计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(2):55−62 [doi: 10.15888/j.cnki.csa.008970]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于人脸表情识别的在线课堂学生专注度分析①

王　林, 赖梦林

(西安理工大学 自动化与信息工程学院, 西安 710048)
通信作者: 赖梦林, E-mail: 2270625123@qq.com

摘　要: 针对人脸表情识别在特征提取时容易丢失大量有用的特征信息, 无法提取更加全面的人脸表情特征的问题, 提出了一种多尺度特征融合网络模型 (DS-EfficientNet). 该模型包括深层网络和浅层网络两部分, 浅层网络用来提取面部表情的细节纹理信息, 深层网络提取表情的全局信息. 并在浅层网络中加入注意力机制, 增强对浅层细节信息的提取能力. 最终在通道上进行特征融合, 融合之后网络可以提取更加丰富的人脸表情信息. 为了减少模型参数, 提高模型的泛化性能, 将全连接层替换为全局平均池化层, 加入批归一化. 本文提出的方法在 Fer2013 和 CK+上进行实验, 识别准确率达到了 73.47% 和 98.84%. 实验证明该方法可以提取人脸更加丰富的表情信息, 模型具有更强的泛化能力.

关键词: 人脸表情识别; 特征融合; 注意力机制; 深度学习

引用格式: 王林,赖梦林.基于人脸表情识别的在线课堂学生专注度分析.计算机系统应用,2023,32(2):55–62. http://www.c-s-a.org.cn/1003-3254/8970.html

Analysis of Students' Concentration in Online Classroom Based on Facial Expression Recognition

WANG Lin, LAI Meng-Lin

(School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: Facial expression recognition is easy to lose a lot of useful feature information during feature extraction and cannot extract more comprehensive facial expression features. In view of these problems, a multi-scale feature fusion network model (DS-EfficientNet) is proposed. The model includes a deep network and a shallow network. The shallow network is used to extract the detailed texture information of facial expressions, and the deep network is used to extract the global information of expressions. An attention mechanism is added to the shallow network to enhance the ability to extract shallow detail information. Finally, feature fusion is performed on channels, and the network can extract more abundant facial expression information after the fusion. In order to reduce the model parameters and improve the generalization performance of the model, the fully connected layer is replaced by a global average pooling layer, and batch normalization is added. The method proposed in this study is tested on Fer2013 and CK+, and the recognition accuracy reaches 73.47% and 98.84%. Experiments show that this method can extract more abundant facial expression information, and the model has a strong generalization ability.

Key words: facial expression recognition; feature fusion; attention mechanism; deep learning

2020 年突然爆发的新冠肺炎疫情对我们的学习、工作和生活都造成了很大影响, 其中传统授课方式学生和老师面对面交流无疑成了一大难题, 各大高校相继选择了线上授课的方式. 在线课堂环境中, 学生和老师通过屏幕进行交流, 老师只能通过学生的面部状态判断学生听课的情况, 通过分析学生的面部表情变化

① 基金项目: 陕西省科技计划重点项目 (2017ZDCXL-GY-05-03)
收稿时间: 2022-06-24; 修改时间: 2022-07-25, 2022-09-01; 采用时间: 2022-09-14; csa 在线出版时间: 2022-11-16
CNKI 网络首发时间: 2022-11-18

可以帮助老师更好地了解学生的听课状态[1], 从而对上课模式进行及时调整. 1971 年美国学者 Ekman 等人[2]通过大量的面部表情别实验首次将人脸分为 6 种基本表情, 包括开心、惊讶、恐惧、伤心、厌恶、生气. 学生上课的情绪可以分为积极情绪、消极情绪以及中性情绪. 积极情绪包括开心、惊讶, 当学生表现出积极情绪时, 表明学生对于上课所讲授的知识处于一种愿意接受的状态, 此时他们正在认真听课, 积极思考. 消极情绪包括伤心、愤怒、恐惧、厌恶, 表明学生对老师所讲授的知识比较反感或者说上课时并没有认真听讲, 此时可判断学生上课的注意力并不集中. 而中性情绪则认为学生听课状态一般. 因此, 分析学生在线课堂的表情状态具有非常重要的意义.

当前表情识别技术发展迅速, 传统的表情识别方法包括局部二值模式 (local binary patterns, LBP)[3]、方向梯度直方图 (histogram of oriented gradient, HOG)[4]、尺度不变特征变换 (SIFT) 等方法, 然而传统的方法受到人为规则的约束, 提取方法比较困难并且特征点提取不完全, 很难提取到人脸表情的深层特征. 随着深度学习的不断发展研究者不断尝试新的方法进行面部表情识别. 文献 [5] 提出了一种基于 DenseNet网络并结合注意力机制与多尺度特征融合, 有效学习到人脸更有效的特征, 提取更丰富的人脸表情. 文献 [6] 将轻量级的卷积神经网络和注意力模型相结合, 解决了非人脸区域的噪声干扰, 避免了模型过拟合. 文献 [7] 运用剪枝算法对 GoogLeNet 网络进行改进, 引入全局最大池化保留人脸位置信息, 在运行速度和准确率上都有很大提升. 文献 [8] 针对现如今大部分方法容易忽略面部细节特征的问题, 提出 FLF-TAWL 卷积神经网络, 该方法将局部特征与全局特征进行融合, 并且可以自适应选择与表情相关的重要区域. 文献 [9] 提出了一种两阶段训练算法 FaceNet2ExpNet, 该算法针对表情识别中的较小数据集, 联合训练前阶段和训练阶段可以提取出更加高级的人脸表情特征.

由于在线课堂的特殊性, 受网络摄像头分辨率和周围光线的影响, 学生的面部表情模糊、表情间特征变化不明显, 并且当前表情识别算法大都模型较大, 参数较多, 难以应用在实时的表情监测中. 针对以上问题本文采用改进的轻量化的卷积神经网络 Efficient-Net 实现学生的面部表情识别.

1 相关理论

1.1 EfficientNet 网络

EfficientNet[10] 是 Tan 等人在 2019 年提出的网络模型, 该网络同时考虑了网络深度、宽度以及分辨率对模型性能的影响, 采用复合模型缩放算法同时关注 3 个影响因素. 与现有分类网络相比, EfficientNet 系列网络模型参数量少, 准确率高. 混合缩放方法的具体公式如下:

$$\begin{cases} \text{depth}: \ d=\alpha^{\phi} \\ \text{width}: \ w=\beta^{\phi} \\ \text{resolution}: \ r=\gamma^{\phi} \\ \text{s.t.}\ \alpha\cdot\beta^2\cdot\gamma^2\approx 2\alpha\geqslant 1,\ \beta\geqslant 1,\ \gamma\geqslant 1 \end{cases} \tag{1}$$

其中, d、w、r分别为缩放网络的深度、宽度和分辨率, α、β、γ为缩放基数; ϕ为混合缩放因子. 首先将ϕ固定为 1, 在约束条件下进行搜索得到 EfficientNet-B0 的最佳参数为$\alpha=1.2, \beta=1.1, \gamma=1.15$.

EfficientNet-B0 由 16 个移动翻转瓶颈卷积模块 (MBConv) 组成, MBConv 包括深度可分离卷积 (DW)、批量处理归一化 (batch normalization, BN)、Swish 激活函数、SE 模块以及 Dropout 层, 其中 Dropout 层可以随机丢弃网络中的某些神经元, 有效防止了过拟合问题. MBConv 模块结构如图 1 所示.

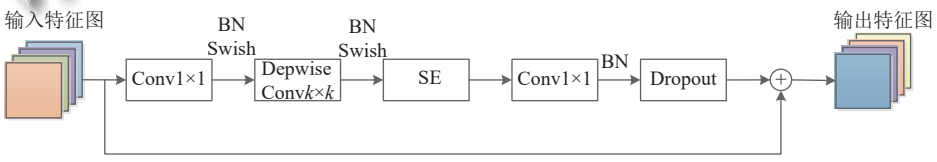


图 1 移动翻转瓶颈模块

1.2 SENet

SENet (squeeze-and-excitation networks)[11] 网络中文可以翻译为压缩激励网络, 主要是为了获得在通道维度上的关注度等级, 以下简称 SE 模块, 网络架构如图 2. 首先对输入特征图的高度和宽度使用全局平局池化 (global average pooling, GAP) 做压缩 (squeeze) 操作, 使空间的高度和宽度降维到$1\times1\times C$, 此时的通道包含全局特征信息. 然后使用两个全连接层 (FC1 和

FC2) 和 ReLU 激活函数做激励 (excitation) 操作建立通道间的联系, 压缩和激励操作如下式所示:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j),\ z \in R^C \tag{2}$$

$$S = \sigma(K_2 ReLU(K_1 z)) \tag{3}$$

其中, z_c表示全局平均池化的结果; u_c表示特征通道; C 为u的通道数; $H \times W$为u的空间维度. S 表示激励操作的结果; z 表示经过平均池化后得到的数值的结合; K_1和K_2为两个全连接层的权值矩阵; σ为 Sigmoid 函数, 将权重参数归一化到 [0, 1]. 最后再将权重参数s_c乘以特征通道u_c, 完成各个通道重要度等级的计算如式 (4):

$$x_c = Fscale(u_c, s_c) = s_c \cdot u_c \tag{4}$$

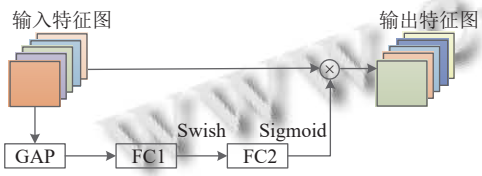


图 2 SE 模块

2 在线课堂学生表情识别模型构建

本文基于 EfficientNet-B0 构建在线课堂学生面部表情识别模型, 首先将网络摄像头捕捉到的图像进行预处理, 裁剪出图像中的人脸并调整为适合 EfficientNet 网络输入的大小格式. 使用 ImageNet 大型数据库预训练好的网络模型, 迁移学习到改进的网络模型中. 改进的 EfficientNet 网络引入多尺度特征融合模块, 可以识别更加丰富的人脸表情特征. 最后通过分类器得到学生表情分类结果并对学生的学习状态进行评估.

2.1 网络结构设计

针对在线课堂学生表情实时监测的场景, 对 EfficientNet-B0 网络结构进行改进. EfficientNet-B0 的最后一层全连接层可以降低特征位置对分类效果的影响, 但是对于在线课堂的环境来说, 人脸的特征信息大多数位于图像中间并且占据大部分像素点, 所以此时的位置信息对分类效果影响不大. 此外全连接层含有大量参数, 很容易丢失部分空间信息, 从而使得空间结构的表达不足.

为解决这一问题本文将全连接层替换为 GAP, 这样就使得输出的每个通道都对应一个类别. GAP 整合了输入图像的全局空间信息, 避免了空间信息的丢失. 相比于全连接层 GAP 参数量大幅减少, 有效避免了网络训练的过拟合问题. 另外为了提升模型的泛化能力, 增加 BN 层. 改进后的网络结构如表 1 所示.

表 1 改进 EfficientNet 网络结构

Stage	网络层	输入尺寸	输出通道数	步长	层数
1	Conv3×3	224×224	32	2×2	1
2	MBConv1, k3×3	112×112	16	1×1	1
3	MBConv6, k3×3	112×112	24	2×2	2
4	MBConv6, k5×5	56×56	40	2×2	2
5	MBConv6, k3×3	28×28	80	2×2	3
6	MBConv6, k5×5	14×14	112	1×1	3
7	MBConv6, k5×5	14×14	192	2×2	4
8	MBConv6, k3×3	7×7	320	1×1	1
9	Conv1×1, BN, GAP	7×7	7	1×1	1

2.2 浅层特征提取

在学生表情识别任务中, 浅层网络的感受野小, 感受野重叠的区域也比较小, 所以网络能捕获更多细节. 浅层特征包含更多的像素点和细粒度信息, 在有遮挡的情况下仍能准确识别学生面部表情. 本文采用 EfficientNet-B0 作为基础网络架构, 浅层网络由 EfficientNet-B0 的前两个移动翻转瓶颈卷积模块组成, 在 EfficientNet 网络中每个 MBConv 模块都含有 SE 注意力模块, SE 模块为学习到的特征信息设置权重等级, 对提取到的有用特征赋予一个较大权重, 而对无用特征赋予一个较小权重. 有效避免了由于面部遮挡对学生表情识别的影响. 因此本文在浅层特征提取部分加入 SE 模块, 将提取到的浅层细节特征重新进行权重分配, 突出人脸表情变化的显著性细节特征.

首先输入图像$I \in R^{C \times H \times W}$, C、H、W 是输入图像的通道数、高度和宽度. 经过 EfficientNet-B0 的前两个 MBConv 模块之后得到浅层特征$Fa \in R^{C_0 \times H_0 \times W_0}$, C_0、H_0和W_0分别是浅层特征的通道数、高和宽. Fa 首先通过 GAP 将特征图压缩为$1 \times 1 \times C$的特征向量, 然后使用全连接层进行通道重要度等级调整, 让网络更加关注表情变化明显的特征区域. 从而解决由于人脸姿态变化、光线原因和手部遮挡等局部遮挡而造成的人脸表情识别精度低的问题. 最后通过最大池化层保留更多的纹理细节信息, 得到浅层特征 Fac.

2.3 多尺度特征融合

人脸表情识别网络的不同深度提取的特征具有不同的特点: 浅层网络具有较小的感受野, 可以提取丰富的细节纹理信息; 深层网络感受野之间重叠的区域增加, 图像信息会进行压缩, 从而提取到的是图像的整体性信息. 因此本文基于 EfficientNet-B0 设计了一个多

尺度特征融合网络 (DS-EfficientNet), 该网络将图像的浅层特征和深层特征结合, 从而学习更加丰富的人脸表情信息.

经过特征融合模块之后的网络结构如图 3 所示. 输入图像到 EfficientNet-B0 的骨干网络中, 得到图像的深层特征 Fb. 将 Fac 和 Fb 采用双线性插值下采样算法使其具有相同的高度、宽度和深度. 然后使用 Concat 通道融合方法将下采样后的深层特征 Fb 和浅层特征 Fac 进行拼接得到融合特征 Fab, 经过特征融合模块之后的特征图拥有更加全面的特征信息.

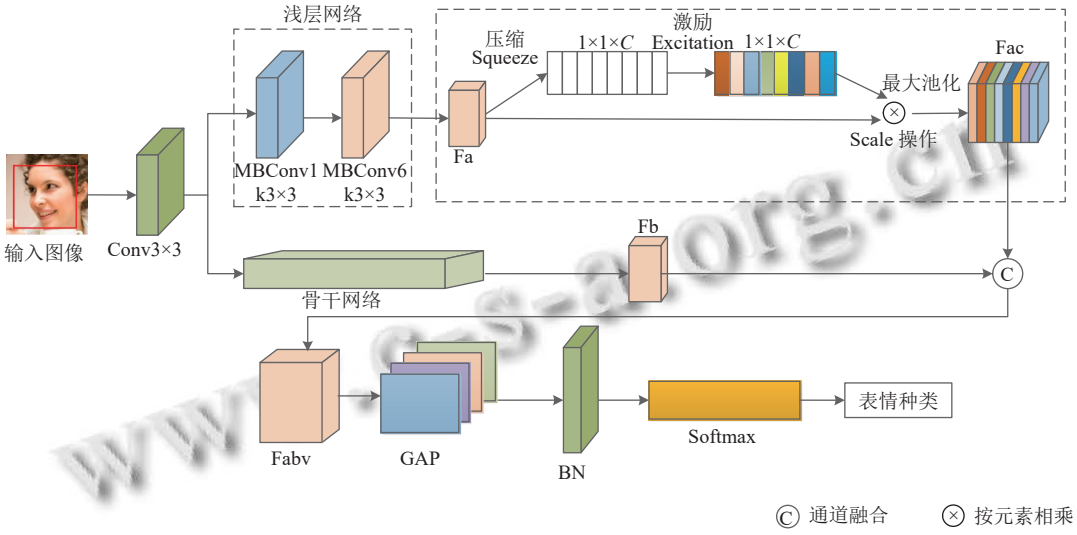


图 3 DS-EfficientNet 网络

2.4 损失函数

在输出层本文采用加权交叉熵损失 (WLoss). 传统的交叉熵损失函数在数据集类别均衡的情况下可以达到很好的分类效果, 但是本文所使用的 Fer2013 数据集高兴的样本有 8 989 张图片而厌恶只有 547 张, 存在严重的类别不平衡问题. 并且 Fer2013 数据集中存在严重的噪声问题, 有很多类别标注不正确的样本, 所以使用传统的交叉熵损失函数就会对人脸表情识别的准确性产生很大影响. 加权交叉熵损失函数就是对类别设置不同的关注度等级, 使模型在训练过程中更加关注小样本类别的学习. 损失函数公式为:

$$WLoss = -\frac{1}{N}\sum_{n=1}^{N} W_n \log(P_{n,i}) \tag{5}$$

其中, W_n为权重系数, $P_{n,i}$表示第 n 个样本类别为 i 的概率, N 表示样本类别数量.

3 实验结果及分析

3.1 数据集及预处理

对于在线课堂学生面部表情识别本文在 Fer2013 和 CK+数据集上进行实验. Fer2013 是 2013 年在 ICML[12] 上提出的人脸表情数据集, 包括 7 种基本表情共 35 886 张人脸表情图像. 数据集分布见表 2.

由于 Fer2013 数据集都来自网络爬取图片, 数据集中的人脸在面部姿态、人脸角度以及年龄之间都存在很大差异, 与真实场景中的人脸状态非常接近很适用于真实环境中的人脸表情检测. 数据集示例如图 4.

表 2 Fer2013 数据集分布

类别	样本数量	训练集	公开验证集	私有验证集
生气	4 953	3 995	467	491
厌恶	547	436	56	55
恐惧	5 121	4 097	496	528
高兴	8 989	7 215	895	879
悲伤	6 077	4 830	653	594
惊讶	4 002	3 171	415	416
中性	6 198	4 965	607	626

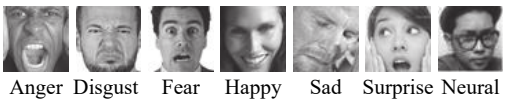


图 4 Fer2013 数据集示例

CK+数据集[13] 是 2010 年在实验室环境下采集得到的, 数据集由 123 名测试者的 593 个视频序列组成, 从其中选择表情表达比较明显的图像进行实验, 共 981 张图像, 包括 7 种基本面部表情类别, 数据集示例如图 5 所示.

图 5 CK+数据集示例

为了提升模型的泛化能力, 增强模型的鲁棒性, 对数据集进行数据增强. 对于 Fer2013 数据集, 将训练集随机裁剪和水平翻转, 将私有验证集和公开验证集图像的中心点及 4 个角进行裁剪来实现数据集的扩充, 并最终在私有验证集上测试实验效果.

对于 CK+数据集, 数据集中不仅仅只包括人脸. 为了减少身体其他部分对分类效果的影响, 首先使用 OpenCV 对原始图片进行人脸检测并裁剪出人脸图像, 然后使用双线性插值算法进行尺度归一化将人脸图像缩放为 224×224 像素大小. 由于 CK+数据集所含图像太少, 对图像进行水平翻转和随机裁剪进行数据扩充. 同时为了保证模型在遮挡条件下的识别效果, 在数据集中通过黑色框遮挡眼睛、嘴巴等部位来模拟在线课堂状态下存在的面部遮挡问题, 如由眼镜、口罩、手托腮等引起的遮挡. 处理方式如图 6 所示.

图 6 遮挡方式

3.2 实验环境与评价指标

本实验采用的操作系统为 Windows 10, 硬件环境为 Intel-CPU-i5-10400F, GPU 为 8 GB 的 NVIDIA Ge Force GTX1070Ti; 软件环境 Python 3.7, 深度学习框架 PyTorch 1.7.1. 为进一步提升模型性能, 训练时采用迁移学习和数据增强方法. 对模型进行 200 个 epoch 的训练, batch_size 设置为 16, 使用加权交叉熵损失, Adam 优化器. 初始学习率设置为 0.01, 设置学习率按照指数形式衰减动态调整学习率大小.

实验使用识别准确率 (accuracy)、模型参数量 (parameter)、混淆矩阵及 *F*1-score 作为评价指标. 准确率指正确分类的样本占总样本的比例. 模型参数量指训练时需要学习的参数个数, 参数量越大说明模型越大. 混淆矩阵指的是由真阳性 (TP)、假阴性 (FN)、假阳性 (FP)、真阴性 (TN) 呈现出的表格. *F*1-score 指精准率 (*precision*) 和召回率 (*recall*) 的调和平均数, 取值范围 0–1, 计算公式为:

$$F1\text{-score} = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

3.3 模型性能结果分析

3.3.1 消融实验

本文针对在线课堂学生表情识别问题, 以 Efficient-Net-B0 为基础, 提出 DS-EfficientNet 网络模型. 为了验证本文提出的各个模块的有效性, 采用以下几种消融实验方案进行实验, (1) 仅使用 EfficientNet网络; (2) 将浅层网络和深层网络拼接进行多尺度融合; (3) 在浅层网络中加入 SE 模块; (4) 将全连接层替换为全局平均池化层 (GAP). 评价指标采用准确率和参数量, 实验结果见表 3.

表 3 表情识别消融实验

序号	多尺度融合	SE模块	GAP	Fer2013 (%)	CK+ (%)	参数量 (parameter)
1	×	×	×	69.83	94.56	1 695 720
2	√	×	×	71.64	96.83	2 036 425
3	√	√	×	72.89	97.77	4 373 248
4	√	√	√	73.47	98.84	1 749 300

由表 3 可以看出, EfficientNet-B0 基础网络在 Fer2013 上的准确率为 69.83%, 在 CK+数据集上的准确率为 94.56%. 当增加多尺度融合模块之后准确率达到了 71.64% 和 96.83%, 相比于基础网络分别提升了 1.81% 和 2.27%, 这表明多尺度融合模块可以将表情粗粒度信息和细粒度信息结合起来进而提升网络性能. 在浅层网络加入 SE 模块之后模型准确率分别达到了 72.89% 和 97.77%, 相较于不加 SE 模块性能分别提升了 1.25% 和 0.94%. 将最后的全连接层替换为全局平均池化和批归一化层之后网络识别率分别提升了 0.58% 和 1.07%, 并且参数量大大减少.

3.3.2 Fer2013 数据集中实验结果及分析

为了说明训练过程, 以 Fer2013 数据集为例, 改进的网络模型在 Fer2013 上训练的损失 (loss) 和精度 (accuracy) 曲线如图 7、图 8 所示.

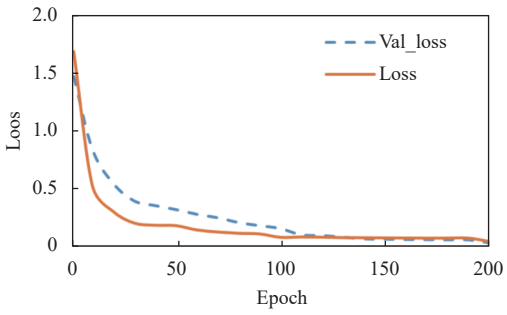


图 7 Fer2013 训练损失曲线

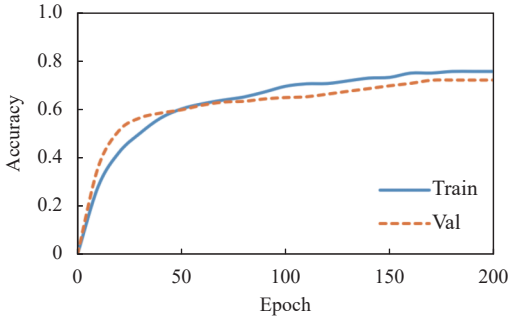


图 8 Fer2013 训练精度曲线

由图 7、图 8 可知改进后的 EfficientNet 网络在训练 200 个 epoch 之后训练集和验证集的损失降到最低, 此时的准确率达到 74.53%.

在 Fer2013 数据集上进行实验时, 为了更好地验证改进后网络的优势, 最终在测试集上验证网络性能, 得到的混淆矩阵如图 9 所示.

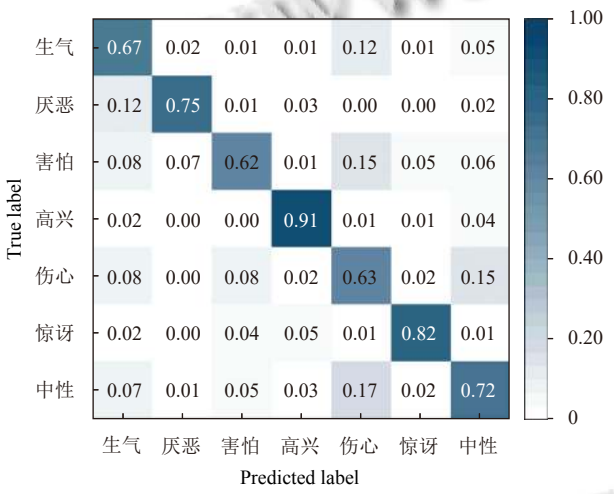


图 9 Fer2013 识别结果混淆矩阵

由混淆矩阵可以看出, 本文所提出的网络模型在高兴和惊讶这两类的识别率分别达到了 91% 和 82%, 而生气、害怕和伤心这 3 类的准确率相对较低. 这是因为对于高兴和惊讶来说, 面部纹理特征更加突出, 所以更容易识别. 害怕表情特别容易和惊讶表情混淆, 因为害怕表情也会出现眼睛睁大、嘴巴张开的情况, 但是害怕时嘴巴张开的幅度没有惊讶时张开幅度大. 此外害怕和悲伤都有皱眉、额头紧皱等相似特征, 害怕类的识别难度最大. 生气、害怕和伤心这 3 类表情都属于消极情绪, 表情之间具有很强的相似性, 从而出现识别率低的情况.

本文将改进的 EfficientNet 与其他算法进行对比分析如表 4 所示, 实验发现本文提出的网络模型识别准确率达到了 73.47%, 与其他算法相比准确率和 *F*1-score 值都取得了较好的效果, 并且参数量只有少量增加.

表 4 Fer2013 数据集实验对比

方法	准确率 (%)	参数量	*F*1-score
VGG16[14]	70.12	139 357 544	0.73
ResNet50[15]	71.15	25 502 912	0.75
Xception[16]	66.80	16 691 895	0.70
Inceptionv3[17]	70.13	23 614 078	0.72
MobileNetv2[18]	72.28	3 447 520	0.74
EfficientNet[10]	71.56	1 695 720	0.73
本文方法	73.47	1 749 300	0.76

3.3.3 CK+数据集中实验结果及分析

改进后的网络在 CK+数据集上的训练损失 (loss) 和精度 (accuracy) 曲线如图 10、图 11 所示.

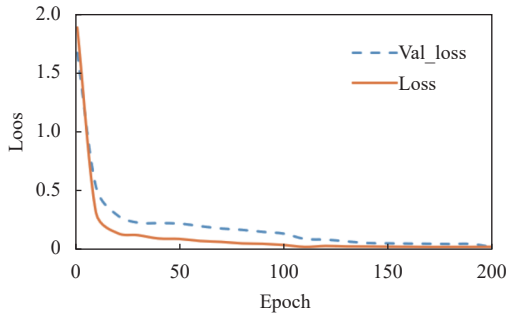


图 10 CK+训练损失曲线

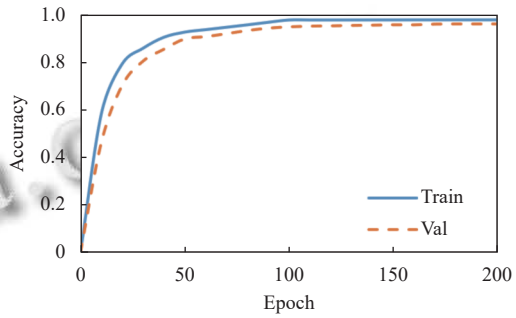


图 11 CK+训练精度曲线

CK+数据集是在实验室环境下, 人脸无遮挡并且没有光线和环境因素的噪声干扰, 所以准确率相比于 Fer2013 高很多. 由于 CK+数据集图片数量较 Fer2013 少很多, 采用十折交叉验证. 按照训练集和测试集 9:1 的比例划分数据集, 其测试集混淆矩阵如图 12 所示. 在 CK+数据集中, 各类表情的准确率相较于 Fer2013 都提高了很多. 由混淆矩阵可以看出高兴、惊讶和生气有较高的识别率, 而愤怒、害怕、伤心和中性识别率相对较低. 这是由于这几类表情之间有相似的特征容易混淆, 并且这几类表情训练样本相较于高兴、惊

讶和生气样本数量较少, 导致训练不充分, 从而识别准确率相对较低.

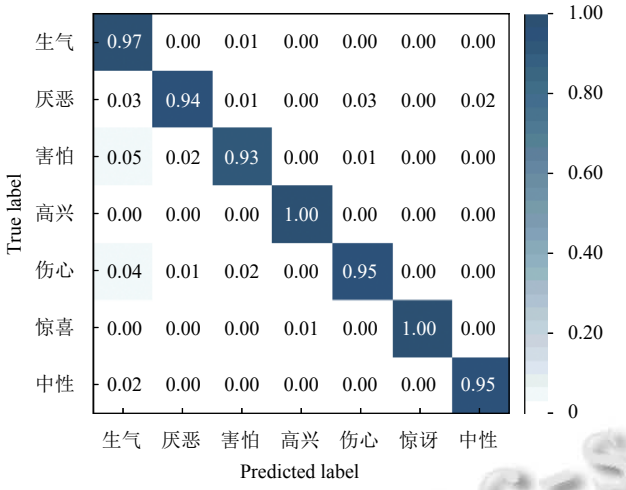


图 12 CK+识别结果混淆矩阵

表 5 是不同方法在 CK+数据集上的对比, 本文所提方法准确率达到了 98.84%, 与其他算法相比本文算法准确率和 *F*1-score 值更高, 参数量只有少量增加.

表 5 CK+数据集实验对比

方法	准确率 (%)	参数量	*F*1-score
VGG16[14]	95.24	139 357 544	0.96
ResNet50[15]	94.13	25 502 912	0.98
Xception[16]	97.48	16 691 895	0.96
Inceptionv3[17]	94.02	23 614 078	0.97
MobileNetv2[18]	92.43	3 447 520	0.98
EfficientNet[10]	96.75	1 695 720	0.97
本文方法	98.84	1 749 300	0.98

3.4 分类效果对比

对改进前后的网络分类性能对比如图 13 所示. 相比于改进前的网络, 改进后的网络在光线昏暗和在有遮挡的情况下仍能实现更好的分类效果.

3.5 在线课堂学生表情状态分析

学生在线上课的表情状态可分为积极情绪、消极情绪以及中性情绪, 如图 14 所示.

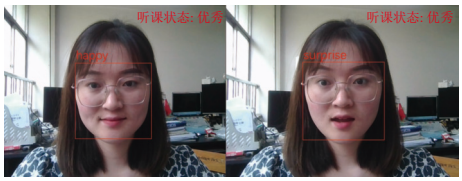
当学生出现积极情绪 (开心和惊讶) 时, 表明学生对知识处于一个乐意接受的状态, 在上课过程中学生学习情绪比较高涨, 此时将学生的上课状态判定为优秀. 当学生课堂中出现消极情绪 (生气、害怕、伤心和厌恶) 时, 表明学生上课时注意力并没有在课堂上或者学生对授课内容反感, 此时将学生的上课状态判定为较差. 当学生出现中性表情时, 表明学生上课过程中听课状态一般, 此时可将学生听课状态判定为良好. 本文所提算法预测单幅图像所需的时间为 39.276 ms, 满足实时性需求.

图 13 原始网络和改进后网络分类效果对比

4 结论

本文针对人脸表情识别中常见的表情细节信息容易被忽略的问题, 提出一种多尺度融合算法. 该算法首先通过浅层网络和注意力模块加强面部表情的细节信息的表达, 同时将图像送入 EfficientNet-B0 的骨干网络提取人脸深层信息. 最后采用特征融合模块将深层特征和浅层特征进行信息融合, 融合之后的网络可以提取人脸更加丰富的表情信息. 为了减少网络模型的参数量, 增强模型泛化性能, 将最后的全连接层替换为全局平均池化和批归一化层, 有效防止了网络模型的

过拟合问题. 最终在 Fer2013 和 CK+数据集上进行实验, 结果表明本文所提出算法识别精度高于其他先进方法. 此外仅通过人脸表情对学生专注度进行分析不够全面, 在后续研究中, 将考虑采用学生的脸部姿态和面部表情相结合的方式进行专注度分析, 并且使用更加接近在线课堂环境下的数据集进行研究, 提高在网络课堂中的应用价值.



(a) 积极情绪

(b) 消极情绪

(c) 中性情绪

图 14 在线课堂表情状态

参考文献

1 郦泽坤, 苏航, 陈美月, 等. 支持 MOOC 课程的动态表情识别算法. 小型微型计算机系统, 2017, 38(9): 2096–2100. [doi: 10.3969/j.issn.1000-1220.2017.09.033]

2 Ekman P, Friesen WV. Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 1971, 17(2): 124–129. [doi: 10.1037/h0030377]

3 Ahmed F, Bari H, Hossain E. Person-independent facial expression recognition based on compound local binary pattern (CLBP). The International Arab Journal of Information Technology, 2014, 11(2): 195–203.

4 Dalal N, Triggs B. Histograms of oriented gradients for human detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005: 886–893.

5 张鹏, 孔韦韦, 滕金保. 基于多尺度特征注意力机制的人脸表情识别. 计算机工程与应用, 2022, 58(1): 182–189. [doi: 10.3778/j.issn.1002-8331.2106-0174]

6 褚晶辉, 汤文豪, 张姗, 等. 一种基于注意力模型的面部表情识别算法. 激光与光电子学进展, 2020, 57(12): 121015.

7 张宏丽, 白翔宇. 利用优化剪枝 GoogLeNet 的人脸表情识别方法. 计算机工程与应用, 2021, 57(19): 179–188. [doi: 10.3778/j.issn.1002-8331.2102-0296]

8 郑剑, 郑炽, 刘豪, 等. 融合局部特征与两阶段注意力权重学习的面部表情识别. 计算机应用研究, 2022, 39(3): 889–894, 918.

9 Ding H, Zhou SK, Chellappa R. FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition. Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition. Washington: IEEE, 2017. 118–126.

10 Tan MX, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 6105–6114.

11 Hu J, Shen L, Albanie S, *et al*. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023. [doi: 10.1109/TPAMI.2019.2913372]

12 Goodfellow IJ, Erhan D, Carrier PL, *et al*. Challenges in representation learning: A report on three machine learning contests. Neural Networks, 2015, 64: 59–63. [doi: 10.1016/j.neunet.2014.09.005]

13 Lucey P, Cohn JF, Kanade T, *et al*. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. San Francisco: IEEE, 2010. 94–101.

14 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego: Computational and Biological Learning Society. 2014, 111(1): 98–136.

15 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. IEEE Conference on Computer Vision & Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

16 Chollet F. Xception: Deep learning with depthwise separable convolutions. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1800–1807.

17 Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2818–2826.

18 Sandler M, Howard A, Zhu ML, *et al*. MobileNetv2: Inverted residuals and linear bottlenecks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 4510–4520.

(校对责编: 牛欣悦)