的神经网络表示（取自 fc7 层），所以数据规模还是比较庞大的。我们使用 LIBLINEAR 训练的时间大约为 1 分钟，交叉验证的准确率达到 99.85

```
nSV = 440
..*.*
optimization finished, #iter = 31
Objective value = -0.019160
nSV = 331
.................*....***.
optimization finished, #iter = 240
Objective value = -0.099208
nSV = 337
.**
optimization finished, #iter = 15
Objective value = -0.005257
nSV = 190
.....................................................................................
.................
optimization finished, #iter = 1000

WARNING: reaching max number of iterations
Using -s 2 may be faster (also see FAQ)

Objective value = -6.356085
nSV = 344
Cross Validation Accuracy = 99.8545%
jiaming@tsong:~/Projects/liblinear-1.96$
```

我们认为这个实验结果可以至少说明以下两点：

1. 我们的算法速度可以达到实际应用要求的速度
2. 神经网络提取出的特征对图片具有很强的代表性

我们认为我们的试验中还有训练集图片之间差别相对较小的问题（因为很多来自同一个视频），使得训练集和测试集的区别并不是非常明显。但是，很高的正确率依然可以表明，这个模型的拟合程度很好，在不同的数据集上依然有很好的潜力。

6.3 图形界面开发

目前图形界面的开发已经初步完成。界面示意图如下：

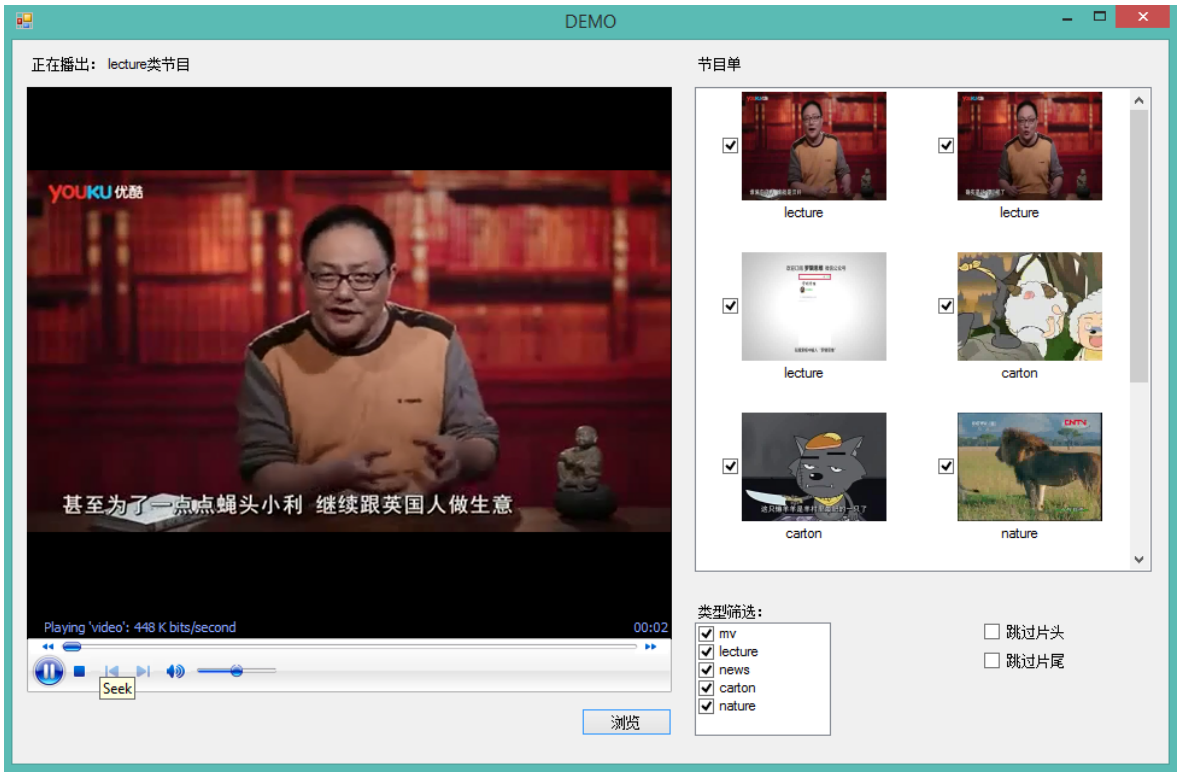


图 1: 用户界面示意图

用户界面已经实现的功能有：

1. 读取算法的输出结果文件 description.txt，获取视频文件信息和视频分段、分类信息
2. 从结果文件中获得视频中含有的节目类型列表
3. 获取视频中每个节目中的一个镜头作为缩略图显示
4. 支持用户对节目类型进行筛选观看，即跳过特定类型节目，或者跳过特定某个节目
5. 节目与节目之间切换连贯，不会产生中断观看的情况

7 中期检查目标

在已经完成内容的基础上，我们还会完成以下内容：

1. 将目前的数据处理方法用脚本的形式连接在一起，使得特征提取和训练能够连续进行
2. 控制训练集和测试集的比例，来调查实验数据是否有缺陷

8 期末检查目标

在中期的检查目标的基础上，额外增加以下的内容：

1. 增加对节目的不同部分的识别与分类功能，提供类似“跳过片头、跳过片尾”功能
2. 通过改进图像和音频的特征来提高实验性能
3. 扩大实验数据集规模，扩大视频种类数，以进一步评估实验方法的可推广性

参考文献

[1] Dhanalakshmi and et al. Classification of audio signals using SVM and RBFNN. Expert Systems with Applications, 2009.

[2] Dhanalakshmi and et al. Classification of audio signals using AANN and GMM. Applied Soft Computing, 2011.

[3] Tong Zhang and C. -C Jay kuo. Hierarchical Classification of Audio Data for Archiving and Retrieving.

[4] Rainer Lienhart and et al. Scene Detection based on Video and Audio Features.

[5] Katharina Morik. Automatic Feature Extraction for Classifying Audio Data. Machine Learning 58, 2005.

[6] Dongge Li and et al. Classification of general audio data for content-based retrieval. Pattern Recognition Letters 22, 2001.

[7] Simon Moncrieff and et al. Detecting Indexical Signs in Film Audio for Scene Interpretation.

[8] Dhanalakshmi and et al. Pattern classification models for classifying and indexing audio signals. Engineering ApplicationsofArtificialIntelligence24(2011)

A 目录结构

- *demo* 图形化界面所在文件夹
- *material* 所有参考文献所在文件夹
- *mklab* 视频 shot/scene 切割模块所在文件夹
- *video* 数据集所在文件夹
- *workspace* 主要工作目录
  - *audioFeature* 音频特征提取
  - *baseline* 单独使用音频信息的结果文件夹
  - *features* 特征文件缓存文件夹
  - *img* 视频图像信息缓存文件夹
  - *seg* 视频切割文件夹
  - *wav* 音频信息缓存