视听信息系统导论课程设计
视频情感分类系统

马栩杰, 陈彦熹, 宋长河

2017 年 1 月 6 日

小组成员及分工

马栩杰主要完成系统结构设计、人脸表情特征提取、人物年龄特征提取、场景特征提取、PCA 降维、SVM 分类器, 陈彦熹主要完成视频图像人脸检测、人脸对准模块、LDA 分类器, 宋长河主要完成音频特征提取模块、特征降维调优、分类器调优。

文件列表

代码:

- `fun_process.m` 程序接口
- `loadAgeNet.m` 加载年龄模型
- `loadPlaceNet.m` 加载场景模型
- `loadEmotionNet.m` 加载表情模型
- `ageGender.m` 封装的 Caffe 接口
- `places.m` 封装的 Caffe 接口
- `emotion.m` 封装的 Caffe 接口
- `inputTransform.m` 将图片转换为模型输入格式
- `faceDetector.m` 人脸检测
- `faceAffine.m` 人脸对准
- `emotionFeature.m` 提取人脸表情特征
- `ageFeature.m` 提取年龄特征
- `placeFeature.m` 提取场景特征
- `my_mfcc.m` MFCC
- `extract_audio_features.m` 音频特征

模型 (model 目录下):

- `ldamodel.mat` LDA 分类器
- `pcaresult.mat` PCA 变换矩阵
- `agegender/` 年龄模型
- `alexnet_emotion/` 表情模型
- `googlenet_places205/` 场景模型

库/外部程序:

- `face_detect_align` 人脸检测程序
- `mfcc/` MFCC 库

1 系统架构

AFEW 的数据集由一个个很短的从电影中截取的片段构成，每个电影片段包含若干连续的视频帧以及对应的音频。从数据中，我们可以发掘出很多有助于我们对视频情感进行分类的信息，例如视频所涉及的人数、人物的表情及表情的变化、人物的性别与年龄、人物的动作、所处的环境、说话时包含的感情色彩、对话的内容、配乐等等。我们主要探索了人物表情、人物性别与年龄、所处场景、音频情感色彩这几项特征与电影片段情感类别之间的关系，并最终选择了主要角色表情、年龄、场景、音频情感特征作为分类的依据。

系统架构如图 1 所示。

忽略视频与音频之间的相互关联, 分别进行处理。

对于视频部分，首先忽略连续的视频帧与帧之间的关联, 对整个视频的所有帧做均匀采样, 取其中 1/5 帧提取场景特征。同时，逐帧检测视频中的人脸，对检测到的人脸首先检测其关键点并对准到预设的标准脸位置，然后剪切出人脸部分，逐人脸通过 CNN 检测其表情特征和年龄特征。如此处理后，视频每帧会提取出一份场景特征，视频中检测到的每个人脸提取出一份表情特征和一份年龄特征。我们注意到人脸检测模块有时可能在非人脸的区域也有响应，而剪切到非人脸区域以后再进行表情检测和年龄检测会引入相当大的误差。但是对于非人脸区域，我们发现表情检测网络的输出往往会有比较明显的特征，其最后一级全连接层的 7 个输出往往都很小，并且没有显著的最大值，从而 Softmax 输出的最大值很小，以此为依据，在此时将人脸检测检出的非人脸图像删去，以提高后续处理的效果。为了使不同视频的特征长度相同, 对特征做逐帧 (场景) 与逐人脸 (表情、年龄) 平均，为避免损失一些比较显著的特征，同时也保留所有帧/人脸的特征各维度上的最大值。

对于音频部分, 我们采用了音频的能量、过零率、

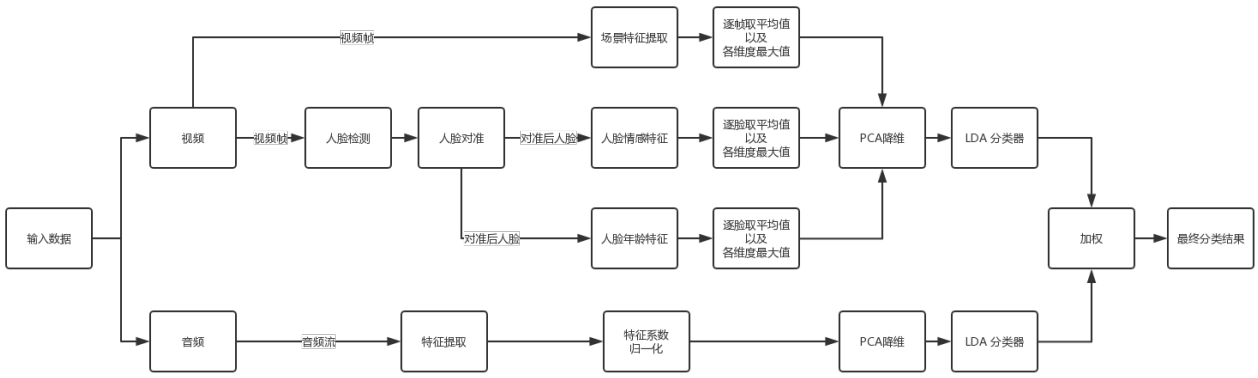

图 1: 系统整体架构

	特征	逐帧/逐人脸	整体	降维后
视频	表情	7*人脸个数	7+7	17
	年龄	8*人脸个数	8+8	
	场景	205*帧数	205+205	
音频	MFCC 等	N/A	64	18

表 1: 特征维度

频谱、MFCC 特征，从中选取出对结果影响显著的数值，并使用音频长度对特征数值做归一化，得到音频特征。

由于此时得到的特征维度比较高，我们首先使用 PCA 方法对数据进行降维，然后分别训练多个针对视频特征的分类器和针对音频特征的 LDA 分类器，最后对分类器输出的置信概率进行加权求和，得到最终的分类结果。

系统各个部分输出的特征维度如表 1 所示。

2 视频特征提取

2.1 人脸检测与对准

采用 [1] 提供的基于 OpenCV 的人脸检测校准工具，对于每一视频帧，利用该工具生成的 5 个人脸关键点坐标，对照标准 5 点坐标作仿射变换并截取，以获得校准的人脸图像。

由于该人脸检测工具可能出现无法成功检测到人脸的情况，在实践中，我们对无法检测到人脸的数据使用 MATLAB 的人脸检测工具再做一次检测。对于依然无法检测到人脸的视频片段，我们选择直接放弃基于人脸的特征。

2.2 人脸表情特征

视频中出现的人物的表情特征对视频情感分类的作用最为显著。经过调研并比较了几种不同的模型的效果，我们使用 CNN 来进行人脸表情分类。

本部分使用的模型来自 [2] 。该模型使用在 ImageNet 上 pretrain 过的 AlexNet，输出调整为 7 维，对应数据集的 7 种情感标注。先用 FER2013 数据集进行一次 finetune，然后再使用 SFEW 数据进行 finetune。原模型在 FER2013 测试集上的正确率为 64.74% 。

在此基础上，为了进一步提升该模型在 AFEW 数据上的效果，我们尝试继续使用 AFEW 数据对模型进行 finetune。在测试时，我们发现在 AFEW 上的再次 finetune 并没有起到任何作用，反而让模型的准确率有所下降。这很可能是因为截取 AFEW 视频片段中的人脸，其表情并不一定与视频的标注相符，由此而引入了显著的噪声。

使用该网络的 Softmax 输出作为人脸的表情特征。

2.3 人物年龄特征

考虑到电影片段中出现角色的年龄也可能与影片的情感相关，我们加入了人物年龄识别模型。

我们使用了 [3] 训练的模型。该模型使用自行设计的有 3 个卷积层、含有 dropout 单元的 CNN 进行分类，在 Adience benchmark 数据集上进行训练，在相应的测试集上可以达到 45.1% 左右的年龄分类正确率。同样使用网络的 Softmax 输出作为年龄特征。

2.4 场景特征

影片片段中人物所处的场景往往也暗示了影片的情感，因此我们也加入了视频场景特征提取模块。

我们使用 [4] 训练的模型，该模型为在 Places205 数据集上训练的 GoogLeNet。直接用一帧的图像作为网络输入，然后得到 205 个场景分类的 Softmax 输出作为场景特征。

3 音频特征提取

根据日常经验，说话人的声音中饱含着情感特征：例如愤怒的声音普遍音量很大，喜悦的声音轻快流畅如此等等。所以对于视频的情感分类，除了图像人脸特征以外，音频特征也是不可或缺的依据。为此我们从时域特征和频域特征两个方面出发，参考 [5] ，并通过实验对比效果，选取了四个最有效的特征，分别是时域的能量和过零率，以及频域的频谱和 MFCC。最终对于每个样例可以处理得到一个拥有 64 个分量的特征向量。

能量是音频时域中最主要的特征之一，我们可以通过它很好的区分说话人的情绪激动程度。主要情感中，愤怒和高兴的能量较大，而悲伤和害怕的能量通常较小。

过零率是音频信号经过零点跳变的频率，所以它很大程度上可以反映音频的频率特征。而频率是区分音频情感的一个很必要的特征。并且考虑到最终分类检测时会有时间限制，所以选择了过零率这种可以较为省时地反映频率的特征。

频域分析最直接最全面的方法就是频谱分析。借鉴资料经验后，选用短时傅里叶分析音频，因为这样可以同时考虑到声音特征的短时性。具体方法是首先把音频序列分成长度 20ms 的段，同时在段与段之间留下 50% 的交叠以减少信息损失。然后将每一段加窗，并重采样到 5 个点，计算其 FFT。最后使用最大值和方差来描述每个 FFT 系数。比较后可以得出：愤怒和高兴等均表现为基频最大值和方差的提高和频谱中高频成分的增加。与此相反，悲伤对应于基频最大值的减小，以及频谱中高频成分的降低。

选取 MFCC 是因为它使用线性的倒频谱表示方法，与人类的非线性听觉系统更加接近，所以被大量使用在音频处理分析中。具体方法是：对于 MFCC 得到的一串数字，以 Hamming 窗为窗函数，每段时间取 13 个 MFCC 系数，得到系数矩阵。其中对每个系数选用 4 个统计量：方差、最大值、最大值与平均值的比值、中位数。

4 分类器设计

4.1 特征降维

由于提取出的特征维数比较高，并且 AFEW 数据集相当小，在当前提取出特征的基础上直接训练分类器是一件相当困难的工作。在实际尝试时，直接使用高维特征训练 SVM 或 LDA 分类器会导致严重的过拟合，导致训练后分类器几乎完全失效。

为了让分类器收敛，使用 PCA 对特征做降维，然后使用降维的特征训练分类器。由于视频特征使用 CNN 的 Softmax 输出作为特征，而音频部分使用 MFCC 库及统计方法提取特征，两者特征的统计特性有显著的差异，因此在实践中我们对视频和音频分别独立地做 PCA 进行降维。在经过降维以后，分类器的训练效果有非常显著的提升，并且在实验时注意到训练出分类器的泛化准确率与降维后的特征维度之间的关系基本呈现一个单峰函数的形态。据此，我们分别取视频和音频在特征降维后准确率达到峰值的维度，用于训练分类器和进行测试。具体维度如表 1 所示。

4.2 分类器

4.2.1 SVM

由于 SVM 本身是二分类器，为了在多分类问题中应用 SVM，我们尝试了训练多个一对多分类器、训练多个二对多分类器并投票、训练多个类两两之间的二分类器并投票这 3 种方法，并且最终得出结论，训练多个二对多的分类器并且投票的效果最佳。具体做法是对标签的两两组合 (L_i, L_j)，将训练集中标签为 L_i 的和标签为 L_j 的数据均标记为 1，其他数据标记为 0，训练 SVM。如此操作，遍历所有的 (L_i, L_j) 组合，我们可以得到 49 个分类器，再用这 49 个分类器的结果的预测结果进行投票，选择得票数最高的分类，即得到多分类的结果。

4.2.2 LDA

同时我们也尝试了采用线性判别分析 (LDA) 对降维后的特征进行分类。LDA 的基本原理是，在贝叶斯最佳解决方案基础上，通过假设条件概率密度 $p(x|y)$ 为正态分布，得到二次判别分析；再作出额外简化的方差齐性假设，且协方差满秩，从而将分类标准转化为观察值的线性组合函数。对于多类情形，使用 Fisher 判别派生出的分析方法，LDA 可以直接进行多类分类。

特征 \ 方法	SVM	LDA
音频	24.77%	30.01 %
仅表情	34.35%	34.68 %
视频	34.04%	36.29 %
视频 + 音频	未做	**41.40** %

表 2: 视频与音频特征测试准确率

4.3 不同特征分类器的组合

在将视频与音频的特征进行组合时，所采用的方法是分别使用视频和音频特征训练分类器，然后对分类后的置信概率进行加权求和。对于 LDA 分类器，分类器可以直接输出每个分类的置信概率。对于 SVM，我们采用对投票结果做 Softmax 后的结果作为分类的置信概率。

在实验中，我们发现对 SVM 投票结果做 Softmax 然后进行加权求和的效果并不好，分类结果非常不稳定，因此放弃该做法；而 LDA 分类器做置信概率的求和则可以表现出显著的效果。实验中，我们使用加权系数 1:0.75 分别对视频和音频的预测分类置信概率进行加和。

5 实验

由于 AFEW 的数据集太小，如何充分地利用数据就成了一个难点。为了保证有足够的数据进行训练，我们额外取得了 AFEW 的验证集，并且不对训练集与验证集做区分。

而为了确保在不事前划定训练集、验证集和测试集的情况下，使用尽可能多的数据进行训练，并且还可以保证测试结果的可信度，我们采取了如下测试方式：随机选择数据集的 90% 进行训练，另外 10% 进行测试，计算测试集上的正确率，重复进行此过程多次并且求多次测试正确率的平均值。当重复次数足够多时，正确率的平均值会收敛到一个稳定的值，这个值就视为模型的泛化准确率。

测试结果如表 2 所示。表中所有数据均为进行 200 次独立重复的训练与测试过程得到正确率的平均值。可以看出，使用多种不同模型的组合确实有助于提高系统整体分类的性能。最佳方法的测试效果可以稳定达到 41.40 %。

图 2 是使用视频与音频特征、LDA 分类器的 Confusion Matrix。其中 Angry、Happy 和 Neural 三个标签的分类正确率远远高于其他类别，而 Disgust

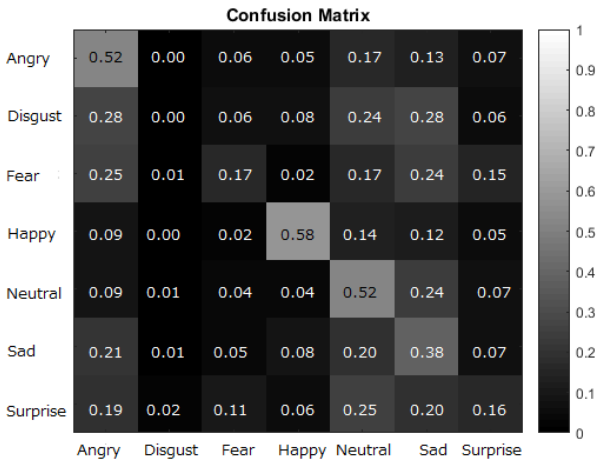

图 2: Confusion Matrix

则几乎无法分类。这与数据集的分布以及表情本身的易辨认性有关。

参考文献

[1] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *ACM International Conference on Multimodal Interaction*, 2016, pp. 445–450.

[2] H. W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *ACM International Conference on Multimodal Interaction*, 2015, pp. 443–449.

[3] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.

[4] B. Zhou, A. L. Garcia, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," *Advances in Neural Information Processing Systems*, vol. 1, pp. 487–495, 2014.

[5] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.