CCML 2021 第十八届中国机器学习会议
The 18th China Conference on Machine Learning
2021年8月6-8日 中国·长沙
AUG 6-8, 2021 CHANGSHA, CHINA

一种基于决策层融合的多模态情感识别方法

韩天翊[1,2] 林荣恒[1,2]
[1]北京邮电大学计算机学院（国家示范性软件学院） 北京 100000
[2]网络与交换技术国家重点实验室（北京邮电大学）北京 100000
Email:tyhan@bupt.edu.cn

介绍

情感是一系列主观认知体验的总称。它是一种心理和生理状态，是各种感觉和行为的结合。情感通常与心情、性格等因素相互作用，也会受到激素的影响。不同的情绪对日常行为会有着不同的导向作用，我们做的每件事都会有不同的情感表达。了解用户使用产品时的情感，可以大大提高服务质量和服务效果。

近年来，随着科学技术的日益发达，情感识别的效率和准确性均有大幅提升。情感识别系统能够帮助客服人员预先了解客户的情感状况，提高服务质量，进而改善客户的服务效率和满意度。但是，情感识别在单一模态下很难进一步提高识别准确率，又往往需要使用高性能的GPU，这使得系统的使用范围很小，情感识别很难在现实生活中表现出应有的效果。因此，必须减少系统的使用条件，设计一个软硬件结合的多模态情感识别系统。多模态融合可以提高系统识别准确率，软硬件结合可以并行加速计算过程，让系统可以在更广泛的范围内使用。

本文设计并实现了一种软硬结合的多模态情感识别系统。该系统使用了语音和面部表情两个模态，通过梅尔倒频谱系数与卷积神经网络对情感进行识别和分类，同时将语音情感识别迁移到神经网络计算棒以降低环境负载。在模态融合时，系统采用决策层融合的方式来提高识别准确率。

方法

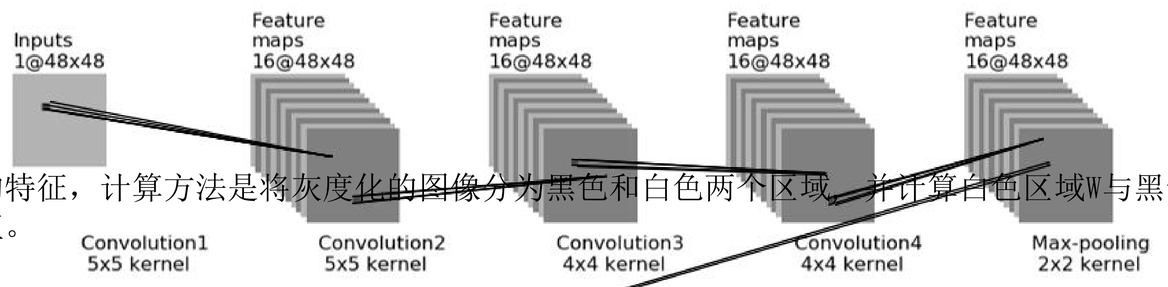
语音情感识别

梅尔频率倒谱系数是目前应用于情感识别中重要的声学特征。MFCC计算公式如下：

其中，ml表示滤波器输出的对数，L是滤波器的个数。

对于音频的情感识别，我们使用AlexNet卷积神经网络[1]。为了满足情感识别过程的环境要求并并行加速计算过程，让系统可以在更广泛的范围内使用，我们将训练后的卷积神经网络运行在Intel神经网络计算棒上。

- 准备好使用的模型，为其配置模型优化器。
- 使用包括特定网络拓扑和权重以及其余相关参数的训练网络当作模型优化器的输入。
- 使用模型优化器转换我们要使用的模型。
- 模型优化器生成网络的中间层，当作推理引擎的输入部分。
- 使用推理引擎来优化目标硬件的推理执行。

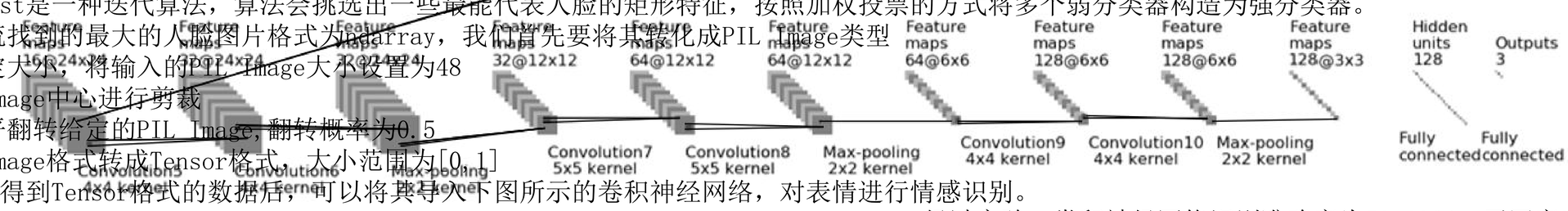


视频情感识别

Haar特征是基于“块”的特征，计算方法是将灰度化的图像分为黑色和白色两个区域，并计算白色区域W与黑色区域B的像素值之和的差值，乘以相应的权重系数T，得到i区域Haar特征值。

Adaboost是一种迭代算法，算法会挑选出一些最能代表人脸的矩形特征，按照加权投票的方式将多个弱分类器构造为强分类器。

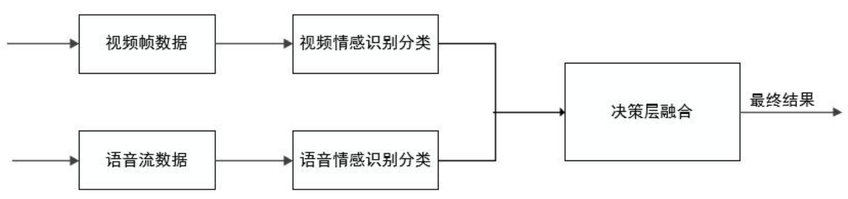
- 从视频流找到的最大的人脸图片格式为ndarray，我们首先要将其转化成PIL Image类型
- 重新设定大小，将输入的PIL Image大小设置为48
- 在PIL Image中心进行剪裁
- 随机水平翻转给定的PIL Image,翻转概率为0.5
- 将PIL Image格式转成Tensor格式，大小范围为[0,1]



在我们得到Tensor格式的数据后，可以将其导入下图所示的卷积神经网络，对表情进行情感识别。

在单模态情感识别结束后，我们得到了分别由语音和视频识别的结果。为了将使系统总体的准确率提高，我们需要将两种模态进行融合。多模态融合就是通过不同模态之间的关联性，将多个角度的数据结合来提高总体识别准确率。决策层融合[18]是在提取不同模态的数据后，各自独立的进行情感分类，最后利用两种模态的结果得到多模态的识别结果。本文中决策层融合采用加权求和的方式。

其中，w_0和w_1分别是语音和视频情感识别的权重，且$w_0+w_1=1$



结果

系统对开心、生气与正常三种表情进行识别。语音识别模型使用营业厅通话录音作为数据集。数据集按照8:2的比例随机划分为训练集与测试集。使用测试集对语音模型测试，其分类准确度为74.5%。同时，为了测试神经网络计算棒对系统性能的影响，选择了不同长度的音频文件进行测试。表1显示了对于相同时长的音频在不同硬件上的运行时间。

Table 1 Comparison table for different hardware

表 1 不同硬件所需时间的对比

Duration of audio	*Performance Testing*	
	CPU	Neural network computing stick
18sec	2sec	1sec
2min 33sec	4sec	5sec
4min	5sec	6sec

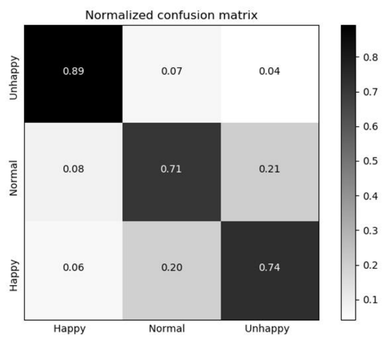
视频模型的数据集结合了Fer2013，CK +和GENKI数据集，训练集和测试集按8：2进行划分。使用支持向量机和深度神经网络作为基线算法。

经过实验，卷积神经网络识别准确率为78.62%，召回率77.98%，优于支持向量机和深度神经网络的识别准确率，表2显示了不同方法对情感的识别结果。取w0为0.3，w1=0.7，系统经过决策层融合，系统准确率提升了3.4%。相比于文献[2]和文献[3]，以及多流隐马尔可夫模型[4]和异步DBN模型[4]，可以得到相似或更高的准确率。

Table 2 Recognition results of different methods

表 2 不同方法对情感的识别结果

Method	Accuracy	Loss
SVM	53.3%	-
DNN	70%	0.7386
CNN	78.62%	0.3520



总结

针对传统情感识别系统准确率较低和往往需要高性能环境的问题，本文设计和实现一种软硬结合的多模态情感识别系统以降低环境负载。语音情感识别首先对每句语音进行语音信号预处理，提取语音的梅尔倒频谱系数特征，将卷积神经网络迁移至神经网络计算棒对其进行情感识别分类。视频情感识别使用haar特征与级联分类器从图片中切割人脸表情,通过卷积神经网络对表情进行识别和分类。实验结果表明，本次研究采用的软硬结合的多模态情感识别方法，系统拥有较高的识别准确率且能够在性能较差的运行环境中保持运行速度。

参考文献

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25: 1097-1105.
[2] 闫静杰,卢官明,李海波,王珊珊.基于人脸表情和语音的双模态情感识别[J].南京邮电大学学报(自然科学版),2018,38(01):60-65.
[3] Datcu D, Rothkrantz L. Multimodal recognition of emotions in car environments[J]. DCI&I 2009, 2009.
[4] Jiang D, Cui Y, Zhang X, et al. Audio visual emotion recognition based on triple-stream dynamic bayesian network models[C]//International Conference on Affective Computing and Intelligent Interaction. Springer, Berlin, Heidelberg, 2011: 609-618.

投稿人 姓名：韩天翊
单位：北京邮电大学