基于通道注意的可变形金字塔表情识别网络

叶耀光[1],陈宗楠[1],陈丽群[1],潘永琪[2],潘家辉[1,3]
(1. 华南师范大学 软件学院,广东 佛山 528225;
2. 华南农业大学 数学与信息学院 软件学院,广东 广州 510642;
3. 琶洲实验室,广东 广州 510320)

摘 要:人脸表情识别是计算机视觉领域里一项热门且有挑战性的任务。由于人脸表情的特性相对固定,从宏观上针对人脸表情的特性进行方法设计能有效提高人脸表情识别的性能。基于这一观点,针对人脸表情特征的不规则特性和空间多尺度互补特性,提出了基于通道注意的可变形金字塔网络。该网络主要由可变形卷积块、空间金字塔池化块和通道注意块构成,其中可变形卷积块有助于网络对人脸表情的不规则特征进行采样;而空间金字塔池化块则加强了网络学习多尺度空间上下文情绪信息的能力;通道注意块则促使网络动态关注更具判别性的情绪特征。该方法在 CK+、JAFFE 以及 Oulu-CASIA 三个实验室环境的人脸表情数据集和 FER2013 以及 RAF-DB 两个野外环境的人脸表情数据集上进行了对比实验和消融实验并取得了有竞争力的结果。从可视化结果上看,该方法提取的特征及关注的人脸区域符合不同表情的呈现特性和人们日常判断表情的规律。

关键词:人脸表情识别;卷积神经网络;可变形卷积;金字塔架构;注意力机制
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1673-629X(2022)11-0064-08
doi:10.3969/j.issn.1673-629X.2022.11.010

Channel-attention-based Deformable Pyramid Network for Facial Expression Recognition

YE Yao-guang[1],CHEN Zong-nan[1],CHEN Li-qun[1],PAN Yong-qi[2],PAN Jia-hui[1,3]
(1. School of Software,South China Normal University,Foshan 528225,China;
2. School of Mathematics and Informatics College of Software Engineering,South China Agricultural University,Guangzhou 510642,China;
3. Lab of Pazhou,Guangzhou 510320,China)

Abstract:Facial expression recognition is a popular and challenging task in the field of computer vision. Since the characteristics of facial expressions are relatively fixed,designing methods for the characteristics of facial expressions from a macro perspective can effectively improve the performance of facial expression recognition. Based on this view,a deformable pyramid network based on channel attention is proposed for the irregular characteristics of facial expression features and the complementarity of spatial multi-scale. The network mainly consists of deformable convolutional blocks,spatial pyramid pooling blocks and channel attention blocks. Specifically,the deformable convolution block helps the network to sample irregular features of facial expressions,while the spatial pyramid pooling block enhances the network's ability to learn multi-scale spatial contextual emotion information,and the channel attention block motivates the network to dynamically focus on more discriminative feature maps to improve the contribution of discriminative emotion information to the facial expression recognition task. Experimental results on both three in-the-lab facial expression datasets (including CK+,JAFFE,and Oulu-CASIA) and two in-the-wild facial expression datasets (including FER2013 and RAF-DB) demonstrate the effectiveness of the proposed method. From the visualization results,the features extracted by the proposed method and the facial regions of interest are consistent with the presentation characteristics of different expressions and the patterns of people´s daily judgment of expressions.

Key words:facial expression recognition;convolution neural network;deformable convolution;pyramid structure;attention mechanism

收稿日期:2022-06-29 **修回日期**:2022-08-30
基金项目:国家自然科学基金面上项目(62076103);广州市重点领域研发计划(202007030005);科技创新 2030-"脑科学与类脑研究"重点项目(2022ZD0208900)
作者简介:叶耀光(1998-),男,硕士生,研究方向为情感计算、计算机视觉、脑机接口;通讯作者:潘家辉(1982-),男,教授,博士,研究方向为脑机接口、模式识别与智能系统。

0　引　言

在现实生活中,人们常常需要交流各自的情绪信息来表达对某种事物的态度。人脸表情作为一种易直接观察、易理解的情绪交流媒介,往往是人与人之间交流和人与机器之间交互的一把利器。近年来,随着人们对于自身情绪心理健康的关注意识逐渐觉醒,同时由于人机交互场景提出的更高级交互需求,即情感交互,情绪识别技术逐渐成为新的研究热点。而人脸表情识别技术更是其中的重头戏,被广泛用于医疗诊断、人机交互、危险驾驶检测、公共安全等领域。

人脸表情的类型往往是复杂多样的,其中,复合表情之间所表达的情绪存在交叉,准确地判断出复合表情是极为困难且不现实的。考虑到基本表情的分类理解、使用频率和呈现规律对于全人类而言是基本一致的,目前人脸表情识别的相关任务一般采用的是心理学家 Ekman 提出的六分类基本离散表情模型[1]。该模型定义了六种基本情绪类别,即愤怒、厌恶、恐惧、愉快、悲伤和惊讶,部分数据集还会标注额外的类别,比如 CK+[2] 数据集还包含平静和轻蔑这两类标注数据。

在人脸表情识别领域中,传统方法使用预先设计的手工特征来表示人脸表情信息,但手工特征的表现模式较为固定,其稳定性和适用性较差,图像中的照明、旋转和噪声在内的变化可能会削弱这种手工制作的特征描述符的能力,往往不能完整地表示人脸表情中的关键情绪信息。于是,深度学习方法逐渐成为人脸表情识别任务的主流解决方案。

事实上,将深度学习应用于人脸表情识别任务中存在一个关键问题,即如何对神经网络基于人脸表情识别任务进行适应性和针对性的设计,让神经网络学习到更具代表性和更符合人脸表情特性的判别性特征。针对这一问题有如下三点思考:(1)考虑到人脸表情中呈现的特征形状往往是不规则的,若能够让神经网络根据人脸特性采样不规则特征,能够有效提高人脸表情识别网络的性能;(2)考虑到不同尺度的人脸特征所包含的情绪信息不尽相同,若能够利用多个尺度特征图的情绪信息进行融合学习,在提高网络尺度不变性的同时,加强网络学习上下文情绪信息的能力;(3)考虑到人脸表情的不同特征图往往具有不同程度的重要性,若能让网络自适应关注关键特征图,有助于提高网络对于判别性特征的提取能力。基于上述三点思考,该文提出了一个基于通道注意的可变形金字塔网络,该网络以 ResNet50 作为原始网络模型,通过将一般二维卷积变换为可变形卷积结构,达到加强网络学习不规则特征能力的目的;通过在 ResNet50 的基本瓶颈块后嵌入空间金字塔池化块对多尺寸特征图进行采样,有助于网络构建互补上下文语义信息;通过在网络中嵌入通道注意块,进一步加强网络对不同语义情绪信息的关注能力。该文的贡献如下:

(1)基于人脸表情的呈现特性,提出了一种可变形金字塔网络。该网络从加强不规则特征和多尺度特征的情绪信息提取能力两方面出发,提高人脸表情特征的判别性情绪信息的含量和质量,进而提高其进行情绪识别的精确性。

(2)根据可变形金字塔网络特性,对应设计一个通道注意块并嵌入于金字塔块之后,减弱多语义多尺度特征冗余信息的干扰,有效加强网络对于判别性特征的学习能力。

(3)在 5 个人脸表情识别数据集上开展广泛的对比实验以评估网络的性能,结果显示该网络存在着较强的竞争力。此外还进行了详细的消融实验和结果可视化,以验证各模块的有效性。

1　相关工作

卷积神经网络广泛应用于图像识别领域,是人脸表情识别任务的一种重要解决方案,如 VGGNet 和残差神经网络 ResNet 等经典网络都曾被用于提取深层面部表征,并在人脸表情识别任务中取得不错的结果[3-4]。然而,若要进一步提高模型用于人脸表情识别任务的性能,还需要基于人脸表情识别任务进行适应性改进。Yang 等人[5] 提出了 DeRL 方法,通过为每个表情人脸重建一个中性人脸来结合情绪对比信息,提取更丰富的表情信息。邓楚婕[6] 通过将面部动作单元数据作为第二类输入数据,使用同一个网络完成人脸表情识别和面部动作单元识别两个任务,进而使得网络对于微弱表情的识别更准确。郑豪等人[7] 同时考虑了样本的情感标签信息和局部空间分布信息,提出了一种用于人脸表情识别的判别式多任务学习方法。王珏[8] 通过超分辨率技术获得包含更多细节的高分辨率表情图像,再进一步进行人脸表情识别。

注意力机制也是一种行之有效的通用方法,它们广泛应用于人脸表情识别任务中来提高网络对于表情特征的辨别能力。Farzaneh 等人[9] 提出了一种深度专注的中心丢失方法,他们通过使用 CNN 中的中间空间特征图作为上下文来估计与特征重要性相关的注意权重。这些基于注意力的混合模型缓解了卷积滤波器在学习远程感应偏差方面的弱点,并扩大了感受野。

根据上述研究情况,可以发现目前提出的人脸表情识别方法已经针对人脸表情识别任务设计许多适应性改进,然而,却少有研究关注不规则人脸表情特征和多尺度人脸表情特征对于人脸表情识别任务的影响。该文设计了一个基于通道注意的可变形金字塔网络,该网络通过可变形卷积结构学习更符合人脸表情的不

规则特征;通过金字塔架构学习多尺度人脸表情特征所含情绪信息的互补性;通过通道注意块关注不同特征图之间的重要性,降低特征图之间情绪信息的冗余度。此外,还在训练阶段采用 Softmax 损失函数和中心损失函数混合而成的损失函数,以减小不同表情类别的类内变化,增大类间变化。

2 基于通道注意的可变形金字塔网络

该文综合考虑了人脸表情的表现特性,结合可变形卷积、金字塔操作和通道注意的优点,提出了一种基于通道注意的可变形金字塔网络,用于完成人脸表情识别任务。该网络的架构和整体处理流程如图 1 所示。首先会对原始图像进行预处理操作,得到相应的人脸图像,作为特征提取器的输入。而特征提取器以 ResNet50 网络作为基线方法,通过嵌入多个可变形卷积核加强网络学习不规则特征的能力,通过多次运用金字塔架构加强网络对不同尺度特征信息的挖掘能力。经过特征提取器进行特征学习后,得到多张人脸表情特征图,由全连接层进行特征信息融合,由 Softmax 函数计算情绪分类概率,得到最终情绪识别结果。

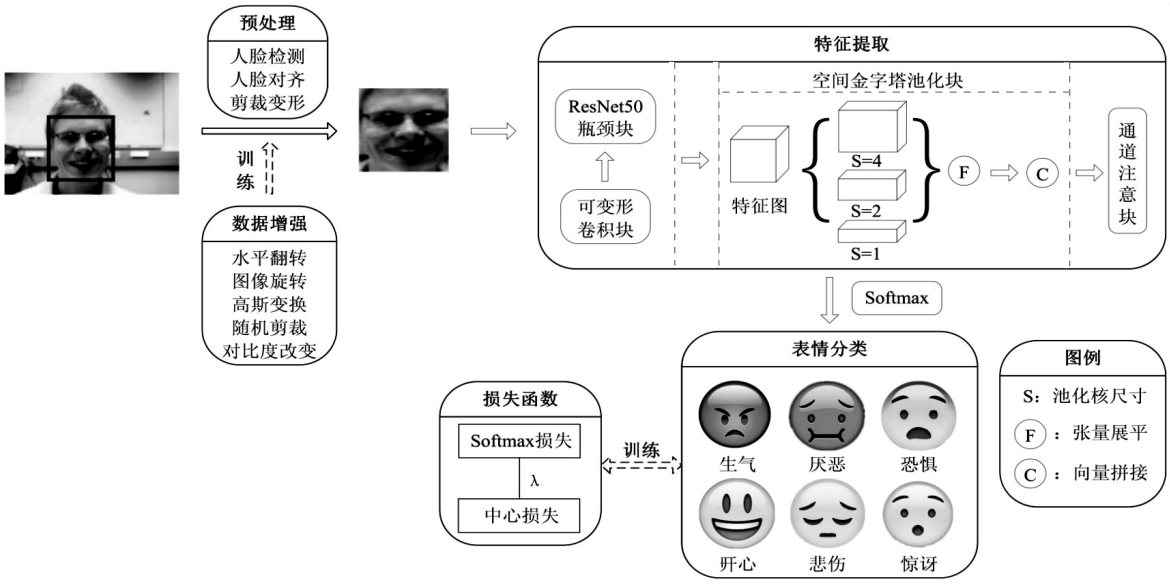


图 1 基于通道注意的可变形金字塔网络架构

2.1 可变形卷积块

人类的视觉认知是由物体的结构因素引导的,通过在卷积神经网络中嵌入可变形卷积模块,可以使卷积神经网络自适应地生成更明确的空间特征表示。可变形卷积的概念首先由代季峰等人[10]提出,他们证明这种方法在目标检测和语义分割任务中是有效的,但是却鲜有研究者验证其在人脸表情识别任务中的有效性。该文通过可变形卷积模块,分离不同表情之间的结构信息,并使网络能够更好地捕获上下文依赖关系。

具体地,在传统 3 × 3 卷积核中,可以将卷积操作视为如下规则采样网格:

$$\mathbb{G}=\begin{Bmatrix}(-1,1) & (0,1) & (1,1)\\ (-1,0) & (0,0) & (1,0)\\ (-1,-1) & (0,-1) & (1,-1)\end{Bmatrix}$$

每个位置的卷积输出值通过采样点的加权和求出:

$$\boldsymbol{y}(\boldsymbol{p}_0)=\sum_{\boldsymbol{p}_n\in\mathbb{G}}\boldsymbol{w}(\boldsymbol{p}_n)\cdot\boldsymbol{x}(\boldsymbol{p}_0+\boldsymbol{p}_n)$$

其中,$\boldsymbol{p}_0$ 表示当前采样点,$\boldsymbol{p}_n$($n\in(1,N)$,N 为采样点个数 $|\mathbb{G}|$)是对规则采样网格 $\mathbb{G}$ 每个采样点位置的枚举。而可变形卷积通过给传统的卷积核中的每个采样点添加一个二维偏移量来达到变形的目的,即对于每个采样位置,其经过可变形卷积的计算公式为:

$$\boldsymbol{y}(\boldsymbol{p}_0)=\sum_{\boldsymbol{p}_n\in\mathbb{G}}\boldsymbol{w}(\boldsymbol{p}_n)\cdot\boldsymbol{x}(\boldsymbol{p}_0+\boldsymbol{p}_n+\Delta\boldsymbol{p}_n)$$

其中,$\Delta\boldsymbol{p}_n$ 即为设置的二维偏移量,这个二维偏移量作为一个额外的参数,需要通过额外的卷积层根据输入图像的特征图进行自适应学习,如图 2 所示。

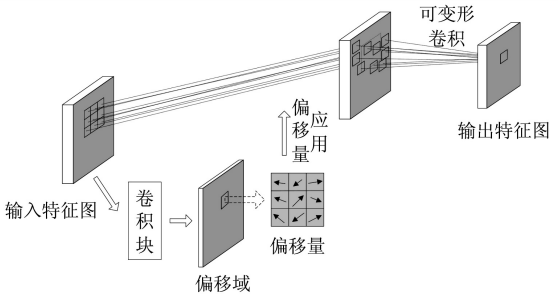


图 2 可变形卷积结构及处理流程

2.2 空间金字塔池化块

池化操作可以聚合输入的特征图的上下文信息,形成特定大小的特征图。而多尺度特征在传统方法中

起着重要作用,例如,尺度不变特征变换算法通常在多个尺度上提取特征[11]。该文希望结合池化和多尺度特征的优势,学习到多尺度的人脸表情空间上下文特征。具体地,在 ResNet50 的每个基本瓶颈块后添加一个空间金字塔池化块,其结构及处理过程如图 3 上半部分所示。该块设置 3 个独立的池化层,其池化大小分别为 1、2、4。基本瓶颈块处理后得到的特征图 $M_o \in \mathbb{R}^{C\times H\times W}$ 将分别由这 3 个不同大小的独立池化层进行池化操作,对应得到 3 张不同大小的池化特征图,每张特征图将分别展平成一维向量,并沿着通道维度进行拼接,由通道注意块进行进一步的处理。空间金字塔池化块的公式可表示如下:

$$M_{\mathrm{SPP}} = \mathrm{FlaCon}(\mathrm{PA}_{n\times n}(M_o))$$

其中, M_{SPP} 表示空间金字塔池化块输出的特征图, M_o 表示基本瓶颈块处理后得到的特征图, FlaCon 表示展平和拼接操作, $\mathrm{PA}_{n\times n}$ 表示池化大小为 n 的平均池化操作,文中 $n \in \{1,2,4\}$ 。

2.3　通道注意块

考虑到空间金字塔池化块提取到的多尺度特征之间的重要性也有所差异,为了确保网络捕捉到更有判别性的表情信息特征,通过注意力机制[12]来探索特征通道之间的相互依赖关系。通道注意块的结构和处理流程如图 3 下半部分所示,其以空间金字塔池化块的输出作为全局通道注意信息,输入至通道注意块中。通道注意块使用两个分别具有 ReLU 和 Sigmoid 激活函数的全连接层来学习通道之间的线性和非线性信息,捕获特征之间的通道依赖关系。

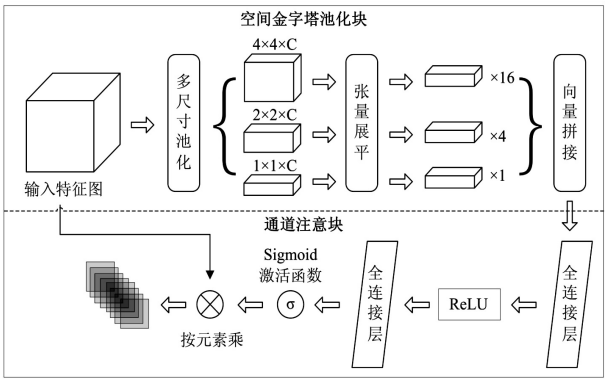


图 3　空间金字塔池化块和通道注意块的结构及处理流程

通道注意块的输出特征图公式如下:

$$M_C = \sigma(F_1(E(F_2(M_{\mathrm{SPP}}))))$$

其中, F 、E 和 σ 分别指全连接层、ReLU 激活函数和 Sigmoid 激活函数。F_1 旨在压缩输入特征的通道信息。在被 ReLU 激活函数激活后,这些特征随后通过卷积层 F_2 增加到原始通道数。该块的最终输出特征通过以下方式获得:

$$M_{\mathrm{out}} = M_C \otimes M_o$$

其中, $\otimes$ 表示按元素进行乘法操作。

3　实验结果与分析

3.1　数据集

该文使用 CK+[2]、JAFFE[13]、Oulu-CASIA[14]、FER2013 和 RAF-DB[15] 五个人脸表情识别数据集来验证该方法的有效性。这些数据集的详细信息如下所述。

CK+数据集是一个实验室环境的人脸表情数据集,包含来自 123 名受试者的 593 个图像序列,每个序列包含 15 张静态人脸图像。该数据集记录了受试者 6 种基本表情,部分受试者还记录了轻蔑和平静表情。在实验中,与大多数相关研究相同,即分别进行 6 种基本表情的分类实验和 7 种表情(6 种基本表情+轻蔑)的分类实验。在实验过程中均选择最后三帧来构建训练集和测试集。

JAFFE 数据集是一个实验室环境的人脸表情数据集,包含 213 张来自 10 名日本女性的照片。其中 30 张图像被标记为中性表情,而其他 183 张图片被标记为 6 种基本表情的其中一种。考虑到该数据集样本数量相对较少,应用数据增强增加数据量,以避免过拟合。

Oulu-CASIA 数据集是一个实验室环境的人脸表情数据集,共有 80 名受试者在可见光及正常照明条件下采集到的 480 个图像序列。每个序列都有 6 种基本表情中的一种。在实验过程中均选择最后三帧来构建训练集和测试集。

FER2013 数据集是一个野外环境的人脸表情数据集,其所含图像是从互联网上收集的,由 35 887 张大小为 48×48 的灰度人脸图像组成,每张图像被标记为 7 种表情(6 种基本表情+平静)中的一种。

RAF-DB 数据集是一个野外环境的人脸表情数据集,由 15 339 张人脸图像组成,这些图像是从各种搜索引擎收集的,包含 7 种表情分类(6 种基本表情+平静)。

3.2　环境、参数及实验设置

该文在训练模型时,使用随机梯度下降算法作为优化方法,初始学习率设置为 1e-5,批处理大小设置为 32。训练过程中使用早停法,设置一个最大平稳次数参数,其值为 6,在模型训练过程中,当模型在验证集上的误差比此前最好结果差时,平稳次数自增,若超过 6,则提前停止训练。设置一个平稳耐性参数,其值为 2,当平稳次数首次超过 2 时,学习率降低至原学习率的 1/10。

为了提高模型的鲁棒性,对数据较少或类别样本

不平衡的数据集进行数据增强:(1)对于输入的人脸图像,在尺寸范围 200 × 200 和 256 × 256 之间进行随机裁剪或填充,以削弱人脸图像中人脸的相对位置的影响;(2)以 50% 的概率水平翻转图像,减弱人脸姿态对结果的影响;(3)在一定范围内随机改变图像的亮度和对比度,减弱光照对结果的影响;(4)使用高斯模糊方法对图像进行模糊处理,减弱清晰度对结果的影响。

3.3 实验结果与对比

为了评估该文提出的基于通道注意的可变形金字塔网络的性能,首先在三个实验室控制的数据集(CK+、JAFFE、Oulu-CASIA)上进行实验。表 1 展示了其他研究和文中方法在这三个数据集上的实验设置及准确率。文中方法在 CK+数据集上进行了 6 分类和 7 分类实验,在 JAFFE 和 Oulu-CASIA 数据集上进行了 6 分类实验。

表 1 实验室环境数据集(CK+、JAFFE 和 Oulu-CASIA)上相关研究方法的实验结果对比

方法	数据	验证策略	不同数据集中达到的准确率/%		
			CK+	JAFFE	Oulu-CASIA
MSCNN-PHRNN[16]	图像序列	10 F	98.50(#7)	-	86.25
CAN[17]	图像序列	10 F	99.03(#6)	-	88.33
PPDN[18]	首帧图+1 峰值图	10 F	99.30(#6)	-	84.59
DeRL[5]	首帧图+1 峰值图	10 F	97.30(#7)	-	88.00
FER-net[19]	-	-	98.33(#7)	98.27	-
MFF-CNN[20]	3 峰值图	10 F	98.80(#6)	96.52	86.51
FDRL[21]	3 峰值图	10 F	99.54(#6)	-	88.26
DAM-CNN[22]	3 峰值图	10 F	95.88(#6)	99.32	-
ALFW[23]	3 峰值图	10 F	97.36(#7)	-	85.42
ALAW[23]	3 峰值图	10 F	97.11(#7)	-	85.83
DDL[24]	3 峰值图	10 F	99.16(#7)	98.00	88.26
文中方法	3 峰值图	10 F	99.33(#6) 99.16(#7)	99.23	89.27

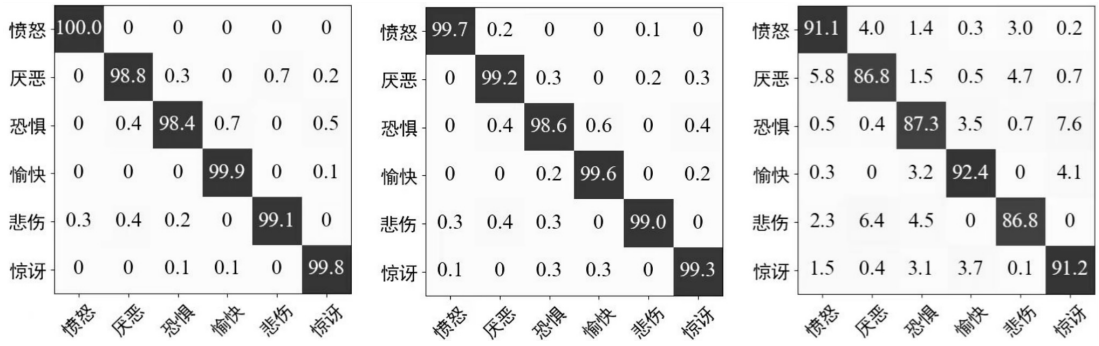
注:数据一栏中的“3 峰值图”指从图像序列中的末尾帧开始往前取 3 帧图像;验证策略中的“10 F”指十折交叉验证;CK+数据集结果中的标记“#6”指该研究获得的准确率对应 6 种基本表情分类任务,而“#7”指该研究获得的准确率对应基本表情和轻蔑 7 种表情分类任务,JAFFE 和 Oulu-CASIA 数据集进行的均是 6 种基本表情的分类任务。表中某些数据栏中的“-”符号表示该研究未说明该部分内容。

考虑到 CK+和 Oulu-CASIA 两个数据集含有数量较多的表情图像序列样本,这两个数据集也被广泛用于动态人脸表情识别任务中。而该文所提出的方法相较于一些动态表情识别方法在准确率方面仍不遑多让,这说明了符合人脸特性的表情特征对于人脸表情识别任务而言也是极为重要的,在一定程度上弥补了缺乏动态情绪信息的不足。综合表现较好的 DDL 方法在 CK+数据集上的准确率结果与文中方法相同,其通过多个子网络来提取表情特征和非表情特征等干扰特征,从而让模型集中在面部表情识别任务上。相对来说,DDL 方法复杂度较高,需要融合 4 个子网络的信息结果;且观察其网络的关注区域,发现其对于每种表情图像,注意力均集中于图像中的人脸五官,难以依据不同表情的特性进行动态关注,因此该文所提出的方法在另外两个实验室环境的数据集上的表现要优于 DDL。

对于 JAFFE 数据集,表现最好的研究是 DAM-CNN[22],其模型由一个用于提取特征的 VGG-Face 网络、用于细化 CNN 特征和突出显著表情区域的显著表情区域描述符和用于生成对多种高级表征的多径变异抑制网络构成。其中的显著表情区域描述符部分中的注意力遮罩层设置为 7 × 7 大小,再通过双线性池化方式上采样得到输入图像的关注热度图,相较于 DDL 方法而言具有更加粗的关注细粒度。该方式所产生的关注热度图虽然能让网络的关注位置集中于人脸部分,但却并不利于关注人脸表情变化时的微小肌肉变化,产生较多冗余信息,导致该方法在 CK+数据集上取得了一个较为落后的效果,而该文所提出的方法综合考虑了人脸表情图像中的特性,利用不规则特征和多尺度特征的情绪信息,加强网络对于多尺度特征的表征能力。

图 4 为文中方法在 CK+、JAFFE 和 Oulu-CASIA 三个实验室环境数据集中计算得到的混淆矩阵。总的来说,该方法在 CK+和 JAFFE 两个数据集中获得的整体准确率较高,因此对各种表情的识别也表现颇佳。而从 Oulu-CASIA 数据集的混淆矩阵中发现,该方法

对于每种表情的分类准确率较为接近,并未对某种情绪有明显的偏向性,这说明了该方法能较好地学习到6 种基本表情的判别性特征。



(a) CK+　　(b) JAFFE　　(c) Oulu-CASIA

图 4　在实验室环境数据集(CK+、JAFFE 和 Oulu-CASIA)上计算得到的混淆矩阵

为了进一步验证该文所提出的方法的有效性,还在 FER2013 和 RAF-DB 两个野外人脸表情数据集上开展了相关实验,其结果分别如表 2 和表 3 所示。

表 2　FER2013 数据集的相关方法及其准确率

方法	准确率/%
多视角图像+改进 CNN[4]	68.00
FER-net[19]	71.56
ALFW[23]	72.58
ALAW[23]	72.67
SDFL[25]	72.14
对齐和非对齐人脸信息融合[26]	73.30
文中方法	74.47

表 3　RAF-DB 数据集的相关方法及其准确率

方法	准确率/%
FER-net[19]	82.46
基于金字塔结构的卷积神经网络[6]	82.65
SCN[27]	87.01
SG-DSN[28]	87.13
DDL[24]	87.71
FDRL[21]	89.47
文中方法	88.91

由于 FER2013 存在图像质量差、标注错误等问题,在不进行数据清洗的情况下,该数据集的准确率一般在 75% 以下。在 FER2013 的相关研究中,SDFL[25] 针对 L2 正则化应用于人脸表情识别中的弊端,提出了一种基于特征稀疏性的正则化方法,该方法结合 L2 正则化和中心损失,能够学习具有更好泛化能力的深层特征,这在一定程度上符合野外人脸表情识别数据的信息稀疏性特性,但由于其所用网络并未相应地针对人脸图像的相关特性作适应性改进,导致其在各个数据集上的准确率无法处于领先水平,而该文所提出的方法则较好地兼顾了这两个问题,除了将人脸表情特性纳入考虑范围之外,也采用了 Softmax 损失函数和中心损失函数混合的损失函数,在一定程度上提高了网络的泛化能力。

对于 RAF-DB 数据集,FDRL[21] 是所列研究中准确率最高的方法,该研究认为情绪信息是表情相似性和表情独立性的组合,其通过特征分解网络首先将基本特征分解为一组面部动作感知潜在特征,以建模表情相似性;通过特征重构网络捕捉潜在特征的特征内和特征间关系,以描述特定情绪的变化,并重建表情特征。尽管 FDRL 所提出的方法实现较为复杂,但该方法的思想在一定程度上也符合表情的呈现特性,针对野外环境下的图像数据有较好效果,使得模型的适用性和鲁棒性有所提高。

3.4　消融实验

针对所提出的基于通道注意的可变形金字塔网络方法的各个关键组件进行了广泛的消融实验,以评估其不同组成部分的有效性,结果如表 4 所示。以 ResNet50 为基线方法,在此基础上逐步加入文中各个模块,并在五个数据集上进行实验,比较最终准确率。其中,CK+数据集上进行的是 6 分类实验。

从消融实验结果上看,空间金字塔池化块在人脸表情识别任务中起到的作用更大,可变形卷积块次之。这在一定程度上说明人脸表情对多尺度空间上下文信息较为敏感,而可变形卷积块也对网络进一步学习不规则特征有所帮助。

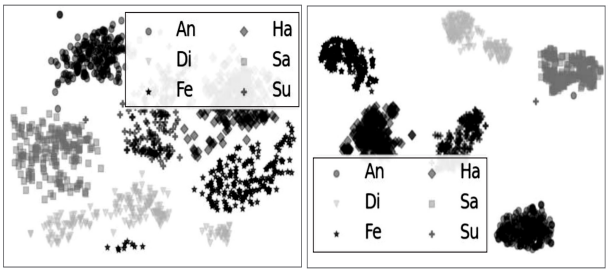
此外,通道注意块也能动态关注网络所提取的特征图,有效降低冗余情绪信息含量,优化情绪识别性能。最终,该文所提出的方法结合了可变形卷积块、空间金字塔池化块和通道注意块,从不规则和多尺度上下文信息这两个符合人脸表情特性的角度出发,挖掘出更多判别性特征,并通过通道注意块抑制冗余特征的重要性,突出判别性特征的贡献度,最终在 5 个人脸表情识别数据集上取得最佳结果。

表 4 关于每个关键组件的消融实验

可变形卷积块	空间金字塔池化块	通道注意块	不同数据集中达到的准确率/%				
			CK+	JAFFE	Oulu-CASIA	FER2013	RAF-DB
			91.39	91.73	79.17	71.13	77.68
√			93.86	93.48	82.49	71.76	79.89
	√		94.91	95.29	83.07	72.33	80.94
		√	93.77	94.58	81.56	71.65	79.57
√	√		97.03	97.34	85.94	73.12	83.76
	√	√	98.33	98.27	87.31	73.82	85.97
√		√	97.62	97.67	86.23	73.01	84.43
√	√	√	99.33	99.23	89.27	74.47	88.91

3.5 结果可视化与分析

使用 t-SNE[29] 分别可视化基线方法和文中方法提取的表情特征向量，如图 5 所示。从基线方法提取的表情特征区分度较低，而从文中提出的方法中提取的特征类内差异较小，类间差异较大。愉快与恐惧、惊讶的类间距离较为接近，而厌恶与悲伤、恐惧的距离较为接近，这个规律在一定程度上符合现实中的表情呈现特性。



(a)基线方法 (b)文中方法

图 5 使用 t-SNE 方法可视化基线方法和文中所用方法在 Oulu-CASIA 数据集上提取到的表情特征

(An：愤怒；Di：厌恶；Fe：恐惧；Ha：愉快；Sa：悲伤；Su：惊讶)

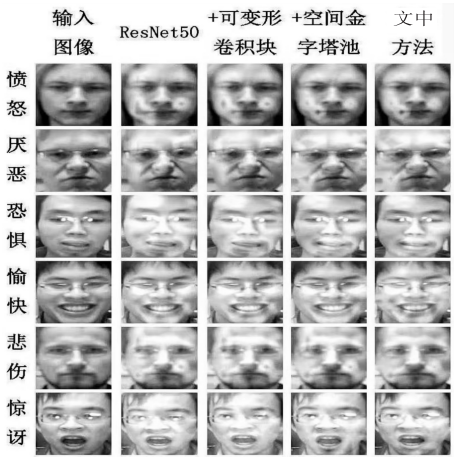


图 6 使用 Grad-CAM 方法可视化模型的关注区域

还使用 Grad-CAM[30] 方法计算得到不同模型在 Oulu-CASIA 数据集中 6 种表情分类的热度图，以展示不同模型对于不同表情的关注区域的差异，其结果如图 6 所示。从图中可以看出，基线方法相较于其他方法而言，其关注的区域面积较大片，且同等关注程度的区域形状相对比较规则；加入了可变形卷积块后，关注的区域会有向外蔓延的趋势，形成不规则的形状；而加入了空间金字塔池化块后，关注的区域保留了上一模型的特点，并显得更加细腻，不同区域的上下文信息存在一定交互，因此更为突出重点区域；继续加入通道注意块后，得到文中方法进一步突出了判别性特征的重要性，而抑制非关键特征的干扰。

4 结束语

基于人脸表情的呈现特性，在经典卷积神经网络 ResNet50 的基础上针对人脸表情识别任务进行合理的适应性改进，提出了一个基于通道注意的可变形金字塔网络。在该网络中，嵌入了多个可变形卷积块、空间金字塔池化块和通道注意块，提高网络对于不规则特征和多尺度空间上下文特征的表征能力。该方法在 3 个实验室环境的人脸表情识别数据集和 2 个野外环境的人脸表情数据集上开展了广泛的对比实验和消融实验，并均取得了领先层次的识别准确率结果，从而验证了该方法和其中各个模块的有效性。从可视化结果上看，该方法促使网络进一步关注与表情强相关的人脸区域，且在一定程度上符合人们对于各种表情显著区域的理解，进而从另一个角度论证了该方法的有效性。当然，该方法还存在着较大的改进空间，未来的研究方向有二：(1)进一步考虑人脸表情的呈现特性，实现更多针对人脸表情识别任务的网络或模块设计，提高判别性特征的质量；(2)结合更多种类的表情特征，比如动态特征，表情相似性与独立性特征，提高判别性特征的含量比，进而更好地完成人脸表情识别任务。

参考文献：

[1] EKMAN P. Strong evidence for universals in facial expressions：a reply to Russell's mistaken critique[J]. Psychological Bulletin，1994，115(2)：268-287.

[2] LUCEY P，COHN J F，KANADE T，et al. The extended Cohn-Kanade dataset (CK+)：a complete dataset for action

unit and emotion - specified expression [C]//2010 IEEE computer society conference on computer vision and pattern recognition - workshops. San Francisco: IEEE, 2010: 94 - 101.

[3] MOHAN K, SEAL A, KREJCAR O, et al. Facial expression recognition using local gravitational force descriptor - based deep convolution neural networks[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-12.

[4] 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.

[5] YANG H, CIFTCI U, YIN L. Facial expression recognition by de - expression residue learning[C]//2018 IEEE/CVF conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 2168-2177.

[6] 邓楚婕. 基于卷积神经网络的人脸表情识别方法研究[D]. 广州: 华南理工大学, 2020.

[7] ZHENG H, WANG R, JI W, et al. Discriminative deep multi-task learning for facial expression recognition[J]. Information Sciences, 2020, 533: 60-71.

[8] 王　珏. 基于深度学习的人脸表情识别研究[D]. 南京: 南京邮电大学, 2021.

[9] FARZANEH A H, QI X. Facial expression recognition in the wild via deep attentive center loss[C]//2021 IEEE winter conference on applications of computer vision (WACV). Waikoloa: IEEE, 2021: 2401-2410.

[10] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]//2017 IEEE international conference on computer vision (ICCV). Venice: IEEE, 2017: 764-773.

[11] ZHANG T, ZHENG W, CUI Z, et al. A deep neural network-driven feature learning method for multi-view facial expression recognition[J]. IEEE Transactions on Multimedia, 2016, 18(12): 2528-2536.

[12] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[13] LYONS M, AKAMATSU S, KAMACHI M, et al. Coding facial expressions with Gabor wavelets[C]//Proceedings third IEEE international conference on automatic face and gesture recognition. Nara: IEEE, 1998: 200-205.

[14] TAINI M, ZHAO G, LI S Z, et al. Facial expression recognition from near-infrared video sequences[C]//2008 19th international conference on pattern recognition. Tampa: IEEE, 2008: 1-4.

[15] LI S, DENG W. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition[J]. IEEE Transactions on Image Processing, 2019, 28(1): 356-370.

[16] ZHANG K, HUANG Y, DU Y, et al. Facial expression recognition based on deep evolutional spatial - temporal networks[J]. IEEE Transactions on Image Processing, 2017, 26(9): 4193-4203.

[17] QU X, ZOU Z, SU X, et al. Attend to where and when: cascaded attention network for facial expression recognition[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6(3): 580-592.

[18] ZHAO X, LIANG X, LIU L, et al. Peak-piloted deep network for facial expression recognition[C]//Computer vision-ECCV 2016. Amsterdam: Springer, 2016: 425-442.

[19] MOHAN K, SEAL A, KREJCAR O, et al. FER-net: facial expression recognition using deep neural net[J]. Neural Computing and Applications, 2021, 33(15): 9125-9136.

[20] ZOU W, ZHANG D, LEE D J. A new multi-feature fusion based convolutional neural network for facial expression recognition[J]. Applied Intelligence, 2022, 52(3): 2918-2929.

[21] RUAN D, YAN Y, LAI S, et al. Feature decomposition and reconstruction learning for effective facial expression recognition[C]//2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Nashville: IEEE, 2021: 7656-7665.

[22] XIE S, HU H, WU Y. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition[J]. Pattern Recognition, 2019, 92: 177 - 191.

[23] XIE W, SHEN L, DUAN J. Adaptive weighting of handcrafted feature losses for facial expression recognition[J]. IEEE Transactions on Cybernetics, 2021, 51(5): 2787-2800.

[24] RUAN D, YAN Y, CHEN S, et al. Deep disturbance-disentangled learning for facial expression recognition[C]//Proceedings of the 28th ACM international conference on multimedia. Seattle: ACM, 2020: 2833-2841.

[25] XIE W, JIA X, SHEN L, et al. Sparse deep feature learning for facial expression recognition[J]. Pattern Recognition, 2019, 96: 106966.

[26] KIM B K, DONG S Y, ROH J, et al. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach[C]//2016 IEEE conference on computer vision and pattern recognition workshops (CVPRW). Las Vegas: IEEE, 2016: 1499-1508.

[27] WANG K, PENG X, YANG J, et al. Suppressing uncertainties for large-scale facial expression recognition[C]//2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle: IEEE, 2020: 6896-6905.

[28] LIU Y, ZHANG X, ZHOU J, et al. SG - DSN: a semantic graph-based dual-stream network for facial expression recognition[J]. Neurocomputing, 2021, 462: 320-330.

[29] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(86): 2579-2605.

[30] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.