doi:10.19562/j.chinasae.qcgc.2024.09.017

基于注意力融合特征增强的座舱表情识别模型*

罗玉涛[1,2],郭丰瑞[1,2]
(1. 华南理工大学机械与汽车工程学院,广州 510640; 2. 广东省汽车工程重点实验室,广州 510640)

[摘要] 针对智能座舱驾驶员表情识别深度学习模型准确率和实时性难以兼顾的问题,提出一种基于注意力融合与特征增强网络的表情识别模型EmotionNet。模型以GhostNet为基础,在特征提取模块内利用两个检测分支融合坐标注意力和通道注意力机制,实现注意力机制互补与对重要特征的全方位关注;建立特征增强颈部网络以融合不同尺度特征信息;最终通过头部网络实现不同尺度特征信息决策级融合。在训练中则引入迁移学习思想和中心损失函数以进一步提高模型的识别准确性。在RAF-DB和KMU-FED数据集实验中,模型分别取得85.23%和99.95%识别准确率,并达到59.89 FPS的识别速度。EmotionNet平衡了识别准确率和实时性,达到了较为先进的水平并具备一定的智能座舱表情识别任务的适用性。

关键词:智能座舱;表情识别;注意力机制;特征增强网络

Cockpit Facial Expression Recognition Model Based on Attention Fusion and Feature Enhancement Network

Luo Yutao[1,2] & Guo Fengrui[1,2]
1. *School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou* 510640;
2. *Guangdong Provincial Key Laboratory of Automotive Engineering, Guangzhou* 510640

[**Abstract**] For the problem of difficulty in balancing accuracy and real-time performance of deep learning models for intelligent cockpit driver expression recognition, an expression recognition model called EmotionNet based on attention fusion and feature enhancement network is proposed. Based on GhostNet, the model utilizes two detection branches within the feature extraction module to fuse coordinate attention and channel attention mechanisms to realize complementary attention mechanisms and all-round attention to important features. A feature enhanced neck network is established to fuse feature information of different scales. Finally, decision level fusion of feature information at different scales is achieved through the head network. In training, transfer learning and central loss function are introduced to improve the recognition accuracy of the model. In the embedded device testing experiments on the RAF-DB and KMU-FED datasets, the model achieves the recognition accuracy of 85.23% and 99.95%, respectively, with a recognition speed of 59.89 FPS. EmotionNet balances recognition accuracy and real-time performance, achieving a relatively advanced level and possessing certain applicability for intelligent cockpit expression recognition tasks.

Keywords: intelligent cockpit; expression recognition; attention mechanisms; feature enhancement network

* 工信部制造业高质量发展专项资金项目(R-ZH-023-QT-001-20221009-001)和广州市科技计划项目(2023B01J0016)资助。
原稿收到日期为2024年02月26日,修改稿收到日期为2024年04月20日。
通信作者:罗玉涛,教授,博士,E-mail:ctytluo@scut.edu.cn。

前言

研究表明,驾驶员在驾驶时的情绪对驾驶安全有着重要影响。比如,对于离散的基本情绪,愤怒、恐惧、伤心、惊讶与厌恶这几种情绪下的高风险驾驶比例较大;中性与高兴情绪则表现出较低的高风险驾驶比例[1]。而人的情绪,往往通过脸部的表情体现出来,故研究智能座舱系统驾驶员表情识别,对于判读并及时调节驾驶员情绪,从而降低事故发生风险有着重要意义。

目前,针对情绪识别的方法主要包括基于生理信号(如脑电图和心电图等)、基于语音信号和基于视觉识别人脸表情3种方法。其中,前两者分别具有须随时佩戴检测设备和识别准确率低的问题[2-3],对于智能座舱实用性有限。而基于识别人脸表情的方法又可以分为基于传统机器视觉和基于深度学习两类,其中深度学习方法无须手动设计特征且在准确性上更具优势[4],成为目前情绪识别方法中的主流。

Nan等[5]以MobileNetV1模型为基础,加入CBAM模块(convolutional block attention module)增强面部表情的局部特征提取并结合中心损失(Center loss)函数以增大类间距离。该方法准确率较基线网络提高了2.87%,但也引入了参数量增大的问题。Xiao等[3]基于迁移学习思想引入在表情识别数据集CK+、FER预训练权重,并设计了基于增强的重采样模块;文献[3]中的模型在其自建道路驾驶数据集上达到了96.4%的准确率,但其主干网络计算量较大,为后续智能座舱部署带来挑战。梁艳等[6]注意到脸部局部器官特征等对于表情的重要作用,采用残差神经网络提取人脸表情的全局特征并与眼睛、嘴巴局部特征融合。该方法在多个数据集上均取得了较高准确率,但同样存在模型复杂度高的问题。Minaee等[7]则聚焦于提高轻量化网络性能,提出一种空间变换模块以聚焦重要面部特征,该模块提高了轻量化模型特征提取能力,但由于骨干网络过于简单,在准确性方面仍有一定提升空间。

当前,智能座舱算法部署的域控制器为Arm架构嵌入式芯片,其算力及存储空间均受到限制,且智能座舱的多功能要求须同时实现多模型部署,这势必导致系统整体运行减缓,因此智能座舱系统对表情识别模型的实时和高效性提出了更高的要求。而目前的研究多聚焦于准确性与实时性的某一方面,较少关注模型性能的平衡。

为在兼顾实时性的同时提高表情识别模型的准确性,针对智能座舱模型部署的需求以及现有方法的不足,本文提出了一种基于注意力融合与特征增强网络的智能座舱表情识别模型EmotionNet。具体地,基于GhostNet[8]设计了一种利用二分支结构融合通道注意力与坐标注意力的特征提取模块并以之构建主干网络;在主干网络后面构建特征增强颈部网络以融合不同尺寸特征图;通过包含双分类头的头部网络实现不同尺寸特征图决策级融合;在训练过程中引入中心损失Center loss[9]和迁移学习方法进一步提高模型的识别准确率。

1 EmotionNet模型架构

EmotionNet模型整体架构如图1所示,包括主干网络(Backbone)、颈部网络(Neck)和头部网络(Head)3个部分。其中,主干网络负责对输入图片进行特征提取,并输出8倍、16倍下采样特征图到颈部网络;颈部网络则对两种尺度特征图进行特征融合及增强;头部网络则基于颈部结果得到两种特征图相应的表情识别结果概率分布,并加权融合得到最终识别结果。图1中各结构输出特征图大小格式为高×宽×通道数;k、s、p分别代表卷积单元或块的卷积核大小、卷积步长以及特征图边缘填充大小;Concat为特征图拼接操作,作用于通道维度;UpSample为上采样操作,即通过线性差值方法在宽、高维度扩展特征图大小,以便进行特征图拼接。

模型中,为增强泛化能力、缓解过拟合并避免深层神经网络带来的梯度爆炸问题,模型卷积单元由卷积层(convolutional layer,Conv2D)或深度卷积层(depthwise convolution)加批量标准化层(batch normalization layer,BN)加非线性激活函数ReLU组成。在本文图片中,用Conv代指普通卷积单元,用DWConv代指深度卷积单元。

1.1 主干网络

在CNN中,注意力机制能够使CNN聚焦于对当前任务更为关键的信息,从而更加高效地分配有限的计算资源。目前主要的注意力机制包括通道注意力、空间注意力、坐标注意力3类。它们关注的特征信息如表1所示。

本文主要采用通道注意力机制压缩激励模块(squeeze-and-excitation module, SE)[10]和坐标注意

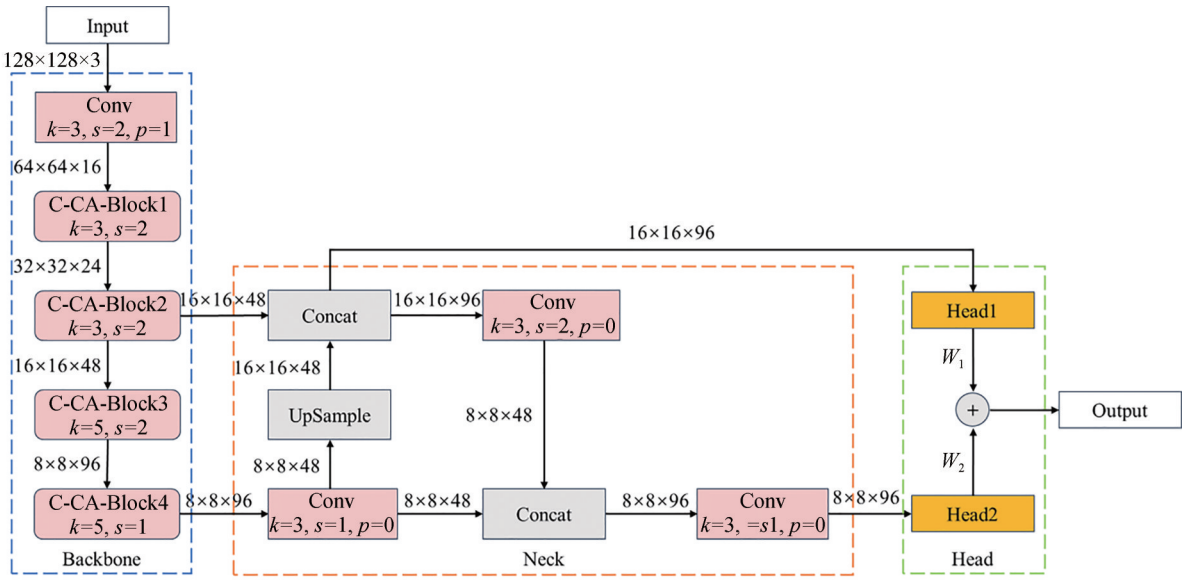


图1　EmotionNet模型整体架构

表1　不同注意力机制关注的特征信息

注意力机制	关注特征信息
通道注意力	重要特征所在通道
空间注意力	重要特征所在位置
坐标注意力	重要特征所在通道及位置

力机制模块(coordinate attention, CA)[11]。为充分利用这两种注意力机制聚焦特征的差异性,本文在GhostNet基础特征提取模块G-bneck模块的基础上提出了如图2所示的C-CA-Block模块,图中m根据模型设计的需要设定。

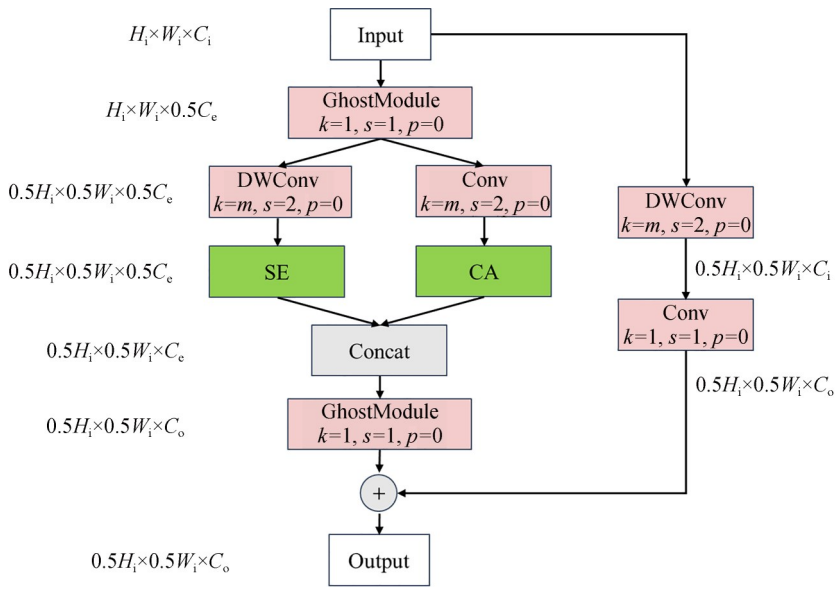


图2　C-CA-Block结构

图中H_i、W_i分别代表输入特征图的高和宽,C_i、C_o、C_e分别代表输入和输出特征图的通道数及模块倒残差结构扩展的通道数。

C-CA-Block整体可以分为主要分支和辅助分支两条支路。对于主要分支,输入首先通过1×1卷积单元GhostModule以较低的计算成本扩展通道数至C_e的一半;随后进入两个特征提取分支,第一个分支沿用G-bneck模块原始结构,采用深度卷积单元提取特征并通过SE模块强调重点特征通道;第二个分支则采用普通卷积单元提取特征,以增强与第一个分支提取特征的差异性,并通过CA模块兼顾重点特征的位置;两个分支的特征提取结果在通道维度

以拼接的方式融合,最后通过1×1的GhostModule降低通道维度至预定大小C_o。

辅助分支保留G-bneck原始结构,其作用为通过1×1卷积单元调整原始输入特征图通道数后与主要分支输出特征图相加,此时若C-CA-Block进行了下采样(卷积步长大于等于2)则须利用步长为2的深度卷积对输入特征图进行下采样。

与GhostNet相比,模型主干网络将卷积模块从16个精简为4个以提高计算速度,并将原有的G-bneck模块替换为C-CA-Block以提高其检测性能。4个C-CA-Block模块参数设定如表2所示。表中卷积核大小和步长均代指C-CA-Block中特征提取卷积单元的相关参数。

表2 主干网络C-CA-Block参数设定

模块	卷积核大小	步长	输入/扩展/输出通道
C-CA-Block1	3×3	2	16/80/24
C-CA-Block2	3×3	2	24/144/48
C-CA-Block3	5×5	2	48/576/96
C-CA-Block4	5×5	1	96/576/96

下面将具体介绍C-CA-Block模块中的各个单元。

1.1.1 GhostModule

在普通卷积中,卷积操作须消耗大量的计算成本。比如假设输入特征图大小为$H_i \times W_i \times C_i$,输出特征图大小为$H_o \times W_o \times C_o$,卷积核大小为$C_i \times k \times k \times C_o$,则不考虑偏移的情况下(下同)本次卷积计算量$F$为

$$F = H_o \cdot W_o \cdot C_o \cdot C_i \cdot k \cdot k \tag{1}$$

另一方面,卷积模块中产生的大量中间特征图(储存在通道维度)中存在一定数量的相似特征图对,这便导致了特征图存在冗余和计算资源浪费。因此如图3所示,GhostModule首先通过普通卷积生成部分固有特征图,随后通过较低计算的成本线性运算(深度卷积)以固有特征图为输入生成与之相似的冗余特征图,最终通过拼接操作得到卷积结果,图中m、n根据模型设计的需要设定。

在GhostModule计算中,假设第一步生成了x个固有特征图,输出特征图通道数为C_o,输出特征图数量与固有特征图数量之比为σ,则为使固有特征图数量与冗余特征图数量之和等于C_o,$x=C_o/\sigma$,且每个固有特征图须进行$\sigma-1$次线性运算,设线性运算卷积核大小为$d \times d$,则普通卷积与GhostModule计算量之比r为

$$r = \frac{F}{F_1 + F_2} \approx \sigma \tag{2}$$

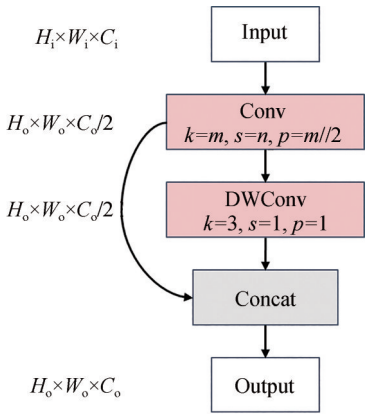


图3 GhostModule结构

式中:F_1为通过普通卷积生成固有特征图所需的计算量;F_2为通过线性运算生成冗余特征图所需的计算量。F_1、F_2的计算如式(3)和式(4)所示;式(2)中的约等号当且仅当$\sigma << C_i$时成立。

$$F_1 = \frac{C_o}{\sigma} \cdot H_o \cdot W_o \cdot C_i \cdot k \cdot k \tag{3}$$

$$F_2 = (\sigma - 1) \cdot \frac{C_o}{\sigma} \cdot H_o \cdot W_o \cdot d \cdot d \tag{4}$$

可见在理想情况下,GhostModule可以将计算量降低到普通卷积的$1/\sigma$,在EmotionNet中,σ取2。

1.1.2 SE模块

SE模块为一种轻量级通道注意力机制模块,通道注意力认为CNN中特征图各通道间的重要性不同,而通道注意力则通过构建可训练的权值矩阵,凸出重要特征通道的比重。

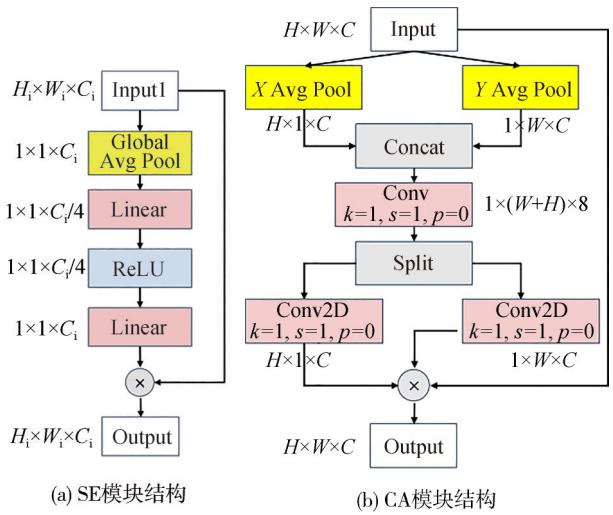
SE模块的结构如图4(a)所示,该模块包含"压缩"和"激励"两个步骤。其中,"压缩"步骤通过全局平均池化压缩特征图宽、高维度信息,"激励"步骤则通过两个全连接层学习各个通道权重,最终将权重与输入特征图相乘,实现对于重要特征通道的关注。

1.1.3 CA模块

CA模块在关注重要特征通道的同时兼顾了特征图的空间信息,即重要像素点在宽、高维度上的位置。CA模块的结构如图4(b)所示,该模块首先在宽、高维度上分别进行全局池化压缩编码以保留高、宽方向上的位置信息;在拼接一对特征图后通过卷积模块压缩通道维度;随后将二者分开分别还原到原本通道数从而得到宽、高方向的权重;最终将权重与输入特征图相乘,实现坐标注意力。

1.2 颈部网络

在主干网络特征提取的过程中,不同深度特征图蕴含特征信息不尽相同,比如浅层网络提取到的



(a) SE模块结构　　(b) CA模块结构

图4　EmotionNet注意力机制模块

大尺度特征图通常包括大量的图像细节信息,而深层网络提取到的小尺度特征图则包括了更高层次的语义信息。

因此为充分利用不同尺度特征图蕴含的特征信息,本文借鉴了特征金字塔(feature pyramid network, FPN)[12]以及路径聚合网络(path aggregation network, PAN)[13]的思想,构建了如图5所示的特征增强颈部网络。该网络以8倍和16倍下采样特征图为输入(大小分别为16×16×48和8×8×96),并通过特征图上采样、拼接、特征提取等操作实现对不同尺度特征图的充分融合。

对于16倍下采样特征图B5,颈部网络通过卷积

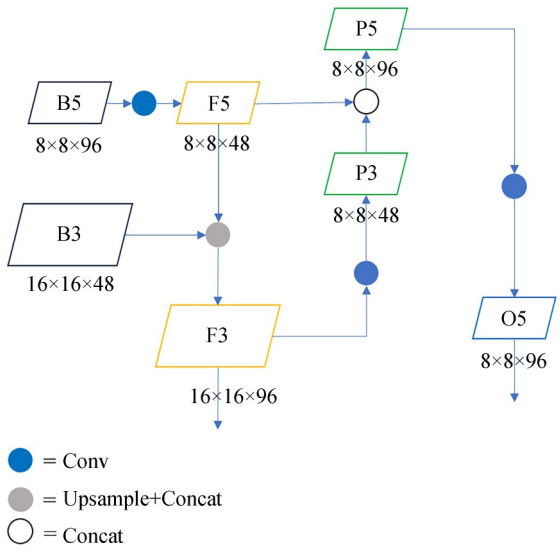


图5　特征增强网络结构

单元压缩其通道数,得到特征图F5;随后通过上采样扩展图片大小并与B3拼接得到16×16×96特征图F3;对F3通过步长为2的卷积操作进行下采样并压缩通道数得到P3;P3随后与F5拼接得到P5并通过卷积单元提取特征得到8×8×96特征图O5;最后将F3、O5输出到头部网络。本文颈部特征增强网络卷积单元采用3×3卷积,从而在实现通道压缩和下采样的同时进一步提取特征。

1.3　头部网络

为进一步利用颈部网络输出的不同尺度特征图蕴含的特征信息,头部网络采用决策级融合方法,由两个结构相同的分类头组成,分别以特征增强网络的F3、O5特征图为输入,单独计算其所属表情的概率分布,并最终通过可训练权重加权相加得到最终结果。

头部网络中的一个分类头结构如图6所示,首先通过步长为1的1×1 GhostModule扩展特征图通道数,随后通过全局平均池化压缩高宽维度,之后通过全连接层调整通道数以得到1×7维特征向量,该向量即代表输入图片对应该表情的概率。

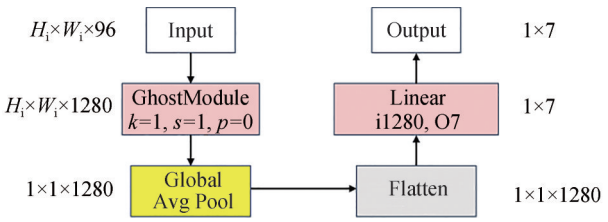


图6　头部网络分类头结构

2　训练过程优化

在驾驶员表情识别中,由于每个人个体间情感表达的习惯差异以及情绪程度的不同(如微愠和狂怒的不同),同一人在表达相同情绪时,其表情可能不尽相同,如图7(a)所示;在表达不同情绪时,其表情也有可能存在相似性,如图7(b)所示;而对于不同个体而言,在表达同一情绪时,其表情亦存在差异性,如图7(c)所示。

在表情识别任务中,上述现象的发生导致数据集类内样本差异增大而类与类之间的样本差异减小,提高了模型区分不同表情类别的难度。而在智能座舱实践中,增加网络深度以增强特征辨析能力则须付出高昂的计算成本代价。针对这一问题,本

(a) 同一人相同情绪(愤怒)的差异表情表达

(b) 同一人不同情绪(左愤怒右厌恶)的相似表情表达

(c) 不同人相同情绪(伤心)的差异表情表达

图7 情绪表情表达的个体差异

文从优化训练过程着手,引入中心损失函数与交叉熵损失函数,以优化数据集类内样本差异,并通过迁移学习方法建立大型表情识别数据集以预训练从而扩大样本范围,以便模型更好地区分类内外差异并降低过拟合。

2.1 中心损失

中心损失是一种聚类算法。该算法在训练过程中对每个类学习一个中心,并降低类中成员到中心的距离,从而改善样本分布情况。其计算公式为

$$L_c = \frac{1}{2}\sum_{i=1}^{m}\left\| \boldsymbol{x}_i - \boldsymbol{c}_{yi} \right\|_2^2 \tag{5}$$

式中:m为训练时样本批次大小;$\boldsymbol{x}_i$代表批次中第i个样本对应的模型输出的特征向量;y_i为第i个样本对应类别;$\boldsymbol{c}_{yi}$代表第i个样本所对应的类别y_i的中心向量。

L_c对于$\boldsymbol{x}_i$的梯度如式(6)所示:

$$\frac{\partial L_c}{\partial \boldsymbol{x}_i} = \boldsymbol{x}_i - \boldsymbol{c}_{yi} \tag{6}$$

在中心损失中,各个类别的中心须随着训练过程而不断更新,对于类别j中心迭代变化率$\Delta \boldsymbol{c}_j$如式(7)所示:

$$\Delta \boldsymbol{c}_j = \frac{\sum_{i=1}^{m}\delta\left(y_i = j\right)\cdot\left(\boldsymbol{c}_j - \boldsymbol{x}_i\right)}{1 + \sum_{i=1}^{m}\delta\left(y_i = j\right)} \tag{7}$$

式中:$\boldsymbol{c}_j$为类别j的中心向量;$\delta(y_i = j)$表示当$y_i = j$,即第i个样本所属类别为j时,该函数取1,反之为0。

在实际使用中,一般将中心损失与其他损失函数搭配使用。在本文中,则与Softmax损失函数L_s共用,并通过超参数λ确定其权重:

$$L = L_s + \lambda L_c \tag{8}$$

$$L_s = -\sum_{i=1}^{m}\log\frac{\exp\left(\boldsymbol{W}_{yi}^{\mathrm{T}}\boldsymbol{x}_i + \boldsymbol{b}_{yi}\right)}{\sum_{j=1}^{n}\exp\left(\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{x}_i + \boldsymbol{b}_j\right)} \tag{9}$$

式中:$\boldsymbol{W}_{yi}$和$\boldsymbol{W}_j$分别代表模型全连接层第yi和第j项的权重矩阵;$\boldsymbol{b}_{yi}$和$\boldsymbol{b}_j$分别代表模型全连接层第yi和第j项的偏置向量;权重和偏置均通过模型训练确定。

参照文献[9],在本文中λ取1以获得较强的聚类效果。

2.2 模型预训练

由于汽车座舱情境下的人脸表情识别数据集较少且数据量和涵盖性别、年龄、种族范围有限,不能充分地反映表情类内外的差异,直接在座舱数据集训练不利于模型泛化。由于汽车座舱场景表情识别与野外、实验室等其他场景表情识别任务有一定相似性,本文借鉴迁移学习思想,通过融合其他情景下的表情识别数据集建立大型预训练数据集,并通过预训练得到预训练模型权重,最终在目标数据集上训练时加载预训练权重,从而变相扩展目标数据集数据量,达到增强模型区分不同表情能力的目的。

预训练数据集所融合其他场景数据集及其样本分布如表3所示。

从表3可见,将所有数据集简单混合后仍存在一定的样本不均衡问题,故本文从各个类别中仅随机保留10 000幅图片(不足则全部保留),得到如表4所示的融合数据集,共67 743幅。由于不同数据集人脸在图片里占据面积大小不同,在预训练之前采用训练好的YOLOv5l人脸检测模型依次检测数据集内图片,该模型能够基于输入图片得到人脸所在区域,从而去除图片背景并统一人脸占据数据集图片面积的比例。预训练按照9:1的比例划分训练集与验证集。

模型预训练环境在Ubuntu18.04操作系统上搭建,Python版本为3.8,CUDA版本为11.0,Pytorch版

表3　预训练数据集所融合数据集

数据集名称/类别	愤怒	厌恶	恐惧	开心	中性	伤心	惊讶	总计
CK+[14]	135	177	75	207	0	84	249	927
FER2013[15]	4 953	547	5 121	8 989	6 077	4 002	6 198	35 887
Affectnet[16](节选)	3 218	2 477	3 175	5 044	5 126	3 091	4 039	26 170
MMAFEDB[17]	8 624	4 542	6 209	39 526	41 081	16 636	11 062	127 680
总计	16 930	7 743	14 580	53 766	52 284	23 813	21 548	190 664

表4　预训练数据集数据分布

数据集名称/类别	愤怒	厌恶	恐惧	开心	中性	伤心	惊讶	总计
融合数据集	10 000	7 743	10 000	10 000	10 000	10 000	10 000	67 743

本为1.7.1;CPU型号为AMD R9 5950X 16C32T,GPU型号为NVIDIA GeForce RTX3090@24G。

模型训练超参数为:单次训练迭代次数(epochs)为300,批量大小(batch-size)为256;训练的优化方法为AdamW,初始学习率设置为0.001,后续的epochs中学习率按余弦退火衰减至0。在训练过程中采用随机裁剪、平移、旋转、亮度对比度随机调整等数据增强方法。

经训练得到预训练模型在融合数据集上的准确率为83.16%。

3　实验与分析

为便于比较各种模型性能以及模型对于座舱场景的适应能力,本文使用表情识别领域常用无约束场景(数据集中图片的选取不受具体场景约束)基准数据集真实世界情感面孔数据库(real-world affective faces database, RAF-DB)[18]和汽车座舱场景表情识别数据集庆明大学驾驶员面部表情数据库(Keimyung University facial expression of drivers database, KMU-FED)[19]作为目标训练集。两个数据集样本分布如表5所示,其样本样例如表6所示。

表5　RAF-DB和KMU-FED样本分布

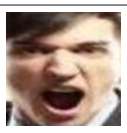
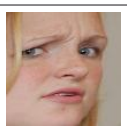
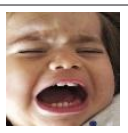
数据集名称/类别	愤怒	厌恶	恐惧	开心	中性	伤心	惊讶	总计
RAF-DB	867	877	355	5 957	3 204	2 460	1 619	15 339
KMU-FED	196	120	200	210	0	180	200	1 106

为较为全面地衡量模型性能,本文采用以下指标:准确率ACC,衡量模型表情识别精度;GFLOPs(giga floating-point operations per second),模型每秒须进行的浮点数计算次数,衡量模型计算量;参数量Parameters,衡量模型占用物理内存大小;FPS(frame per second),模型每秒处理的图片数量,衡量模型识别速度。

为更好地研究模型在智能座舱计算平台的识别性能,实验在车载嵌入式域控制器NVIDIA Jetson Orin上进行,该平台搭载Ubuntu20.04系统,并配置python版本3.8,pytorch版本1.11.0,CUDA版本为11.0。

在RAF-DB训练时,其训练集与验证集之比为4:1,设置模型迭代次数100 epochs;批量大小64;优化方法为AdamW,并设置初始学习率0.001并余弦退火至0,权重衰退1×10^{-3}。由表5可见,RAF-DB存在严重的数据不均衡问题,为改善数据分布,对"愤怒""厌恶""恐惧"3类施加随机旋转,随机平移,随机中心裁剪,随机裁剪,随机亮度、对比度、色调变化,随机灰度化等数据增强方法以增广数据集。在训练过程中则进一步对所有数据施加上述数据增强策略,以模拟驾驶过程中可能存在的光照变化,驾驶员脸部姿态变化以及脸部受到遮挡或不完全入镜等

表6　RAF-DB和KMU-FED样本样例

数据集名称/类别	愤怒	厌恶	恐惧	开心	中性	伤心	惊讶
RAF-DB							
KMU-FED							

情况,从而提高模型对于不同干扰情况的适应性。

在KMU-FED训练时划分训练集、验证集与测试集之比为6∶2∶2,并通过YOLOv5l人脸检测器提取图片中人脸区域。模型迭代次数20 epochs,其余训练参数与在RAF-DB训练时相同。

EmotionNet座舱内驾驶员表情识别结果示例示意图如图8所示。图中驾驶员脸部区域由YOLOv5l人脸检测器检测取得。

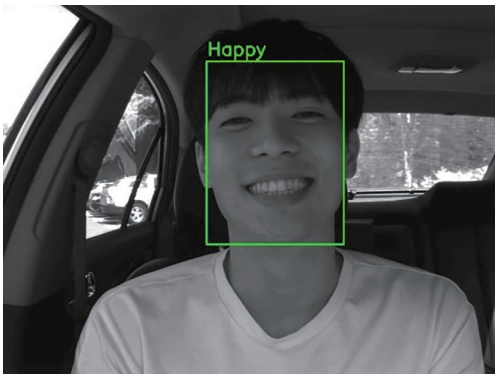


图8 EmotionNet表情识别结果示例示意图

3.1 对比实验

EmotionNet与基线(Baseline)GhostNet和当前优秀轻量化模型MobileNetV3-small[20]、EfficientNetV2-S[21]以及复现的其他优秀表情识别模型A-MobileNet[5]和Deep-Emotion[7]在RAF-DB和KMU-FED数据集上的对比结果如表7所示。可见得益于对于重要特征关注能力的提高和训练过程的改善,EmotionNet在RAF-DB和KMU-FED数据集上均取得了最高的检测准确率,分别达到85.23%和99.95%。在RAF-DB数据集上,与准确率第2名的EfficientNetV2-S相比,EmotionNet具有更高的检测速度和计算量及参数量优势;与第3名A-MobileNet相比,EmotionNet虽然检测速度较低,但更为节约计算资源和存储空间。

在KMU-FED数据集上,Deep-Emotion同样取得了较高的检测准确率,并具备极大的实时性优势,但由于其模型结构过于简单,在较大规模数据集中提取特征能力稍显不足。

表7 对比实验结果

模型	ACC/RAF-DB	ACC/KMU-FED	FPS	GFLOPs	Parameters
GhostNet	81.14%	97.32%	33.39	0.052	3.91 M
MobileNetV3-small	80.51%	81.39%	46.88	0.023	1.68 M
EfficientNetV2-S	85.12%	97.77%	20.72	0.949	20.19 M
A-MobileNet	84.49%	87.40%	79.81	0.756	5.26 M
Deep-Emotion	76.46%	99.55%	406.50	0.159	2.28 M
EmotionNet	85.23%	99.95%	59.89	0.345	4.75 M

3.2 消融实验

为研究所提出模型各部分改进对于模型性能的影响,本文通过消融实验验证模型结构各部分改进以及训练过程优化的影响。消融实验以RAF-DB为目标数据集,实验环境及超参数设置与对比实验相同。

消融实验的结果如表8所示,表中"√"表示采用了对应列的改进方法,第1行为不进行任何改进的GhostNet模型,最后1行为采用本文所述所有改进方法的EmotionNet模型。

从表8可以看出,EmotionNet整体较Baseline在准确率和检测速度上提高了4.09%和26.5 FPS,虽然带来了计算开销增大0.293GFLOPs和参数量增大0.84 M的问题,但总体仍属于轻量化模型范畴,且与需要共同部署的目标检测网络等资源占用较大模型相比可以忽略不计。如表8第3行所示,将GhostNet网络的G-bneck模块替换为C-CA-Block模块后,模型以较低的速度代价将准确率从78.78%提升到80.02%,而计算量和参数量几乎没有增加,表现出较高的计算效率。特征增强网络带来了最大的准确率提升,达3.20%,证明了融合不同尺度特征的有效性。然而这一网络也付出了较大计算代价,引入了FPS大幅降低和计算量参数量增加的问题。双分类头决策级融合对准确性的提升较小仅有0.13%,其原因可能在于颈部网络已经实现了对于不同尺度特征图较为充分的利用,导致决策级融合仅能够起到有限的补充作用。

在训练过程方面,中心损失的引入能够小幅提高模型0.17%准确率,且不会导致计算开销的增大。迁移学习方法则带来了较大的准确率提升,达1.71%,证明了在正式训练中加载在更大规模数据集上预训练所取得的模型权重,可以在一定程度上"继承"在大型数据集上的训练结果,从而提高模型对表情的识别能力。

表8　消融实验结果

模块精简	C-CA-Block	特征增强网络	双分类头	中心损失	迁移学习	ACC	FPS	GFLOPs	Parameters
						81.14%	33.39	0.052	3.91 M
√						78.78%	110.55	0.035	1.82 M
√	√					80.02%	104.45	0.035	1.82 M
√	√	√				83.22%	63.16	0.326	4.67 M
√	√	√	√			83.35%	51.87	0.345	4.75 M
√	√	√	√	√		83.52%	51.87	0.345	4.75 M
√	√	√	√	√	√	85.23%	51.87	0.345	4.75 M

值得注意的是,与GhostNet相比,EmotionNet模型计算量增大但速度却加快。其原因可能主要有两个方面:一方面是因为模型经过精简,减少了特征提取模块数量,卷积单元数量远少于GhostNet;另一方面,由表8可知,模型计算量的增加主要在于普通卷积单元组成的颈部网络部分,而在GPU加速计算过程中,与深度可分卷积相比,普通卷积有着更高的计算并行度,虽然计算量更大但计算速度却不会相差太多。

GhostNet与EmotionNet在RAF-DB数据集对于7种类别表情的识别准确率混淆矩阵分别如图9(a)和图9(b)所示,其中每一行代表数据集样本真实类别,每一列代表模型基于数据集样本预测的表情类别。可见经改进后,模型对于除"恐惧"以外的表情识别准确率均有所提高。

	愤怒	厌恶	恐惧	开心	中性	伤心	惊讶
愤怒	75%	10%	2%	59%	3%	2%	2%
厌恶	8%	45%	4%	8%	16%	13%	6%
恐惧	5%	3%	54%	5%	7%	7%	18%
开心	1%	1%	0%	92%	4%	1%	1%
中性	1%	3%	1%	5%	79%	8%	3%
伤心	3%	4%	0%	5%	12%	75%	1%
惊讶	3%	2%	2%	3%	6%	2%	82%

(a) GhostNet混淆矩阵

	愤怒	厌恶	恐惧	开心	中性	伤心	惊讶
愤怒	81%	5%	2%	3%	6%	1%	1%
厌恶	8%	52%	2%	7%	16%	12%	3%
恐惧	6%	4%	50%	5%	6%	12%	18%
开心	0%	1%	0%	94%	3%	1%	1%
中性	2%	2%	1%	3%	85%	5%	2%
伤心	1%	2%	2%	4%	8%	82%	1%
惊讶	2%	1%	1%	2%	5%	4%	84%

(b) EmotionNet混淆矩阵

图9　EmotionNet与GhostNet在RAF-DB数据集混淆矩阵

结合表5中的数据分布,"开心"表情具有最大的数据量,在改进前后均有很高的识别准确率;"中性"和"伤心"数据量较高,在改进后亦取得了相对较大的准确率提高;"惊讶"表情数据量稍少,但由于此类表情表达相似性较高,较易于与其他表情区分,改进前后均有较高的准确率。

在数据量较少的"愤怒"、"厌恶"和"恐惧"类别中,"愤怒"表情情况与"惊讶"较为相似,样本与其他类别具有较高的差别,故识别准确率较高,但仍低于数据量更多的表情。由于不同类别间表情表达的相似性,"厌恶"样本主要被误分类为"中性"和"伤心",而"恐惧"样本则主要被误分类为"伤心"与"惊讶"。部分被错误分类的"厌恶"和"恐惧"样例及它们易混淆类别的图片样例如表9所示。模型特征提取能力的提高和训练过程的改善虽然在一定程度上改善了部分样本较少的类别分类准确率,但由于数据量有限,不能从根本上解决这一问题,导致这两个类别识别准确率较低。

表9　部分被误分类表情示例

样例真实类别	样例1	样例2	样例3	样例4	样例5	易与示例类别混淆表情类别样例		
厌恶						 中性	 中性	 伤心
恐惧						 惊讶	 惊讶	 伤心

改进后“恐惧”准确率降低的原因可能为该类别数据量在验证集中最少,导致模型在训练过程中更倾向于提高数据量更大、更易区分类别的准确率从而确保验证集整体准确率的提升,而对于数据量较少的类别则关注更低。模型注意力机制的增强也进一步助长了对于数据量更大样本特征的关注,加剧了这一倾向。

4 结论

针对智能座舱表情识别深度学习模型准确率和实时性难以兼顾的问题,本文提出了基于注意力融合与特征增强网络的智能座舱驾驶员表情识别模型EmotionNet。其中,本文在G-bneck基础上增加了一条普通卷积加通道注意力的特征提取分支,从而构建融合通道注意力和坐标注意力的C-CA-Block,并以之构筑主干网络;增加融合8倍、16倍下采样特征图的特征增强颈部网络,融合不同深度特征;最后将8倍、16倍下采样特征图分别输入两个分类头处理并将结果加权相加从而实现对不同尺度特征图决策级融合。在训练阶段,则通过迁移在大型一般情景融合数据集上取得的预训练权重和Center loss损失进一步提高了模型区分不同表情的能力。在一般场景数据集RAF-DB和座舱场景数据集KMU-FED上本文方法均取得较高的检测精度且实时性亦较Baseline GhostNet有一定提高,具备一定的智能座舱表情识别任务的适应性。在后续的研究中,将进一步探索不同注意力机制与特征提取方法的结合以及降低特征增强网络计算成本的方法,并通过TensorRT优化方法进一步提高模型检测性能。

参考文献

[1] 李文博, 刘羽婧, 张峻铖, 等. 驾驶员情绪-驾驶风险机理分析[J]. 机械工程学报, 2022, 58(22): 379-394.
LI W B, LIU Y J, ZHANG J C, et al. Analysis of the influence mechanism of driver's emotion on driving risk[J]. Chinese Journal of Mechanical Engineering, 2022, 58(22): 379-394.

[2] TAMANANI R, MURESAN R, Al-DWEIK A. Estimation of driver vigilance status using real-time facial expression and deep learning[J]. IEEE Sensors Letters, 2021, 5(5): 1-4.

[3] XIAO H, LI W, ZENG G, et al. On-road driver emotion recognition using facial expression [J]. Applied Sciences, 2022, 12(2): 807.

[4] KODHAI E, POOVESWARI A, SHARMILA P, et al. Literature review on emotion recognition system[C]. 2020 International Conference on System, Computation, Automation and Networking (ICSCAN). IEEE, 2020: 1-4.

[5] NAN Y, JU J, HUA Q, et al. A-MobileNet: an approach of facial expression recognition[J]. Alexandria Engineering Journal, 2022, 61(6): 4435-4444.

[6] 梁艳, 温兴, 潘家辉. 融合全局与局部特征的跨数据集表情识别方法[J]. 智能系统学报, 2023, 18(6): 1205-1212.
LIANG Y, WEN X, PAN J H. Cross-dataset facial expression recognition method fusing global and local features [J]. CAAI Transactions on Intelligent Systems, 2023,18(6): 1205-1212.

[7] MINAEE S, MINAEI M, ABDOLRASHIDI A. Deep-emotion: facial expression recognition using attentional convolutional network[J]. Sensors, 2021, 21(9): 3046.

[8] HAN K, WANG Y, TIAN Q, et al. GhostNet: more features from cheap operations[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1580-1589.

[9] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition[C]. Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, part VII 14. Springer International Publishing, 2016: 499-515.

[10] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[11] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13713-13722.

[12] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.

[13] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8759-8768.

[14] KANADE T, COHN J F, TIAN Y. Comprehensive database for facial expression analysis[C]. Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (cat. No. PR00580). IEEE, 2000: 46-53.

[15] GOODFELLOW I J, ERHAN D, CARRIER P L, et al. Challenges in representation learning: a report on three machine learning contests[C]. Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20. Springer berlin heidelberg, 2013: 117-124.

[16] MOLLAHOSSEINI A, HASANI B, MAHOOR M H. AffectNet: a database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing, 2017, 10(1): 18-31.

[17] MAHMOUDIMA. MMA facial expression[DB/OL]. (2020-01-01)[2024-2-15]. https://www.kaggle.com/mahmoudima/mma-facial-expression? select=MMAFEDB.

(下转第1686页)