doi: 10.7690/bgzdh.2024.09.009

基于 CNN 的在线多媒体英语教学情感交互研究

梁　珊
（陕西交通职业技术学院教务处，西安 710000）

摘要：针对多媒体英语教学中情感缺失的问题，提出一种基于人脸表情识别的智能网络教学系统模型。应用主成分分析（principal component analysis，PCA）提取在线学习者视频中面部表情的重要特征帧；基于 CNN 架构的面部情绪识别网络判断和理解学习者的情绪状态，根据学习者的具体情绪状态给予相应的情绪鼓励或情绪补偿策略。仿真结果表明：VGG16 和 ResNet50 比较，该算法平均检测率为 78.28%，平均识别准确率为 81.78%，性能明显较优。

关键词：多媒体英语教学；情感；人脸表情识别；卷积神经网络
中图分类号：TP393　　**文献标志码**：A

A Study of Emotional Interaction in Online Multimedia English Teaching Based on CNN

Liang Shan
(*Academic Affairs Office*, *Shaanxi College of Communication Technology*, *Xi'an* 710000, *China*)

Abstract: In order to solve the problem of emotion absence in multimedia English teaching, an intelligent network teaching system model based on facial expression recognition is proposed. Principal component analysis (PCA) is applied to extract the important feature frames of facial expressions in online learner videos; the facial emotion recognition network based on CNN architecture judges and understands the emotional state of the learner, and gives corresponding emotional encouragement or emotional compensation strategies according to the specific emotional state of the learner. Simulation results show that the average detection rate of the proposed algorithm is 78.28%, and the average recognition accuracy is 81.78%, compared with VGG16 and ResNet50.

Keywords: multimedia english teaching; emotion; facial expression recognition; convolutional neural network

0　引言

在线学习[1-2]是利用多媒体计算机技术和网络技术实现优化教育的一种教育方式。在线学习在资源共享和互动性方面远远优于传统教育。由于在线教育的一个关键要素是师生分离、学校与学生分离；因此，在线教育环境中缺乏师生互动及情感交流[3-4]，面部表情、声音和手势带来的情感信息在学习传输过程中缺失。所有这些情感信息的缺失都会影响师生之间的情感互动。一方面，学生很难感受到老师对他们的关注，在学习中容易产生困惑和懒惰的情绪；另一方面，教师也难以理解学生的感受，难以有效控制学生的学习过程。

针对在线学习技术，大量学者对其进行研究，并取得了丰硕成果。文献[5]通过阐述建模语言、块的结构、权限管理和挖掘规则等构建资源区块链基本要素，探讨区块链技术在在线学习资源管理方面的应用。文献[6]深入分析注意力及在线学习中的表现规律，提出适用于在线授课的注意力策略框架。文献[7]提出了基于在线学习的个性化学习路径推荐系统。文献[8]基于大数据分析，设计了教学目标设定、有效教学内容推送、有效教学方法实施和有效教学干预的在线教学模式。

大多数在线教育只是用计算机以及网络等为代表的多媒体来取代传统媒体，用先进的信息技术为简单的交流工具，利用网络技术进行“搬书”或“电子课”教学，并在互联网上发布一些文本教学内容或实践问题。这种“电子教科书”没有充分利用多媒体交互技术，也没有充分发挥网络的作用，缺乏情感激励。单调的文本信息传输取代了丰富多彩的课堂教学。学习者只看到与教材类似的流媒体教材，而看不到教师“情感教学角色”的表现。学习者与计算机之间的交互仅依赖于键盘和鼠标。计算机不仅没有视觉功能、语言功能和听觉功能，且不具备理解和适应人们情绪的能力。当学习者长时间面对这样一个没有情感的冷漠的电脑屏幕，感觉不到互动的乐趣和情感的刺激时，就会产生厌烦情绪，从

收稿日期：2024-05-19；**修回日期**：2024-06-23
第一作者：梁　珊（1987—），女，陕西人，硕士。

而影响学习者的学习效果。

笔者针对多媒体英语教学中情感缺失的问题，提出一种基于人脸表情识别的智能网络教学系统模型。该模型通过捕捉和识别在线学习者的面部表情，判断和理解学习者的情绪状态，根据学习者的具体情绪状态给予相应的情绪鼓励或情绪补偿策略，从而在一定程度上帮助学习者弥补在线学习中情感缺失问题。

1　模型介绍

首先说明在线多媒体英语教学中特征帧的生成和选择；然后讨论分类模型及其相应的参数和体系结构。

1.1　特征帧提取

主成分分析(PCA)算法用于在线多媒体英语教学中特征帧的生成。PCA 是计算机视觉中广泛应用的降维技术，通过线性投影选择降维，最大化分离所有投影样本。

令$\{x_1, x_2, \cdots, x_N\}$为 n 维空间中 N 个样本图像的集合。假设线性变换将 n 维空间映射到 m 维特征空间，其中 $m<n$。进一步，令特征向量 $\boldsymbol{y}_j \in \mathbb{R}^m$ 表示线性变换，则有：

$$\boldsymbol{y}_j = \boldsymbol{Q}^{\mathrm{T}} x_j,\ j = 1, 2, \cdots, N 。 \tag{1}$$

式中：$\boldsymbol{Q} \in \mathbb{R}^{n \times m}$ 为正交矩阵。此外，散射矩阵 $\boldsymbol{S}_T$ 计算如下：

$$\boldsymbol{S}_T = \sum_{k=1}^{N} (x_k - \mu)(x_k - \mu)^{\mathrm{T}} 。 \tag{2}$$

式中 $\boldsymbol{\mu} \in \mathbb{R}^n$ 为平均面部图像。

进一步，应用线性变换 $\boldsymbol{Q}^{\mathrm{T}}$ 获得新特征向量$\{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$的散射矩阵，记为 $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{S}_T\boldsymbol{Q}$。为使投影样本的散射矩阵的行列式最大化，选择投影 $\boldsymbol{Q}_{\mathrm{opt}}$ 如下：

$$\boldsymbol{Q}_{\mathrm{opt}} = \arg\max_{\boldsymbol{Q}} \left|\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{S}_T\boldsymbol{Q}\right| = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_m] 。 \tag{3}$$

式中 $\boldsymbol{w}_i$ 为 $\boldsymbol{S}_T$ 的第 i 个 n 维特征向量。该特征空间的第一个正交维捕获数据库中最大的方差，而其最后一个维捕获数据库中最小的方差。这些特征向量与原始图像具有相同的维数将被视为特征帧。

笔者对每个视频的帧序列应用 PCA 来选择一组具有代表性的特征帧。为选择最优子集，执行步骤如下：

1) 选择非零特征值的相应特征向量来创建最优特征空间；

2) 丢弃最底部 40%的特征向量；

3) 假设前 3 个特征向量受光照条件的影响，从而降低分类性能。因此，使用除前 3 个特征向量外的所有特征向量；

4) 计算保证能量 e 大于典型阈值的最小特征向量数。令 e_i 为第 i 个特征向量的累积能量，则有：

$$e_i = \sum_{j=1}^{i} \lambda_j \Big/ \sum_{j=1}^{k} \lambda_j 。 \tag{4}$$

式中：k 为非零特征值的数量；λ 为特征值；

5) 计算第 i 个特征向量的拉伸值 s_i，定义为第 i 个特征值 λ_i 与最大特征值 λ_I 的比值，表示如下：

$$s_i = \lambda_i / \lambda_I 。 \tag{5}$$

需注意，拉伸值 s_i 的典型阈值为 0.01。此外，大多数关键帧选择方法依赖于运动差异。运动差异可以计算为面部不同部分(如嘴唇和眉毛)之间的运动量。然而，这种方法需要详尽的数值计算。笔者使用从单个视频中获得的每个主成分解释的方差量，情绪面部表情视频中的变化对应于时域中面部表情的局部变化。这意味着方差解释了与面部表情动力学相关的运动差异。

1.2　预训练

为加快训练效率，笔者选择了 2 个预先训练的 2D CNN，即 VGG16 和 ResNet50，并使用通过不同关键帧选择方法获得的选定输入样本进行微调。

1.2.1　VGG16

VGG16 模型中，输入图像通过一系列卷积层，应用大小为 3×3 的卷积核，步长为 1。卷积层后接大小为 2×2 和步长为 2 的最大池化层。卷积层堆栈后接 3 个全连接层(fully-connected，FC)。最后一层是 softmax 层，使得所有隐藏层进行非线性校正。

为调整预训练 VGG16 网络的参数，使用随机梯度下降(stochastic gradient descent，SGD)进行优化，动量设置为 0.9，批大小设置为 32。通过将预先训练的 VGG16 的分类层替换为具有与目标类数量相同单位数量的 softmax 层执行微调。首先，初始化网络权重。然后，固定预训练网络卷积层的权重，仅微调 3 个 FC 层，通过保留了预先训练过的 CNN 的早期特征，从而更具体地提取细节。

1.2.2　ResNet50

ResNet50 比 VGG 网络深 8 倍，但复杂度较低。其中，卷积层大多具有大小为 3×3 的卷积核。但是，

对于相同大小的输出特征图，这些层具有相同数量的卷积核。如果特征图大小减半，则卷积核的数量将增加一倍，从而保留了每层的时间复杂性。卷积层内执行一个全局平均池层和一个具有 softmax 功能的 FC 层。

同理，笔者使用 SGD 调整预先训练的 ResNet50 网络的参数，参数与 VGG 相同。此外，权重初始化、正则化参数不变，且使用与 VGG16 相同的批大小。

2 面部情绪识别网络

2.1 网络结构

网络中的 CONV 层为 3×3(卷积核大小)。随着网络层数加深，每个 CONV 层学习的卷积数量增加了一倍。此外，所提方法使用 MSRA 方法初始化 CONV 和 FC 层，从而使网络能够更快收敛。进一步，笔者使用 ELU 代替 ReLU 提高分类精度。在上述体系结构中，在每个 CONV 层之后应用激活，然后进行批处理规一化。网络具体参数如表 1 所示，所用层的详细信息解释如下：

卷积层：卷积层的目的是充当特征提取层。CONV 层计算连接到局部输入区域的神经元的输出。它们中的每一个计算的权重是它们在输入体积中链接到的一个小区域之间的点积。当每个神经元的输入与前一层的局部感受野相连接时，从前一层提取局部特征。每个神经元从不同的局部感受野接收数据，但同一特征图中的每个神经元使用相同的卷积核。

池化层：池化层执行空间维度(宽度、高度)向下采样操作。随着卷积层数量的增加，特征图的数量也会增加，从而导致特征尺寸的急剧增加。如果使用所有特征来训练 softmax 分类器，将导致巨大的维数。因此，采用了一个池化层来降低特征维数。池层起向下采样的作用，从而收缩要素特征而不更改其编号。

FC 层：FC 层计算分类分数。FC 层中每个神经元与前一层中的所有神经元相连。

softmax 层：卷积神经网络的最后一层是 softmax 分类器。给定一个输入，通过 softmax 层每个神经元输出一个介于 0 和 1 之间的概率，从而实现面部表情的多分类任务。

为避免过度拟合，所有网络都采用了数据增强的方法进行训练。并非所有的增强方法都适用于面部图像，尤其是特征帧；因此，实验中使用的数据增强形式仅包括垂直和水平翻转、缩放和旋转 45°。

表 1 网络模型结构

层类型	大小	核大小	步长
输入	48×48×1	–	–
CONV	48×48×32	3×3	32
CONV	48×48×32	3×3	32
POOL	24×24×32	2×2	–
CONV	48×48×64	3×3	64
CONV	48×48×64	3×3	64
POOL	24×24×64	2×2	–
CONV	48×48×128	3×3	128
CONV	48×48×128	3×3	128
POOL	24×24×128	2×2	–
FC	64	–	–
FC	64	–	–
FC	6	–	–
softmax	6		

2.2 分类损失计算

给定一组来自 k 个不同数据集的图像，其中标签空间不同，定义 y_j^k 为空间 Ω^k 第 j 类的实际标签。分类的目的是学习一个非线性映射，以 CNN 的形式表示，使每个样本的交叉熵损失函数最小化，因此有：

$$L(\hat{y},y)=\frac{1}{N}\sum_{j\in\Omega^k}\left[y_j\log\hat{y}_j+(1-\hat{y}_j)\log(1-\hat{y}_j)\right]。\quad(6)$$

式中：N 为分类的数量；y_j 为第 j 类的实际标签；$\hat{y}_j$ 为 sigmoid 函数，定义如下：

$$\hat{y}_j=1/(1+e^{-\hat{p}_j})；\quad(7)$$

$$\hat{p}_j=f(Wh+b)。\quad(8)$$

式中：$\hat{p}_j$ 为 CNN 最后一层输出；$f(\bullet)$ 为该层的激活函数；W 和 b 为该层的权重和偏差；h 为最后一层的隐藏表示。

3 仿真与分析

本实验的操作平台主要基于 windows10 操作系统。由于使用了 Python 语言，因此平台需要安装与 Python 相关的环境，并使用图像框架 OpenCV 和深度学习框架 Keras。实验时 90%的数据库用于训练，其余 10%用于测试。为克服过拟合问题，对数据进行了 5 次交叉验证，以便所有类别的至少一名学生的所有视频参与训练或测试程序。

3.1 数据集

实验中使用的数据库是 CASME2 数据库。CASME2 总共有 357 个表情样本，其中 300 个宏表情，57 个微表情。表 2 为 CASME2 数据集统计情况。

表 2　CASME2 数据集统计情况

类型值	数量	统计情况	
		均值/ms	标准差
宏表情	300	1 303	651
微表情	57	419	66

同时，考虑到 CASME2 微表情数据库中非真实微笑的表情样本较少，因此使用摄像机采集 50 名学生的微笑表情作为实验测试数据。网络训练时相关参数为：学习率 10^{-4}，学习率衰减周期 15，学习衰减率 0.1，训练次数 150，Dropout 率 0.2，批大小 32。

3.2　训练过程

图 1 所示为所提模型与 VGG16 和 ResNet50 模型平均准确率对比曲线。可以看出，所提模型性能明显更优。本文中方法平均准确率为 81.78%，ResNet50 为 80.35%，VGG16 为 80.02%。

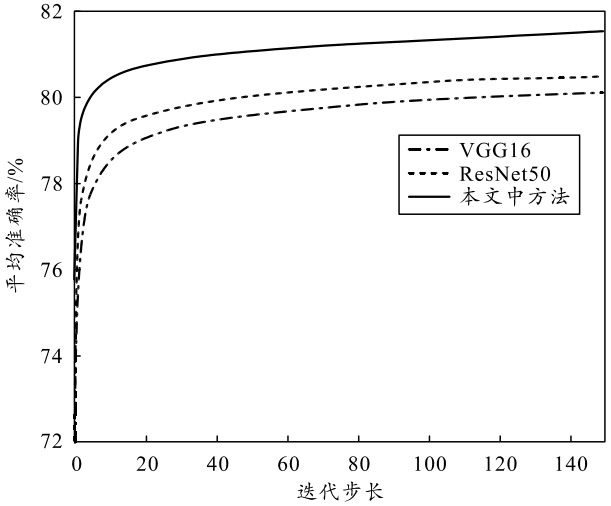

图 1　不同模型评价准确率对比结果

3.3　性能分析

根据仿真结果及数据反馈，对实验中 10 名学生的识别结果进行处理，实验结果如图 2 所示。检测率是指学生能够被检测到的表情的比例(在本实验中，表明学生表现出中性、真实或不真实的笑)。漏检率是未检测到学生表情结果的比例。识别准确率是指检测到的表情类别识别的正确结果所占的比例。识别错误率是指检测到的案例中类别错误识别的结果所占的比例。可以看出，所提模型平均检测率为 78.28%，平均识别准确率为 81.78%。实验结果表明算法识别效果明显。

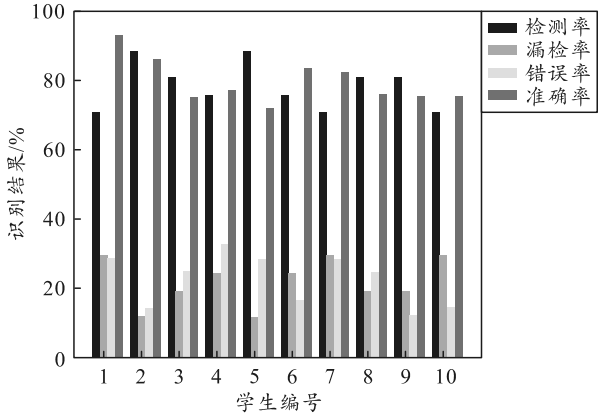

图 2　不同学生表情识别仿真结果

4　结论

针对多媒体英语教学中情感缺失的问题，提出一种基于人脸表情识别的智能网络教学系统模型。它以情感计算为理论基础，以表情识别为核心技术。笔者希望通过情感计算的相关理论和技术，特别是面部表情识别技术，建立一个能够识别面部表情的情感计算模型，并通过捕捉和识别在线学习者的面部表情来判断和理解情绪状态，从而完成学习交互过程，提升学习者在线学习效果。

参考文献：

[1] 李文昊, 陈冬敏, 李琪, 等. 在线学习情感临场感的内部特征与关系模型[J]. 现代远程教育研究, 2021, 33(4): 82-91.

[2] 张鲜华, 王少瑜, 李丹. 在线教育中教师自我表露对学生学习成效的影响[J]. 华北水利水电大学学报(社会科学版), 2021, 37(4): 64-71.

[3] 吕欣贤, 许吉婷, 毕晓璇, 等. 以山东省高校为例的大学生在线学习有效性调查及影响因素分析研究[J]. 中阿科技论坛(中英文), 2021(7): 132-136.

[4] 安玉新, 王永东. 在线教学模式下大学生学习障碍及学习效果探究[J]. 中国现代教育装备, 2021(11): 57-59.

[5] 杨雅都, 孙力. 基于区块链技术的在线学习资源管理模式研究[J]. 软件导刊, 2021, 20(7): 149-155.

[6] 李金玲, 汪凤麟. 在线课堂学生学习注意力提升策略研究[J]. 中国教育信息化, 2021(13): 22-26.

[7] 李彬. 基于在线学习的个性化学习路径推荐模式研究[J]. 电脑知识与技术, 2021, 17(18): 34-35, 40.

[8] 张汉萍. 基于大数据分析的在线教学有效性提升策略与实施路径[J]. 武汉职业技术学院学报, 2021, 20(3): 55-59.