DOI:10.19651/j.cnki.emt.2106676

基于通道注意力机制的人脸表情识别机器人交互研究*

张 波 兰艳亭 鲜 浩 方 炜
(中北大学 电气与控制工程学院 太原 030051)

摘 要：为提升机器人与人类之间的交互能力,实现人机交互更加智能、自然。提出了基于通道注意力机制的人脸表情识别方法,以双足人形机器人 NAO 为实验平台,设计了能进行人脸表情识别的人机交互系统。首先,通过 RAF-DB 数据集对注意力机制的人脸表情识别算法进行训练,训练结果显示,模型可以对 7 种基本表情(高兴、生气、恶心、恐惧、伤心、惊讶和自然)进行识别,其准确率可以达到 76.21%。其次设计 NAO 机器人面对不同表情时的交互语音和动作,最后,对整个人机交互系统进行测试。测试结果显示,当 NAO 机器人接收到电脑端识别的情绪后,会像人类一样说话和做出动作。

关键词：深度学习;人脸表情识别;NAO 机器人;人机交互;通道注意力机制;情绪识别;表情分类;RAF-DB
中图分类号：TP391.4 **文献标识码**：A **国家标准学科分类代码**：520.2

Research on robot interaction of facial expression recognition based on channel attention mechanism

Zhang Bo Lan Yanting Xian Hao Fang Wei
(School of Electrical and Control Engineering, North University of China, Taiyuan 030051, China)

Abstract: In order to enhance the interaction between robots and humans, the human-computer interaction is more intelligent and natural. A facial expression recognition method based on channel attention mechanism is proposed. Using the biped humanoid robot NAO as an experimental platform, a human-computer interaction system capable of facial expression recognition is designed. First, the facial expression recognition algorithm of the attention mechanism is trained through the RAF-DB data set. The training results show that the model can recognize 7 basic expressions (happy, angry, nauseous, fear, sad, surprised and natural). Its accuracy rate can reach 76.21%. Secondly, design the voice and actions of the NAO robot when facing different expressions, and finally, test the entire human-computer interaction system. The test results show that when the NAO robot receives the emotions recognized by the computer, it will speak and act like a human.

Keywords: deep learning; facial expression recognition; NAO robot; human-computer interaction; channel attention mechanism; emotion recognition; emoticon classification; RAF-DB

0 引 言

随着深度学习的发展,逐渐成为人机交互技术领域的热点技术之一。人脸表情识别(FER)技术在人机交互、商业、医疗、娱乐、心理学,疲劳驾驶[1]等领域都有广阔的应用前景。在未来,新型的人机交互系统除了传统的键盘,鼠标之外还有语音、姿势、人脸表情等,构建识别情绪的人机交互系统离不开成熟的人脸表情识别技术,面部表情识别是这些系统中特别重要的一部分。但构建这样的系统仍然面临许多挑战。比如,在实际应用中,肤色对人脸检测的影响[2]。光照强度的变化、噪声的影响以及人脸表情的复杂性和多样性等[3]。在最近的几年里,随着计算机科学与计算机理论得到了飞速的发展。人工智能(AI)和模式识别(PR)也开始盛行,吸引了越来越多的研究者进入这一领域。人脸表情识别作为 AI 领域的一个分支,具有重要的研究价值,并且已经成为一个研究热点。为了解决人脸识别

收稿日期:2021-05-13
* 基金项目:山西省科技攻关项目(20140311027-2)、中北大学 17 届研究生科技立项项目(20201772)资助

中实际遇到的问题。学者们不断提出新的深度学习算法来提高图像的识别率和鲁棒性。

传统的人脸表情识别方法主要的有两种：第 1 种是基于几何的方法，第 2 种是基于整体的识别方法。传统识别方法依赖于前期人工提取特征的优劣，人为干扰因素较大[4]。与传统方法不同，在 2006 年，Hinton 等提出深度信念网络(deep belief networks，DBNs)，使得深度学习重新引起重视。2012 年，Krizhevsky 等[5] 于 2012 年在 ImageNet 图像数据集上使用 AlexNet 卷积神经网络结构取得惊人的成绩，其识别率远超传统的识别方法。深度学习的出现打破了表情识别中传统的先特征提取后模式识别的固定模式，可以同时进行特征提取与表情分类，而且深度学习对特征的提取是通过反向传播算法与误差优化对权值进行迭代更新，所以能够提取出人类预想不到的关键点和特征[6]。史涛等[7] 针对尺度不变特征变换(SIFT)算法在特征提取过程中运算量过大、非主要特征数据冗余、匹配率低等问题，提出一种基于 SIFT 稀疏深度信念网络算法模型。

张娜等[8] 提出一种基于 PCA 与 LDA 融合的人脸识别算法，增强人脸识别技术的鲁棒性。Saurav 等[9] 提出了双集成 CNN(DICNN)模型，该模型整合并优化了两个自定义轻量级 CNN 模型，实现在准确性和计算效率之间取得合理的平衡。文献[10]提出了增强可分离卷积通道特征的轻量化的卷积神经网络表情识别模型。这些方法虽然对准确率有一定的提升，但是模型越复杂，参数量也越大。对计算机的计算能力要求也越高。无法对嵌入式设备进行部署。

上述研究成果有效的推动了人脸表情识别的发展，但仍有一些不足，比如应用中总是效果不理想。因此设计了 ResNet 和 Xception 的融合网络模型，通过 Xception 减少 ResNet 模型中的参数，并加入注意力机制 Senet 提高 Xception 的通道相关性，提高人脸表情的识别率。同时以软银 Aldebaran Robotics 公司研制的双足机器人 NAO 为开发平台，将两者结合，搭建一个基于通道注意力机制的人脸表情识别机器人交互系统。测试结果表明，本系统能正常识别人类的 7 种基本表情并实现语音和动作的交互。

1　通道注意力融合网络模型设计

1.1　Resnet 网络模型简介

残差神经网络(ResNet)是由微软研究院的何恺明等提出的[11]。ResNet 网络取得了 2015 年的 ILSVRC(ImageNet Large Scale Visual Recognition Challenge)竞赛冠军。残差神经网络提出了恒等映射(identity mapping)问题。可以解决网络模型过深导致的“退化现象(degradation)”，并针对退化现象发明了“快捷连接(shortcut connection)”，其结构如图 1 所示。

从图中我们可以看到与正常的网络结构相比，右侧多了一个快捷连接，直接把前两层的输出与本层的输出求和，通过激活函数作为本层的最终输出。残差输入为 x，输出为 $H(x)$ 则可得到：

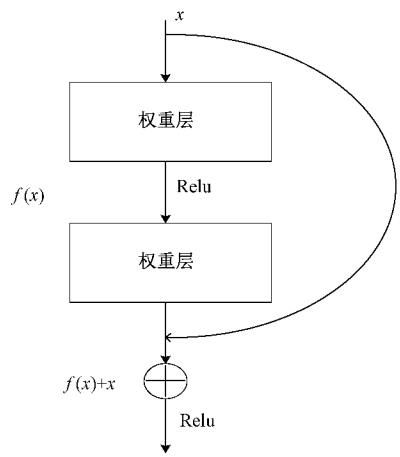


图 1　残差结构

$$H(x)=F(x)+x \tag{1}$$

其中，$F(x)$ 就是需要学习的残差。

1.2　Xception 网络模型简介

Xception 网络模型是 Chollet[12] 提出的，其主要思想是把通道之间的相关性和空间相关性分开处理，完全分离每一个通道，不考虑通道之间的联系。在 Inception[13] 网络的基础上将 3×3 的卷积换成 1×1 的卷积，从而实现 k 个通道的完全分离。Extreme Inception 结构如图 2 所示。

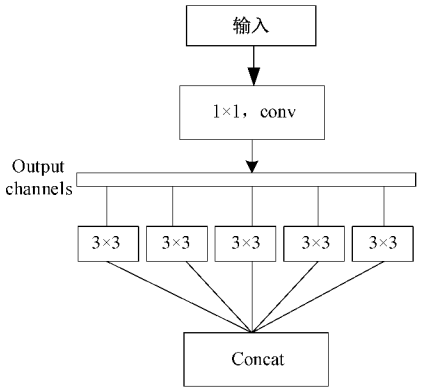


图 2　Extreme Inception 结构图

通过此方法加宽了网络，降低了网络的参数量和复杂度。同时加入类似 Resnet 的残差连接机制，加快了网络收敛速度，整个网络分为 Entry，Middle 和 Exit 3 部分。Xception 整体架构如图 3 所示。在 ImageNet 数据集上，Xception 比 Inception-v3 的准确率稍高，同时参数量有所下降，但是 Depthwise Separable Convolution 其计算过程较为零散，在训练迭代过程中效率更慢一些。

1.3　通道注意力融合网络模型

提出了一种通道注意力相关性的融合 Resnet-Xception-Senet(RXS)模型，网络架构如图 4 所示。此网络在 Resnet-Xception 网络的通道上进行改进，借鉴 Resnet 网络的恒等映射，和 Xception 的通道分离减少参数，但是

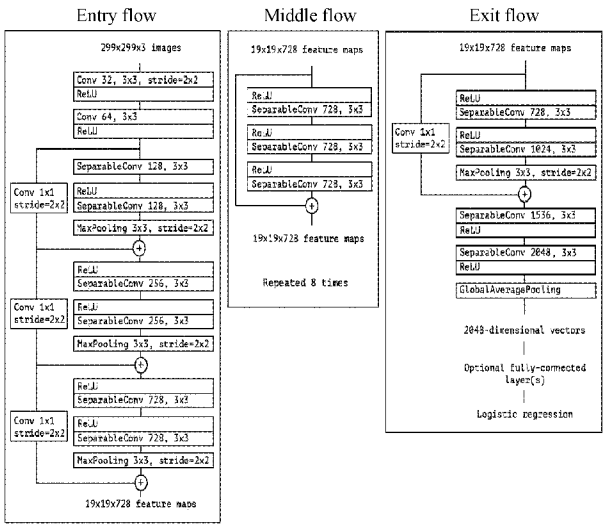


图 3　Xception 整体架构

考虑 Resnet 恒等映射过程中忽视了通道的作用,依然是上几层的通道分布状态。Xception 中的可分离卷积却把通道之间的相关性完全分离,丢失了相应的提取特征,因此加入了注意力机制的 SEnet[14] 网络。Senet 首先通过 GAP (Global Average Poolin)压缩(squeeze),然后通过两次全连接层,得到网络各个通道的权值,通过 sigmod 激活函数把通道权值范围限定到 0～1 之间,最后与可分离卷积模块输出对应通道相乘,得到重要的通道,忽视不重要的通道信息。

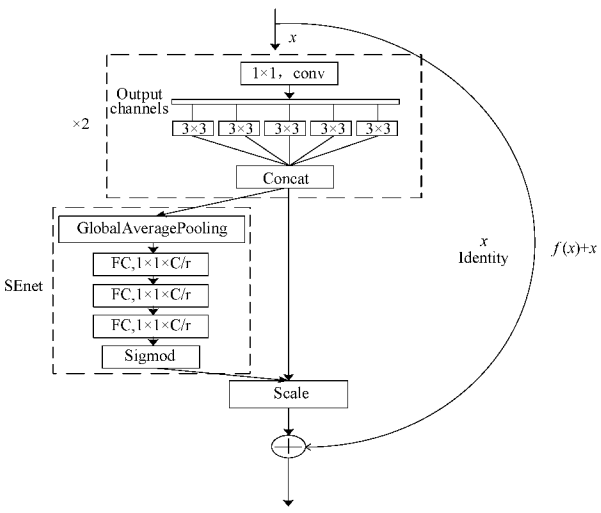


图 4　通道注意力融合网络架构

2　NAO 机器人交互设计

NAO 机器人使用的是 NAOqi 系统。在程序执行中是通过一个代理程序(broker)去加载需要的库文件。每个库文件中又包含一个或者多个的模块。例如需要用到的“ALVideoDevice”视觉模块,“ALAudioDevice”声音模块,“ALMtion”动作模块。

2.1　视觉模块

NAO 机器人拥有两个 2D 摄像头,分别分布在头部的前额和嘴部,其分布如图 5 所示。摄像头可以获取最高 1 280×960 分辨率的图片和视频流,已广泛应用于教育、辅助医疗等领域[15 16]。由于摄像头采用的是 MT9M114 图像传感器,此设备直接输出的照片格式为 YUV。但是在进行人脸表情识别时,采用的是 RGB 格式的色彩空间。因此需要订阅“ALVidcoDevice”模块,将 YUV 格式变为 RGB。考虑 NAO 机器人的 CPU 计算能力有限,我们选择了分辨率为 640×480。这个分辨率不会使人脸图片信息损失过多,也不会让机器人的处理器过载。

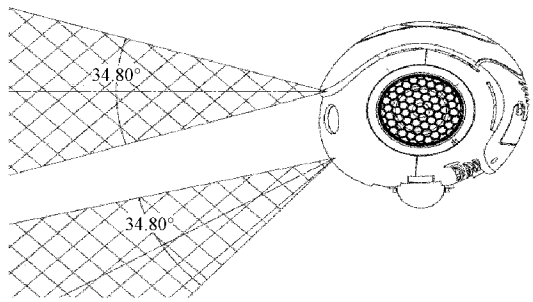


图 5　NAO 机器人的摄像头分布

2.2　语音模块

NAO 机器人的头部安装了 4 个麦克风和 2 个扬声器,其中麦克风的接受频率范围是 150 Hz～12 kHz,保存的文件格式为 WAV 或者 OGG。在 NAOqi 系统中“ALAudioDevice”声音模块管理音频的输入和输出,因此需要读取麦克风采集的数据时,必须订阅“ALAudioDevice”声音模块。

2.3　动作模块

在机器人交互过程中,根据 NAO 摄像头采集的人脸图片,在电脑端经过模型预测后,输出到 NAO 机器人,机器人根据不同的表情做出不同的交互动作,其中“ALMtion”动作模块包括了与机器人的动作相关的方法。采用官方自带的 Choregraphe 图形化编程软件编写机器人动作。运用时间轴指令盒进行制定动作的设计,可以不必使用复杂的机器人运动学知识和程序编程。时间轴如图 6 所示。

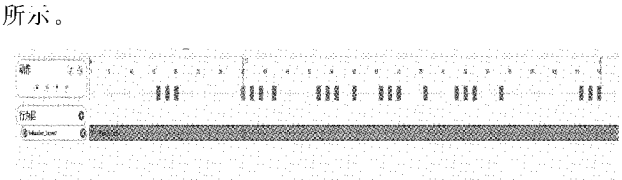


图 6　时间轴

3　实验结果及交互测试

3.1　模型训练

使用 TensorFlow2.3 深度学习平台进行模型训练,操

作系统为 Windows 10，电脑 CPU 为英特尔 Core i5-9400F，内存为 16 GB，GPU NVIDIA GeForce GTX 1060 6 GB，为了加速训练过程，使用了 CUDA（compute unified device architecture）和 CUDNN。采用 RAF-BD（real-world affective faces database）人脸表情数据集进行实验。

1）实验数据集

RAF-DB 数据集[17]。该数据库（RAF-DB）是一个大规模的面部表情数据库，与实验室 JAFFE 数据库不同，数据库中的图像对受试者的年龄，性别和种族，头部姿势，光照条件，遮挡（例如眼镜，面部毛发或自我遮挡），进行过处理操作（例如各种滤镜和特殊效果），其部分图例如图 7 所示，包括：29 672 个真实世界的图像，两个不同的子集：单标签子集，包括 7 类基本情感；复合标签子集，包括 12 类复合情感，每个图像 5 个准确的地标位置，37 个自动地标位置，边框，种族，年龄范围和性别属性注释，基本情绪和复合情绪的基线分类器输出[18]。该数据库已分为训练集（12 271 张）和测试集（3 068 张），两个集中的表达式都接近相同的分布。

图 7　RAF-DB 数据集部分图例

2）数据增强

由于 RAF-BD（real-world affective faces database）人脸表情数据集的训练样本较少，在训练过程中很容易出现过拟合现象。因此采用了数据增强。数据增强是利用数据库中现有的图片，通过随机变化生成可信任的图像来增加训练数据。

本次实验使用 kcars 框架中的 ImageDataGenerator 模块对数据集进行数据增强。其中数据增强的具体参数值如表 1 所示。

表 1　ImageDataGenerator 模块参数表

参数	值
rotation	10
width	0.3
height	0.3
shear	0.1
zoom	0.3
horizontal_flip	True
fill_mode	nearest

其中“rotation”参数代表图片的旋转角度值，旋转范围为 0°～180°。参数“width”和“height”是代表图像在水平或垂直方向上平移的范围。参数“shear”是随机剪切变换的范围，对图片不同区域进行剪切。参数“zoom”是对图像随机缩放的范围。参数“horizontal_flip”是随机将一半图像水平翻转。参数“fill_mode”是用于填充新创建像素的方法。

3）网络超参数设置

在训练过程中，除了网络结构会响应输出的准确率，其中网络的超参数也是至关重要。网络学习过程中就是通过优化器更新参数，使损失函数达到最小值，由于优化参数对目标函数的依赖各不相同，如果只设置统一的全局学习率，对于梯度很大的学习过程，由于步长过长收敛速度会很慢，当学习率设置过大时，已经优化好的参数反而会震荡，出现不稳定的情况。因此首先选择了目前效果较好的 Adam[19]优化器优化训练过程，实现学习率的自适应。学习率设置为 0.001，在模型训练过程中，验证损失在 5 轮内都没有更好的收敛，则会对学习率进行调整降低，最低限制降为 0.000 01。为了防止网络过早的过拟合，还加入了 dropout 层，其参数设置为 0.3，可以在更大的范围内进行搜索最小值。网络的 Batch 为 64，由于本实验属于多分类问题，目标函数采用交叉熵损失函数（categorical_crossentropy loss）。

4）模型结果分析

为了验证设计的模型效果，在 RAF-BD（real-world affective faces database）数据集上进行不同经典模型对比实验。其中经典模型分别为结构非常简洁的 VGG19，模型较深的 ResNet101V2，轻量级的 MobileNetV2[20]，极端可分离卷积的 Xception，与文献[10]中的改进模型。不同网络模型的参数量如表 2 所示。

表 2　不同网络模型的参数量

模型	参数量（B）
VGG19	22 123 591
ResNet101V2	55 735 815
MobileNetV2	10 452 039
Xception	33 970 735
文献[10]	7 836 279
通道注意力融合算法	2 075 335

从表 2 可以看出，本文提出的网络模型参数量在所列出的对比模型中是最少的，仅仅只有大约 1.9 M。ResNet101V2 网络和 VGG19 网络的参数量大约为 55 M 和 22 M，通道注意力融合算法在训练过程中每个 epochs 用时也是最少，从而效率也是最高的，这也证明了加入的 SEnet 网络中两个全连接层构成的瓶颈层，并没有过多的增加网络的参数量，实现了网络模型的轻量化。

在模型的实现过程中，VGG19、ResNet101V2、MobileNetV2、Xception 四个模型使用了迁移学习模型，采

用了 kears 中预训练好的模型,冻结基础层,重新训练了顶层的全连接层,文献[10]的模型是通过手动复现。不同模型在 RAF-BD、FER2013 数据集下学习 20 个 epochs 后,得到的准确率如表 3 所示。

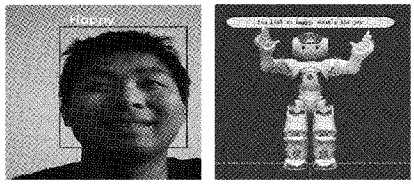
表 3　不同模型训练后的准确率

模型	RAF-BD 准确率
VGG19	66.31%
ResNet101V2	65.32%
MobileNetV2	52.17%
Xception	75.06%
文献[10]	71.57%
通道注意力融合算法	76.21%

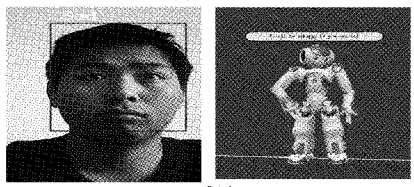
从表 3 可以看出,与现有的模型相比,提出的通道注意力融合模型在 RAF-BD 数据集的准确率达到了 76.21%。通道注意力融合算法得到了最高的准确率。证明通道注意力融合模型在人脸表情识别任务中有一定的优势。

3.2　人机交互测试

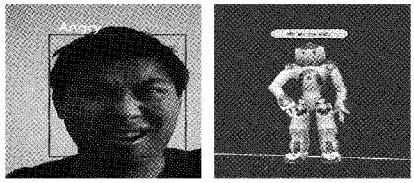
由于训练好的人脸表情模型是在 Python3.6 版本下完成的。而 NAO 机器人只能支持 Python2.7 版本,无法进行直接部署,因此本次交互实验通过无线连接方式进行。连接过程中需要注意,机器人第 1 次与计算机连接时需要插入网线,获取相应的 IP 地址,然后通过 IP 地址登录网页,设置路由器进行无线连接。机器人正常连接网络后,通过无线把程序下载到机器人中,PC 端通过调用机器人的摄像头,获得人脸图像,送入训练完成的模型中,模型预测后的结果再回传给 NAO 机器人,机器人根据识别到的情绪,做出相应的语音和动作。选取了比较具有代表性的开心,难过,生气 3 类表情,其交互结果如图 8 所示。



(a) 开心



(b) 难过



(c) 生气

图 8　部分情绪交互结果

从图 8 中可以看到,左侧的图片为摄像头采集后模型预测的表情结果。右侧图片为机器人做出的语音和相应的动作,为了可视化语音效果,连接至虚拟机器人,对语音结果直接可视化显示。当检测到开心表情时,机器人会说和你互动,并会说出“你这么开心,是遇到什么好事了吗?”。检测为难过表情时,机器人会做出低头难过的动作,并说“你难过我也会不开心哟”。检测到生气表情时,机器人会手叉腰,并说“你为什么生气呢”。通过多次测试,交互结果显示,机器人都能稳定的识别情绪,并作出相应的动作交互。

4　结　　论

提出了一种基于通道注意力机制的人脸表情识别机器人交互系统,首先通过注意力机制 SEnet 模块,可分离卷积模块和残差学习模块设计网络模型,以提高面部表情识别准确率。SEnet 模块通过注意力机制对通道加权,通过学习自动获取每个通道的重要程度,抑制不重要的特征通道。可分离卷积模块极大的减少了模型的参数量。残差学习模块可以有效的减轻网络的退化问题。其次把训练好的模型应用到双足 NAO 机器人中,为机器人设计了语音和动作交互功能,让机器人预测人脸表情后可以进行自然的交互。本系统在养老、医疗、服务业有巨大的市场。但是在应用过程中由于 NAO 机器人的成本过高,导致系统受众少,在后面的工作中会移植到成本更低的设备中。使得此交互系统能被更多人使用。

参考文献

[1]　李锐,蔡兵,刘琳,等. 基于模型的驾驶员眼睛状态识别[J]. 仪器仪表学报,2016,37(1):184-191.

[2]　吴昊,胡敏,高永,等. 融合 DCLBP 和 HOAG 特征的人脸表情识别方法[J]. 电子测量与仪器学报,2020,34(2):73-79.

[3]　XI Z, NIU Y, CHEN J, et al. Facial expression recognition of industrial internet of things by parallel neural networks combining texture features[J]. IEEE Trans. Ind. Inf, 2021,17: 2784-2793.

[4]　李勇,林小竹,蒋梦莹. 基于跨连接 LeNet-5 网络的面部表情识别[J]. 自动化学报,2018,44(1):176-182.

[5]　KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2012, 25(2): 1097-1105.

[6]　叶继华,祝锦泰,江爱文,等. 人脸表情识别综述[J]. 数据采集与处理,2020,35(1):21-34.

[7]　史涛,秦琴. 基于 SIFT 稀疏深度信念网络人脸识别研究[J]. 国外电子测量技术,2019,38(3):88-92.

[8]　张娜,刘坤,韩美林,等. 一种基于 PCA 和 LDA 融合的人脸识别算法研究[J]. 电子测量技术,2020,43(13):72-75.

[9] SAURAV S, GIDDE P, SAINI R, et al. Dual integrated convolutional neural network for real-time facial expression recognition in the wild[J]. The Visual Computer, 2021: 1-14.

[10] 梁华刚,雷毅雄.增强可分离卷积通道特征的表情识别研究[J/OL].计算机工程与应用:1-13[2021-03-22]. http://kns.cnki.net/kcms/detail/11.2127.TP.20210113.1520.014.html.

[11] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[12] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1251-1258.

[13] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[14] 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.

[15] 徐桂芝,赵阳,郭苗苗,等.基于深度分离卷积的情绪识别机器人即时交互研究[J].仪器仪表学报,2019,40(10):161-168.

[16] WANG Z, LI B, DE SILVA C W. Use of fuzzy neural network in industrial sorting of apples[J]. Instrumentation,2019,6(4):37-46.

[17] LI S, DENG W, DU J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017: 2852-2861.

[18] 张哲源.基于子空间嵌入及迁移学习的表情识别研究[D].广州:广东工业大学,2020.

[19] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. ArXiv preprint, 2014,ArXiv: 1412.6980.

[20] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: Inverted residuals and linear bottlenecks[J]. IEEE, 2018: 4510-4520.

作者简介

张波,硕士研究生,主要研究方向为智能控制。

E-mail:723550024@qq.com

兰艳亭(通信作者),副教授,主要研究方向为仪智能控制等。

E-mail:lytcyb@foxmail.com