基于注意力改进残差网络结构的表情识别方法

张 智[1] 魏 蘅[2]

[1](武汉科技大学计算机科学与技术学院 湖北 武汉 430065)

[2](武汉科技大学智能信息处理与实时工业系统湖北省重点实验室 湖北 武汉 430065)

摘 要 针对目前 CNN 在复杂图像中特征提取不充分的问题,提出一种基于注意力的改进残差网络的表情识别网络。设计一个双流网络在完成粗特征表情识别的同时检测关键点,并使用注意力机制增大关键点周边特征的权重。随后以残差网络为基础模型,改进残差块之间的跳跃连接方式,并将残差块中的普通卷积改进为分组卷积来强化特征提取能力。最后联合两个表情识别网络进行分类,实验结果验证了该模型方案有着更卓越的性能。

关键词 人脸表情识别 残差网络 注意力机制 分组卷积

中图分类号 TP391.4 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.08.023

FACIAL EXPRESSION RECOGNITION METHOD BASED ON MULTI-CHANNEL RESIDUAL NETWORK

Zhang Zhi[1] Wei Heng[2]

[1](*College of Computer Science & Technology, Wuhan University of Science & Technology, Wuhan 430065, Hubei, China*)

[2](*Province Key Laboratory of Intelligent Information Processing and Real-time Industrial Systems, Wuhan University of Science & Technology, Wuhan 430065, Hubei, China*)

Abstract To solve the problem of insufficient feature extraction of CNN in complex images, an improved residual network based on attention is proposed for facial expression recognition. A dual stream network was designed to detect the key points while completing the coarse feature facial expression recognition, and the attention mechanism was used to increase the weight of the features around the key points. Based on the residual network model, the jump connection between residual blocks was improved, and the ordinary convolution in residual blocks was improved to block convolution to enhance the feature extraction ability. Two facial expression recognition networks were combined for classification. The experimental results show that the model scheme has better performance.

Keywords Facial expression recognition Residual network Attention mechanism Group convolution

0 引 言

心理学家 A. Mehrabia 认为,人类在平时生活中传达的信息总量里只有 7% 来自于语言,而有 55% 来自于我们的表情。因此面部表情识别一直是图像识别中的一个热门话题。这是因为它具有广泛的潜在应用,如行为分析、生理学、人机交互等。

表情识别的整个过程大致可以划分为三个基本步骤,即数据预处理、特征提取、分类,其中特征提取是最关键的一步。人脸图像预处理主要包括人脸矫正、检测。在处理输入图像时,需要检测出图像中的人脸区域,随后利用人脸矫正将人脸图像旋转成正面。特征提取的方法主要分为传统特征提取和深度特征提取,在传统的特征提取中,表情图像中的特征点往往由人工标注,因此特征提取的质量受到很大程度的人为影

收稿日期:2021 - 03 - 06。国家自然科学基金项目(61673304);国家社会科学基金重大计划项目(11&ZD189)。**张智**,副教授,主研领域:人工智能,机器学习。**魏蘅**,硕士。

响,导致最后的表情识别模型鲁棒性较差。随着深度学习的不断发展,深度特征提取受到了越来越多的关注,在深度特征提取过程中,特征点的提取不再需要人为参与,网络会根据提供的数据集样本自主地学习图像特征,提取出对表情分类有用的表情特征。

在人脸表情图像中,表情的改变往往会带动着眼睛、鼻子、嘴巴等关键点也发生位置上的改变,而反过来,这些关键点的变化也能够用来判断人脸表情,因此表情识别模型可以通过人脸关键点来进行识别。然而目前的表情研究大多是对图像进行处理后直接提取出全局或局部特征,然后对表情图像进行分类,少数的表情研究通过提取关键区域的特征进行表情识别:在处理繁杂表情图像时运用表情类别信息来协助进行关键点的定位,然后根据关键点提取周边的关键特征,虽然这些方法都取得了不错的识别效果,但是它们往往是通过添加先验信息的方式来建立关键点检测和表情识别之间的联系,对数据质量要求较高且有很大的人为干预误差。

针对这一问题,本文设计了一个双流网络,在完成粗特征表情识别的同时提取关键点,使用关键点生成注意力图增大关键点周边特征的权重,减小干扰区域的特征响应值。随后以残差网络为基础模型,改进了残差块之间的跳跃连接方式,并将残差块中的普通卷积改进为分组卷积来强化模型的特征提取能力,完成关键点特征的融合。最后使用 SVM 训练两个网络的 softmax 分数得出最终识别结果。在 CK + 、JAFFE、Fer2013 这三个公开数据集上验证了本文方法的有效性。本文的创新点为:

(1) 设计了一种新的双流网络完成关键点检测和粗特征表情识别,并使用注意力机制更加准确地提取关键点周边的特征以减小干扰区域的特征响应值,有效地提取了图像的粗特征和细特征。

(2) 在残差网络的基础上,改进了残差块之间的跳跃连接方式并使用分层卷积替换原有的普通卷积来进一步强化模型的关键点特征提取能力。

1　相关工作

近年来,人脸表情识别受到了人们广泛的关注,研究者们逐渐发现人脸表情识别的细节主要集中于眼睛、眉毛、嘴巴等关键区域中,局部关键区域特征表情识别逐渐成为研究的热点。

杨飚等[1]在 VGGNet 的基础上,利用 WMDNN 对两通道人脸图像进行处理,包括人脸灰度图像和相应的局部二值模式(LBP)人脸图像。然后通过对局部 VGG16 网络进行微调,提取人脸灰度图像的表情相关特征,并利用 ImageNet 数据库训练的 VGG16 模型初始化其参数。最后两个通道的输出以加权方式融合并利用 softmax 分类计算最终识别结果。李勇等[2]使用一种改进的 LeNet-5 模型有效地提取并融合表情图像的高层次特征和低层次特征,实验结果准确率有明显提升。王琳琳等[3]认为局部关键点特征是表情识别的关键,因此基于深度置信网络(DBN)提出了一种融合局部特征的表情识别方法。该方法提取眼睛和嘴巴的特征,对其分别提取 HOG 特征,融合后进行分类,获得了不错的效果。Ozbey 等[4]同样认为关键点特征与表情识别具有很大的关联性,在关键点周围提取了 LBP 特征用于表情识别。Munasinghe 等[5]认为在表情变化时,人脸关键点之间的距离也会对应地发生变化,因此根据特殊关键点之间的距离作为识别特征,且使用随机森林法来提取距离特征完成最终分类。孔英会等[6]使用关键点检测技术识别出人脸关键子区域并提取出关键区域特征信息,并设计了一种改进的 LGC 算子充分地描述局部区域形变后进行表情分类识别。

残差网络在卷积神经网络的基础上使用了相当深的深度来提高准确性,并缓解了在深度神经网络中增加深度带来的梯度消失问题,由此也受到了表情识别研究者们的关注。

杜进等[7]结合残差结构的思想,提出了一种基于改进残差网络的表情识别方法,引入具有生物真实性的激活函数来替代已有的整流线性单元函数,并将其作为卷积层激活函数对深度残差网络进行了改进,最后取得了较高的识别率结果,但是该方法没有针对面部表情识别的特点,对其神经元本身的模型参数进行研究和分析,过于注重降低能耗而忽视了系统的准确性和识别速度。钱勇生等[8]在残差网络的基础上进行改进,提取了多个角度下的人脸表情特征,使用可分离的卷积替代普通卷积以达到减少计算量的目的,但该方法需要大量的不同角度表情图像,在时效性和实用性上略有不足。

2　模型设计与实现

2.1　模型总体架构

在图像识别领域中,常用的图像特征包括整体特征和局部特征,其中,整体特征有良好的不变性,表示起来更加直观,但是计算量偏大。局部细特征的计

算量小,在遮挡的情况下也不会因为部分特征的消失而影响其他特征的检测匹配,但特征提取需要更多时间。

本文在关键点表情识别中,为了同时兼顾关键点特征与整体特征,以增强特征的区分力与抗噪性能,设计了一个双流网络模型,模型架构如图 1 所示,双流网络结构分别是关键点检测网络和粗表情识别网络,而图像在此之前分别经过了多个卷积和池化,这样使得网络可以更集中于关注人脸的关键特征,随后将检测到的关键点执行高斯分布生成关键点注意力图,辅助增大图像关键特征权重。使用经改进后的残差网络进行特征的提取和融合,该网络模型充分地利用了卷积网络层与层之间的连接的同时,在不增加计算量的前提下更好地提取出人脸表情的细微特征,对融合后的关键点特征进一步提取更高层次的特征并降维后得出关键点表情识别结果。最后结合上粗表情识别网络的识别结果用于表情分类。

其中,改进残差网络中的残差块具体结构如图 2 所示,首先使用分组卷积代替传统卷积,将网络块的输入先进行拆分,再卷积操作,最后相加并输入到下一个块中,在不增加参数量和运算量的前提下增加特征图数量。同时改进了残差块之间的跳跃连接方式,将前一个的残差块状态直接连接到这一层的末尾并相加后输出,从而形成连续记忆机制,更好地利用前后残差块的特征信息。

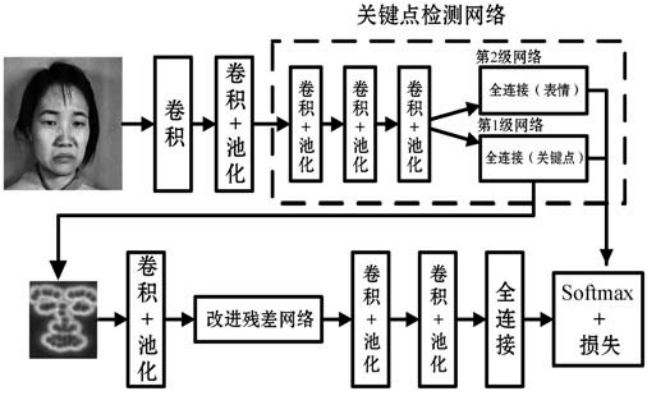

图 1 人脸表情识别模型

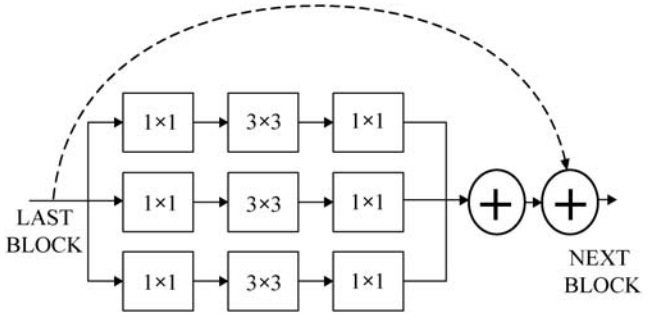

图 2 改进残差块

2.2 算法设计及网络结构

本文提出的基于注意力改进残差网络结构的表情识别模型具体算法设计如算法 1 所示。

算法 1 本文算法

输入:人脸图像。

输出:人脸表情识别结果。

1. 输入人脸图像
2. 使用 DTP 动态任务优先级算法分别为关键点检测网络和表情识别网络设置权重
3. 通过双流网络中的关键点检测网络检测关键点并提取特征,同时在 Hourglass 结构的监督下进行关键点优化
4. 选取检测的关键点中合适的值作为均值执行正态分布,生成置信图
5. 依据置信图做最大值池化,生成关键点注意力图并与同尺寸的特征图相乘,增大其区域内的关键点特征权重
6. 将提取到的特征作为输入,传入改进残差网络中,最大程度地利用每层的特征信息
7. 将不同层的特征提取出来,使其在全连接时与网络的输出层进行融合
8. 通过卷积和池化操作提取出关键点特征中更深层次的特征信息后降维
9. 得出关键点表情识别结果参数后结合上粗特征表情识别结果参数,使用 SVM 训练两个 softmax 分数得出最终识别结果

2.3 算法主要实现

(1) 关键点注意力机制。本文的目的主要是通过关键点的特征结合上全局粗特征进行表情识别,因此需要借助注意力机制[9]来辅助完成。人类观察图像时首先会快速浏览整体图像,并识别到需要集中关注的焦点区域,再对焦点区域投入更多注意力资源,以获得更多的图像细节并弱化其他无用区域的干扰信息。在本文方法中,如图 1 所示,双流网络中的关键点检测网络和表情识别网络之前进行多次卷积池化操作,以减小干扰区域的特征响应值,使整个模型更加关注关键点区域周边的特征信息。其中最开始的两个卷积设计成 5×5 和 3×3。使用 5×5 卷积核是为了融合输入图像中不同区域的特征信息。又因为要防止关键点提取步骤中卷积层数过多导致的网络过拟合[10]的问题,所以后面的卷积都是设计采用的 3×3 和 1×1 的卷积核。

在关键点特征提取中,本文的目的是尽可能地减少特征提取的范围,以达到只提取少量特征就能最大化检测出人脸表情的目的,使网络模型的效率最大化。因此本文利用提取的关键点生成对应的位置注意力图来达到我们的目的,流程如图 3 所示。

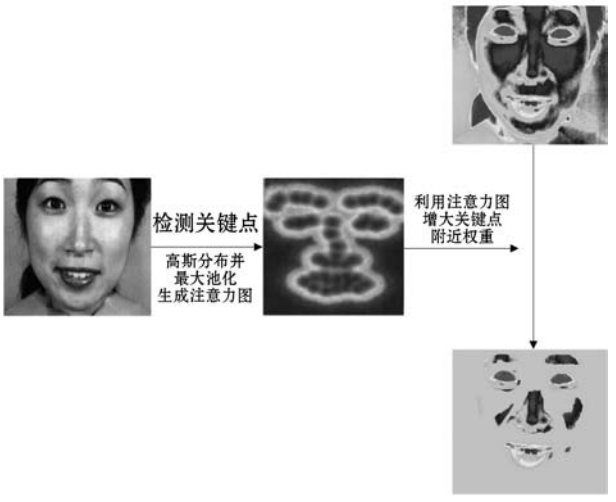

图 3　关键点注意力实现流程

当关键点检测网络获得表情关键点后,将关键点中合适的值作为方差进行高斯分布,生成置信图,其中关键点周围区域具有较高权重,而除此之外的非关键点区域则权重较低,在得到相应的权重矩阵后,将其尺寸调整到和特征图相应的大小并与之相乘,最后进行最大值池化得到关键点注意力图以辅助进行特征提取,减少其他噪声特征的干扰。

(2) 分组卷积。在本文中,为了在不增加计算量的前提下,进一步提高网络的特征提取能力,将普通残差块中的卷积改成分组卷积[11]。一般的卷积如图 4(a)所示,输入特征图的尺寸大小为 $W \times H \times C$,分别对应特征图的宽高和通道数,单个卷积核的尺寸为 $k \times k \times C$ 分别是单个卷积核的宽、高、通道数,输出特征图的尺寸为 $W' \times H'$,输出通道数等于卷积核数量,输出的宽、高与卷积步长相关。一般卷积的参数量(Params)和运算量(FLOPs)分别是:

$$P_{\text{params}} = k^2 C \tag{1}$$

$$FLOPs = k^2 CW'H' \tag{2}$$

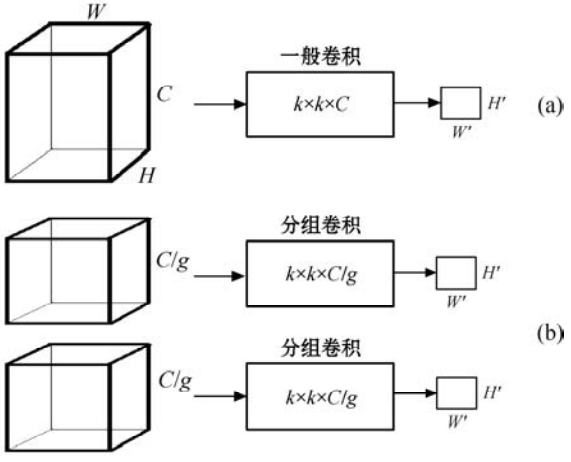

图 4　一般卷积与分组卷积流程

分组卷积(Group Convolution),是对输入的特征图进行分组,然后对每组分别进行卷积,如图 4(b)所示。输入特征图的尺寸大小为 $W \times H \times C/g$,单个卷积核的尺寸为 $k \times k \times C/g$,g 是单个卷积核被分为的组数,输出特征图的尺寸为 $W' \times H' \times g$,共生成 g 个特征图。

由计算可得分组卷积的参数量和运算量分别为:

$$P_{\text{params}} = k^2 \times \frac{C}{g} \times g = k^2 C \tag{3}$$

$$FLOPs = k^2 \times \frac{C}{g} \times W' \times H' \times g = k^2 CW'H' \tag{4}$$

由式(4)、式(5)可知,尽管分组卷积分成了 g 个特征图,但是它的参数量和运算量和普通的卷积相同。因此,在同等条件下,使用分组卷积可以生成大量的特征图,即能够编码更多信息,强化模型的特征提取能力,让残差块可以提取更多的表情细节信息,最终实现不影响识别速度而提高识别准确率的表情识别网络。

3　实　验

本文是在 JAFFE、CK + 、Fer2013 这 3 个公开数据集上进行测试,CK + 数据集包含了 123 个人,共 593 个序列,CK + 数据集除了中性外还包含 7 种表情:愤怒、蔑视、厌恶、恐惧、高兴、悲伤和惊讶。除此之外该数据集还标出了每幅人脸图像中关键点的坐标,因此非常适合本模型,另外两个数据集则是没有。

在实验过程中,由于 CK + 包含关键点坐标,因此将 CK + 数据集通过旋转、调整分辨率来扩大样本数。并将 CK + 中不同表情类别、不同人的表情样本数量平均分配,目的是提高在训练过程中模型对不同类型、不同数据集的训练时的鲁棒性,最后用于训练 Hourglass 关键点检测网络。

由于 JAFFE 和 Fer2013 数据集中未标记出人脸关键点坐标。因此首先使用多任务卷积神经网络(MTCNN)[12]标记人脸区域,然后使用的 Hourglass 网络给每个人脸图像进行关键点坐标值定位,以作为本文表情识别网络中关键点检测任务的真实值,Hourglass 网络架构有效地利用了图像多个尺度的空间信息,可以很好地应用于人脸关键点检测任务。

3.1　实验过程

首先使用 DTP 动态任务优先级算法[12]给关键点检测网络和表情识别网络计算了权重,并将输入图像分别从图片左上、左下、右上、右下角以及图像的中间直接裁剪成 48 × 48 的尺寸,根据关键点生成同样是 48 × 48 尺寸的关键点注意力图,正态分布的标准差预设为 4 像素,将生成的注意力图的尺度调整到 12 × 12 像素,并与特征图相乘减少其他噪声特征的干扰。

DTP 计算得出关键点检测任务的最佳权值范围在 0.08 ~ 0.15 附近,因此需要经过实验对比得出最终系数值,图 5 是调整关键点检测任务加权系数得到的平均识别率。

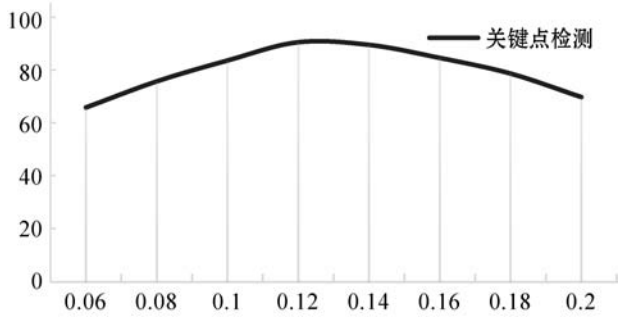

图 5 任务权重实验图

图 5 中横坐标表示加权系数值,纵坐标表示相对应的权重系数达到的平均识别率。从图 5 可知,一开始随着关键点任务加权系数的增加,关键点检测愈发精准,表情识别效率大大增加,识别率在关键点任务权重 0.12 时达到了顶峰,随后随着权重增大,整体识别率开始下滑,因此设置关键点检测网络的权重为 0.12,相应地将表情识别网络设置为 0.88。

本文提出的注意力改进残差网络模型在数据集上训练过程中损失如图 6 所示,其中损失函数为交叉熵损失。

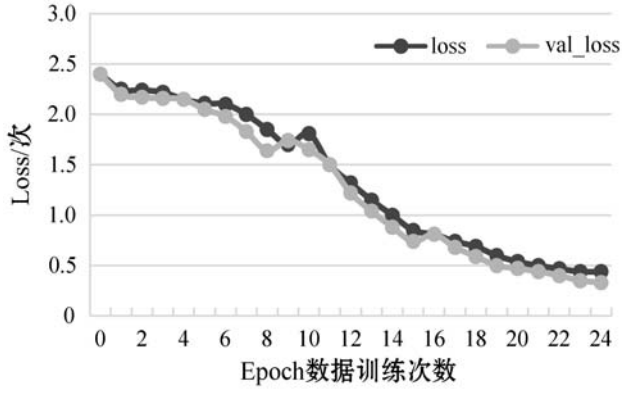

图 6 训练损失

3.2 实验结果分析

在使用 Hourglass 网络后,为了验证使用关键点生成注意力图后指导网络识别的有效性,将该方法与检测到关键点后立即进行关键点特征提取进行对比实验。实验结果如表 1 所示。

表 1 三个选定数据集在不同通道数的模型中的性能精度比较

模型	1 000 幅测试图耗时/s	准确率/%
注意力图指导	53.7	92
直接提取特征	53.5	87

使用关键点生成注意力图后指导网络识别在耗时上和直接提取关键点特征相近,但在识别准确率上有明显的提升,其原因是关键点周围的特征被更有效地利用了。

最后将本文提出的注意力改进残差网络表情识别方法与各种深度学习模型相比。

多任务 + SVM:Wang 等[13]提出的一种传统方法的表情识别模型,该模型使用主动形状模型(ASM)算法检测出关键点坐标后,提取其周边的局部纹理特征,该方法有效地建立了关键点与表情分类之间的联系。

DBN + 局部特征融合:王琳琳等[3]基于深度置信网络(DBN)提出的一种局部特征融合的表情识别方法。该方法着重提取了眼睛和嘴巴的特征,对其分别提取 Log-Gabor 特征与二阶梯度方向直方图特征,融合后进行分类,获得了不错的效果。

VGG + 双通道:Yang 等[1]在 VGGNet 的基础上提出的一种表情识别方法,利用 WMDNN 处理人脸灰度图像和相应的局部二值模式(LBP)人脸图像。然后提取人脸灰度图像的表情相关特征,利用 ImageNet 数据库训练的 VGG16 模型初始化其参数。最后完成表情分类。

深度 ResNet:卢官明等[14]基于残差网络的思想,提出的一种基于深度残差网络的人脸表情识别方法。该方法利用残差学习单元来改善深度 CNN 模型训练寻优的过程,提高了模型的准确率并减少收敛的时间开销。

低功耗 ResNet:杜进等[7]提出的一种改进残差网络的表情识别方法,该方法将网络中整流线性单元函数替换成具有生物真实性的激活函数,并将其作为卷积层激活函数对深度残差网络进行了改进,最后取得了较高的识别率结果。

实验结果如表 2 所示,可以看出本文的方法相较于其他识别方法,在 JAFFE、CK + 、Fer2013 这三个主流人脸表情数据库前,注意力改进残差网络模型的识别率都要优于各种深度学习单通道模型,由此验证了本文方法的有效性。

表 2 与其他表情识别方法对比(%)

识别方法	JAFFE	CK +	Fer2013
多任务 + SVM	88.73	93.68	69.42
DBN + 特征融合	92.2	97.1	71.1
VGG + 双通道	93.8	97.62	72.7

续表 2

识别方法	JAFFE	CK +	Fer2013
深度 ResNet	91.32	92.14	72.21
低功耗 ResNet	88.75	91.51	68.11
本文方法	**95.61**	**97.75**	**73.6**

4　结　语

本文在考虑到关键点与表情识别的关联性后,提出了一种基于注意力改进残差网络的表情识别方法,通过双流网络提取关键点并进行粗特征表情识别,利用关键点生成注意力图增大关键点周边特征的权重、减小干扰区域的特征响应值。以残差网络为基础模型,改进了残差块之间的跳跃连接方式,并使用分组卷积来替代残差块中的普通卷积,以更好地提取关键点特征。最后完成特征融合和表情识别。在公开数据集上,将该方法与其他主流方法进行对比研究,结果表明其准确率有较大提高。

参　考　文　献

[1] Yang B, Cao J M, Ni R, et al. Facial expression recognition using weighted mixture deep neural network based on double-channel facial images[J]. IEEE Access,2018,6:4630 – 4640.

[2] 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.

[3] 王琳琳,刘敬浩,付晓梅. 融合局部特征与深度置信网络的人脸表情识别[J]. 激光与光电子学进展,2018,55(1):196 – 204.

[4] Ozbey N, Topal C. Expression recognition with appearance-based features of facial landmarks[C]//26th Signal Processing and Communications Applications Conference, 2018: 1 – 4.

[5] Munasinghe M I. Facial expression recognition using facial landmarks and random forest classifier[C]//17th International Conference on Computer and Information Science, 2018:423 – 427.

[6] 孔英会,陈咨彤,车辚辚. 基于关键子区域及特征提取的表情识别[J]. 科学技术与工程,2017,17(34):257 – 262.

[7] 杜进,陈云华,张灵,等. 基于改进深度残差网络的低功耗表情识别[J]. 计算机科学,2018,45(9):303 – 307.

[8] 钱勇生,邵洁,季欣欣,等. 基于改进卷积神经网络的多视角人脸表情识别[J]. 计算机工程与应用,2018,54(24):12 – 19.

[9] Bilen H, Fernando B, Gavves E, et al. Action recognition with dynamic image networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 40(12): 2799 – 2813.

[10] Xu Q, Zhang M, Gu Z H, et al. Overfitting remedy by sparsifying regularization on fully-connected layers of CNNs[J]. Neurocomputing,2019,328(7):69 – 74.

[11] Wang X J, Kan M N, Shan S G, et al. Fully learnable group convolution for acceleration of deep neural networks [C]//IEEE Conference on Computer Vision and Pattern Recognition,2020:9041 – 9050.

[12] 龚跃,张真真,黄小珂,等. 基于动态双向优先级的任务分配与调度算法[J]. 计算机应用,2009(4):1131 – 1134.

[13] Wang X, Liu X G. Learning the discriminate patches from the key landmarks for facial expression recognition[C]// IEEE International Conference on Smart City/SocialCom/SustainCom,2015:345 – 348.

[14] 卢官明,朱海锐,郝强,等. 基于深度残差网络的人脸表情识别[J]. 数据采集与处理,2019,34(1):50 – 57.

(上接第 161 页)

[14] Wang J W, Ding J, Guo H W, et al. Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images[J]. Remote Sensing,2019,11(24):2930 – 2951.

[15] 刘凤,李华,南方哲,等. 优化特征提取的多目标交通标志检测方法[J]. 计算机工程与设计,2021,42(5):425 – 432.

[16] Liu F, Qian Y R, Li H, et al. CAFFNet: Channel attention and feature fusion network for multi-target traffic sign detection[J]. International Journal of Pattern Recognition and Artificial Intelligence,2021,35(7):2152008.

[17] Fu C Y, Liu W, Ranga A, et al. DSSD: Deconvolutional single shot detector[EB]. arXiv:1701.06659,2017.

[18] Li Z X, Yang L, Zhou F Q. FSSD: Feature fusion single shot multibox detector[EB]. arXiv:1712.00960,2017.

[19] Cheng G, Zhou P C, Han J W. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing,2016,54(12):7405 – 7415.

[20] Lin T Y, Dollar P, Girshick R, et al. Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition,2017:2117 – 2125.

[21] Woo S, Park J C, Lee J Y, et al. CBAM: Convolutional block attention module[C]//European Conference on Computer Vision,2018:3 – 19.

[22] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//IEEE International Conference on Computer Vision,2017:2999 – 3007.