图像生成和深度度量学习的身份感知面部表情识别方法

章思远¹⁾, 肖世明¹⁾, 张蓬²⁾, 黄伟^{3)*}

¹⁾ (南昌大学信息工程学院 南昌 330031)

²⁾ (西北工业大学计算机学院 西安 710072)

³⁾ (南昌大学信息化办公室 南昌 330031)

(n060101@e.ntu.edu.sg)

摘 要: 为解决面部表情识别中不同图像的背景信息和身份特征会干扰分类准确率的问题, 提出一种将图像合成技术和深度度量学习相结合的身份感知人脸表情识别方法, 通过在面部表情识别任务中创建相同身份下的表情组, 对人脸图像特征进行比较分类. 其结构中对抗生成网络, 目标在于学习表情信息并生成表情组; 特征提取网络用于将图像转化成为可进行度量学习的特征向量; 马氏度量学习网络能够有效地对一对特征值进行比较与分类. 该方法在常用面部表情识别数据集 CK+和 Oulu-CASIA 上取得了 98.653 2%和 99.824 8%的平均分类准确率, 并在 Oulu-CASIA 数据集上超过当前最好方法 10%以上. 通过与目前最新方法的比较, 证实了该方法在面部表情识别中的有效性和进步性.

关键词: 深度度量学习; 图像生成; 身份感知; 面部表情识别

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2021.18462

Identity-Aware Facial Expression Recognition Method Based on Synthesized Images and Deep Metric Learning

Zhang Siyuan¹⁾, Xiao Shiming¹⁾, Zhang Peng²⁾, and Huang Wei^{3)*}

¹⁾ (School of Information Engineering, Nanchang University, Nanchang 330031)

²⁾ (School of Computer Science, Northwestern Polytechnical University, Xi'an 710072)

³⁾ (Informatization Office, Nanchang University, Nanchang 330031)

Abstract: Facial expression recognition (FER) is a challenging task because the external environment and identity characteristics could affect the classification results directly. To settle down the above-mentioned challenges, this paper proposed an identity-aware facial expression recognition method which combined images synthesis techniques and deep metric learning, and made facial images features compared then classified by creating expression groups under the same identity in FER task. There are three parts in our method. The first part is a generative adversarial network, which aims to learn expression information and synthesis the expression groups. the second part is the feature extraction network, which transforms the image into feature vectors that could be used for metric learning. The third part is Mahalanobis metric learning network that could compare and classify a pair of feature values effectively. The average accuracy of proposed method reached 98.653 2% and 99.824 8% on two well-known FER dataset named CK+ and Oulu-CASIA, with more than 10% higher than the method proposed currently. By comparing with several

收稿日期: 2020-06-30 ; 修回日期: 2021-03-20. 基金项目: 国家自然科学基金(61862043, 61971352); 江西省主要学科科学技术带头人培养计划领军人才项目(20204BCJ22011); 陕西省自然科学基金(2018JM6015). 章思远(1996—), 男, 硕士研究生, 主要研究方向为计算机视觉; 肖世明(1995—), 男, 硕士研究生, 主要研究方向为计算机视觉; 张蓬(1978—), 男, 博士, 教授, 博士生导师, 主要研究方向为计算机视觉、模式识别; 黄伟(1983—), 男, 博士, 教授, 博士生导师, CCF 会员, 论文通讯作者, 主要研究方向为计算机视觉、模式识别.

state-of-the-art methods, the experimental results confirmed that the proposed-method was effective and progressive in FER task.

Key words: deep metric learning; image synthesis; identity-aware; facial expression recognition

在基于人类面部图像研究中,面部表情识别(facial expression recognition, FER)被广泛认为是表征人类情绪有效且直接的方法^[1-3]. 现有的表情分类标准通常将人类表情分为6种基本类别:快乐(happy), 生气(angry), 悲伤(sad), 恐惧(fear), 厌恶(disgust)和惊奇(surprise). 由于通过人类面部表情的表达能很好地对人类感情进行分析,因此对6种基本表情分类的准确率常被直接用于评估FER模型.

FER研究最大的挑战来源于表达人身份的多样性和复杂性,其中影响较大的干扰因素包含性别、年龄和不同身份的面部特征. 文献[4-5]讨论了即使是在不同身份的相同类别面部表情图像中分类,多分类模型也会受到不同身份的因素干扰,导致误分面部表情图像. 旨在避免以上身份因素干扰的表情识别被称为身份感知性FER(identity-aware FER).

本文提出图像生成和深度度量学习的身份感知FER方法,真实图像通过starGAN^[6]生成所有标签的面部表情图像,将其并归为一个表情组后,对表情组内部的图像进行相同身份下真实图像和生成图像两两间的特征比较,并用马氏度量学习(Mahalanobis metric learning, MML)^[7]计算出相似度最高的一对图像. 该方法可以描绘出不同表情标签和非完整表情图像的关系,加大了数据规模,避免因身份特征的差异性影响FER的结果. 在CK+^[8-9]和Oulu-CASIA^[10]2个常用FER数据集进行了实验,并与其他FER方法进行了比较.

1 相关工作

1.1 FER

FER受很多因素影响,包括拍摄背景、光线强度和面部角度等,但影响最大的还是身份之间的差异性,这导致了对FER的研究一直是很具有挑战性课题. 近年来,FER的研究主要包含对6种基本面部表情分类和对面部动作单元(action unit, AU)分类2个方向. Hu等^[11]提出了基于卷积神经网络(convolutional neural networks, CNN)的多任务学习模型,采用对抗域对齐、参数共享和残差网络(ResNet)的策略构建模型. Zhao等^[12]采用更加丰富的过渡表情图像,将同标签图像序列的轻微表情

和强烈表情同时作为基于GoogleNetCNN的输入,由此固化表情强度的不变性,使得分类更加具有泛化性. FN2EN模型^[13]通过学习面部知识和表情知识,具备了一定的身份信息学习和良好的数据集迁移能力. Fabian Benitez-Quiroz等^[14]在表情分类的研究中考虑了AU对6种基本面部表情的映射关系,提出基于AU的面部表情分类方法. 虽然基于AU的分类在将来仍拥有很大的潜力,但是其不足是现有的数据集大多缺乏专用的AU标签^[15]. 此外,标记大样本的AU数据集,工作量巨大且专业性强.

基于身份感知性的表情识别是近年来提出的一种重要思想,由文献[4]可知,为使不同身份的同类表情之间有更紧密的关系,可采用缩小类内差异和扩大类外差异的度量方法使模型对面部表情分类更加精确. 去表情剩余学习(de-expression residue learning, DERL)的分类方法^[16]将包含表情的图像转化为无表情的中性图像,降低了图像身份信息对表情干扰. Yang等^[5]采用条件生成对抗网络(condition generative adversarial networks, CGAN)^[17]的生成器生成相同身份条件下的不同面部表情,通过判别器得出特征最一致的一对真假图像,由此得到预测标签. 本文借鉴了以上的图像生成思想,并且在实验部分与文献[5]进行了比较.

除依靠图像生成方法提升FER能力外,采用时序性输入及面部关键点信息[18-19],以及拆分面部图像为局部区域[20-22]等的其他FER方法,也有助于模型学习到更多的表情信息,提升分类能力.

1.2 GAN

通常,GAN^[23]模型包含生成器和判别器2类深度学习网络. 通过2类网络间的博弈,模型的参数逐渐趋于与真实分布相近的解. 自GAN出现到现在,很多旨在完善其模型让生成器拥有生成更具真实性和多样性图像的能力. 文献[17,24]为GAN加入了条件信息,得到质量更好的生成结果,其贡献是将图像与目标标签之间建立关联;Arjovsky等^[25]采用Wasserstein距离来衡量分布差异,提升了GAN模型的稳定性和鲁棒性.

CycleGAN成功地分离了风格和内容关系^[26],其生成器采用编码器-解码器模型,能很好地完成

2 种图像风格之间的转换. 但是 CycleGAN 局限在只能处理极有限的 2 种风格转换; 此外, 它的双生成器和双判别器在实际训练上耗时大, 调参难度高. 为解决 CycleGAN 的局限性, starGAN 利用 Reconstruction 方法循环训练生成器, 实现一对多的风格迁移^[6], 并且该方法在面部表情以及容貌特征上得到了很好的实验效果.

2 身份感知的表情识别方法

如图 1 所示, 本文提出的模型结构包含 3 个部分. 第 1 部分是 starGAN 模型. 在 FER 任务中, 生成图像的关键在于多种标签间的图像风格转换. 为解决多领域迁移学习问题, 本文采用图像到图像的生成对抗模型——starGAN, 运用其生成器引入目标标签, 构造同身份下完整的面部表情组. 第

2 部分是借鉴 VGG 结构含 5 个卷积层的 FER 特征提取网络, 其作用为提取生成图像和真实图像的特征, 并将相同标签的一组特征向量运用均方误差计算损失, 进一步缩小同身份下同类表情之间的特征差异. 第 3 部分是马氏度量学习网络, 其将第 2 部分输出的结果进行 MML, 并将基于面部表情多分类任务转化为比较一对图像特征相似性的二分类任务. 如图 1 所示, 真实图像由 Step1 输入到由 CNN 构成的特征提取网络, 获得真实图像的特征向量; 生成图像依次单独由 Step2 分别输入该网络, 得到 6 个生成图像的特征向量; 然后, 采用深度马氏度量网络将其分别与真实图像的特征向量计算差异, 通过均方误差(mean-square error, MSE)和 hinge-based loss 优化网络参数; 最终预测结果为生成图像组与真实图像间特征向量差异最小的标签(输出值 Y_p 最大).

图 1 本文模型结构

2.1 starGAN

starGAN 与其他 GAN 模型(如 GAN^[23], CGAN^[17], WGAN^[27])相似, 都可以完成图像生成任务. 但其拥有更加突出的图像生成能力和适应多领域图像到图像转化任务的能力. 如图 2 所示, 本文使用几种常用 GAN 模型生成器生成的面部表情图片样例, 通过列于列之间对比可见, starGAN 生成器的生成效果不论是在生成的细节上还是目标标签的效果上, 均优于其他 GAN. 此外, 与原始 GAN 生成器由噪声解码生成图像不同的是, starGAN 生成器的输入也是图像, 这使得输出图像和真实图像之间具有不同领域上的良好关联性. 而 WGAN 虽然在生成图像的多样性和真实性上比原始 GAN 模型更优异, 但本质上仍然没有引入标签参数控制网络生成特定领域的图像. 与 WGAN 相比, CGAN 引入了标签参数控制图像的生成, 却无法构建生成图像与真实图像间的联系, 导致生

图 2 4 种 GAN 模型在 CK+数据集上的生成图像

成面部图像的效果较差. 采用 starGAN 进行图像生成不仅扩增了原始训练集的数据量, 还使每幅真实图像拥有所有目标领域的生成图像.

starGAN 结构如图 3 所示, 它包含一个生成器和一个判别器, 通过生成器重构参数的学习, 生成更加真实的图像. 在神经网络的设计上, 生成器借鉴 CycleGAN 采用的残差块结构; 判别器采用含有 LeakyReLU 作为激活函数的 PatchGANs^[28] 结构. starGAN 包含对抗损失、分类损失和重构损失 3 个损失函数. 下面分别给出它们的计算方法.

图 3 starGAN 模型结构和训练步骤

starGAN 采用对抗损失函数^[27](adversarial loss)

$$L_{\text{adv}} = E_x [D_{\text{src}}(x)] - E_{x,c} [D_{\text{src}}(G(x,c))] - \lambda_{\text{gp}} E_{\hat{x}} \left[\left(\|\nabla_{\hat{x}} D_{\text{src}}(\hat{x})\|_2 - 1 \right)^2 \right] \quad (1)$$

其中, D 表示判别器网络; G 表示生成器网络; x 表示输入图像; c 表示图像的目标标签; $\hat{x}$ 表示一对真实图像和生成图像间随机插值的采样; D_{src} 表示通过判别器输出的图像真假概率分布; λ_{gp} 表示梯度惩罚项的参数. Wasserstein 距离和梯度惩罚项能稳定 starGAN 的训练过程, 并加快训练速度.

分类损失(classification loss)采用判别器分类函数^[24]

$$L_{\text{cls}} = E_{x,c} [-\ln D_{\text{cls}}(c|x)] \quad (2)$$

其中, c 表示图像 x 对应的目标域标签; D_{cls} 表示通过判别器输出的标签概率分布.

为提升 starGAN 生成器的重构质量和生成质

量, 采用 Cycle consistency 方法^[28]的重构损失(reconstruction loss)

$$L_{\text{rec}} = E_{x,c,c'} [\|x - G(G(x,c),c')\|_1] \quad (3)$$

其中, c 表示图像 x 目标域标签; c' 表示图像 x 的原始标签.

starGAN 的 3 个损失函数相互协调, 其总损失函数为

$$L_{\text{stargan}} = L_{\text{adv}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{rec}} L_{\text{rec}} \quad (4)$$

其中, λ_{cls} 和 λ_{rec} 分别为分类损失项和重构损失项的参数.

2.2 FER 特征提取网络

通过第 2.1 节 starGAN 的生成器, 模型会得到一组完整标签的生成图像. 分别选择生成图像和原始的真实图像一对一比较, 通过缩小相同标签的一对图像之间的距离, 提升网络的拟合能力. 本文方法的第 2 部分采用 CNN, 目的是提取出每一对图像的特征向量, 并将它们传给第 3 部分进行度量学习.

为减少模型的参数和训练时间, 使模型的训练集准确率达到 100%, 特征提取网络包含了 5 个卷积层. 如图 4 所示, 卷积层之间采用批标准化(batch normalization, BN)层^[29]. ReLU 激活函数是负半轴为 0 值的非线性激活函数, BN 层可将 ReLU 函数的输出结果强制拉回到均值为 0、方差为 1 的正态分布区间, 以避免梯度消失问题. 本文方法采用最大池化策略缩小特征图的尺寸, 最后一层采用自适应的平均池化层(图 4 中 Adaptive Ave pooling), 得到大小为 $1 \times 1 \times 512$ 的特征向量. 在网络训练中, 采用目前常用的 MSE 函数优化 FER 特征提取网络, 该损失函数用于训练时拉近一对相同标签图像的特征距离. 图 5 所示为真实图像及其生成图像组在一个已训练完成的 FER 特征提取网络中的平均特征差异, 各缩写分别为生气(An), 蔑视(Co), 厌恶(Di), 恐惧(Fe), 快乐(Ha), 悲伤(Sa)和惊奇(Su), 计算差异采用经平均化的曼哈顿距离(Manhattan distance, MD), 即

$$d(x, y) = \frac{1}{N} \sum_i |x_i - y_i| \quad (5)$$

其中, x 和 y 分别表示 2 个特征向量; N 表示特征向量的长度. 由图 5 可知, 同标签的特征向量距离值比其他不同标签的特征向量距离值更小, 而采用 MSE 函数能进一步缩小相同标签特征向量之间的距离.

组图像序列,每组图像序列均是无表情到有表情的渐变面部图片.其中,只有327个是有表情标签的序列.CK+数据集标签种类为7种,比原始6种基本表情多出蔑视(contempt)表情.由于大部分身份下的表情图像标签缺失,故CK+数据集是标签数目非均衡的数据集.

图6 2个数据集的图像样本

Oulu-CASIA数据集是一个需要申请获得并包含6种表情标签的常用数据集.其包含2个子集VIS(visible light)和NIR(near infrared),VIS包含Strong, Weak, Dark共3种类型的图像集,它们分别表示在光照良好条件下、弱光照条件下和无光照条件下的图像.每种类型的图像集均有80个文件,每个文件包含同一身份的6种表情序列,属于标签数目均衡的数据集.与CK+相似,Oulu-CASIA的每个表情序列也由无表情到有表情渐变的一组面部图像构成.

CK+和Oulu-CASIA数据集均包含不同人种、肤色、性别以及年龄的面部表情图片.Oulu-CASIA数据集的人像包含佩戴眼镜,CK+数据集包含复杂的光照条件和彩色图像的情况.

3.2 实验设置

为使实验和现有方法更具有比较性,进行了与文献[4-5]相同的数据集预处理操作.在CK+数据集中选取带标签的每组图像序列最末3幅图像,采用MTCNN人脸检测算法^[32]定位面部,然后自动截出实验所需的表情图像.在Oulu-CASIA数据集中,实验使用VIS子集下的Strong图像集,如图6b所示,采集的图像均来自数据集Oulu-CASIA的

C_PreProcess文件夹,实验也是取每个表情文件组最后3幅图像.然后,对2个数据集进行如下操作:处理成灰度图并缩放图像像素分辨率为 $64 \times 64 \times 1$;对每张图片进行旋转操作,旋转角度分别为 -15° , -12° , -9° , -6° , -3° , 0° , 3° , 6° , 9° , 12° 和 15° ,初步扩增图像;将整理好的数据集随机分成10份,作为训练集比测试集为9:1的交叉验证使用.预处理后CK+数据集为9 922幅图像,Oulu-CASIA为15 840幅图像.

starGAN的训练参数采用文献[6]中给出的适用于人脸任务的设置值:图像输入大小 $64 \times 64 \times 1$, batch size为16,迭代次数为200 000,生成器和判别器学习率为0.000 1,生成器优化器采用Adam(系数为0.5),判别器优化器采用Adam(系数为0.999),分类损失参数为1,重构损失的参数为10,梯度惩罚项参数为10,生成器生成图像大小设置为 $64 \times 64 \times 1$,判别器与生成器更新频次比值为5.实验的编程框架采用PyTorch,训练机器使用Nvidia TITAN V GPU,完成一个生成器模型的训练耗时约6 h.

starGAN生成器完成训练之后,进行训练集的图像生成,扩增得到的训练数据作为FER特征提取网络的输入初始数据.FER特征提取网络和MML网络合并训练,模型的参数设置如下:batch size为8,迭代数30次,初始学习率为0.01,每10次迭代衰减系数为0.1,优化器采用SGD;设定损失函数的参数 $\lambda_{MSE}=100$, $\lambda_{l2}=0.1$.神经网络输入大小为 $64 \times 64 \times 1$.一个模型的训练耗时约2 h.

CK+和Oulu-CASIA数据集实验均采用10倍交叉验证,实验结果取10次实验结果的平均值.

3.3 CK+数据集结果与分析

CK+数据集包含7个标签的面部表情,本实验是一个7分类的任务.采用本文方法进行实验,CK+数据集的准确率达到98.6532%.如表1所示,本文将实验结果和近几年的8种FER方法(LBVCNN^[33], RN+LAF+ADA^[11], PPDN^[12], IACNN^[4], IA-gen^[5], ppfSVM^[18], DTAGN^[19], DCMA-CNNs^[22])以及仅采用starGAN判别器分类的结果进行了比较.通过在CK+数据集实验的准确率可见,仅有PPDN模型比本文方法的准确率略高.值得注意的是,PPDN所采用的训练数据包含非峰值表情图像,比本文方法原始训练数据需求多1倍;LBVCNN, ppfSVM和DTAGN方法输入均为1组图片序列,用于训练的数据包含大量身份信息,其训练采用的数据量比本文方法实际采用的

数据量高出十余倍; 尽管 IACNN 与 IA-gen 考虑了身份信息对表情分类的影响, 但是 IACNN 在对拉远身份信息的处理方法上, 没有采用图像生成法还原身份的完整性高, 而采用 CGAN 的 IA-gen 在图像生成的质量以及特征提取上没有本文的 starGAN 和深度马氏度量方法全面. 与未考虑身份信息的机器学习方法^[11,18,22]相比, 本文方法的准确率亦更高.

表 1 10 种方法在 CK+数据集上的平均准确率比较

方法	准确率/%	实验策略	输入
LBVCNN(joint) ^[33]	97.380 0	10 折交叉验证+旋转	图像序列
RN+LAF+ADA(FT) ^[11]	96.400 0	5 折交叉验证	图像
PPDN ^[12]	99.300 0	10 折交叉验证	图像
IACNN ^[4]	95.370 0	8 折交叉验证	图像
IA-gen ^[5]	96.570 0	10 折交叉验证+旋转	图像
ppfSVM ^[18]	96.870 0	10 折交叉验证	图像序列
DTAGN(joint) ^[19]	97.250 0	10 折交叉验证+旋转	图像序列
DCMA-CNNs ^[22]	93.460 0	10 折交叉验证	图像
starGAN (Discriminator) ^[6]	67.280 0	10 折交叉验证+旋转	图像
本文(无 MSE loss)	98.266 5	10 折交叉验证+旋转	图像
本文	98.653 2	10 折交叉验证+旋转	图像

表 2 描述了本文方法在 CK+数据集实验结果的混淆矩阵, 斜对角数值分别代表生气(An)、蔑视(Co)、厌恶(Di)、恐惧(Fe)、快乐(Ha)、悲伤(Sa)和惊奇(Su)分类的准确率. 通过数据可见, 蔑视标签的数值低于其他标签, 原因是蔑视标签的数据量小于其他 6 种标签的数据, 导致模型缺少相关标签的训练数据. 通过对比发现, 无 MSE 损失函数的模型分类准确率略低于有 MSE 的模型, 证明本文第 2.2 节的 MSE 损失函数可以从拉近相同标签特征距离的角度上提升分类准确率. 其有效性同样被第 3.4 节 Oulu-CASIA 实验结果验证.

表 2 本文方法在 CK+数据集上的混淆矩阵

	An	Co	Di	Fe	Ha	Sa	Su
An	98.093 9	0.520 5	0.292 5	0	0.292 0	0.510 0	0.291 2
Co	0.175 4	95.811 0	0	0.506 6	0.650 8	2.856 2	0
Di	0	0	99.682 9	0.211 0	0.106 2	0	0
Fe	0	0	0.133 3	98.546 0	0.924 0	0.396 7	0
Ha	0	0	0	0	99.909 7	0.045 7	0.044 6
Sa	0.779 3	0.263 2	0	0	0	98.957 5	0
Su	0	0.257 8	0.042 9	0	0.084 6	0.043 5	99.571 3

3.4 Oulu-CASIA 数据集结果与分析

Oulu-CASIA 数据集包含 6 个表情标签. 本实验及比较方法均采用 VIS 下的 Strong 子集. 实验结果同近几年的 7 种模型(LBVCNN^[33], RN+LAF+

ADA^[11], PPDN^[12], FN2EN^[13], IA-gen^[5], ppfSVM^[18], DTAGN^[19])进行比较, 如表 3 所示. 值得关注的是, IA-gen 也是一种采用对抗生成模型的 FER 方法, 本文方法比先前最佳方法 IA-gen 在准确率上提升了超过 10 个百分点, 在 Oulu-CASIA 数据集上取得目前最高的分类准确率.

表 3 9 种方法在 Oulu-CASIA 数据集上的平均准确率比较

方法	准确率/%	策略	输入
LBVCNN(joint) ^[33]	82.410 0	10 折交叉验证+旋转	图像序列
RN+LAF+ADA(FT) ^[11]	87.100 0	10 折交叉验证	图像
PPDN ^[12]	84.590 0	10 折交叉验证	图像
FN2EN ^[13]	87.710 0	10 折交叉验证	图像
IA-gen ^[5]	88.920 0	10 折交叉验证+旋转	图像
ppfSVM ^[18]	83.130 0	10 折交叉验证	图像序列
DTAGN(joint) ^[19]	81.460 0	10 折交叉验证+旋转	图像序列
starGAN (Discriminator) ^[6]	62.651 5	10 折交叉验证+旋转	图像
本文(无 MSE loss)	99.806 9	10 折交叉验证+旋转	图像
本文	99.824 8	10 折交叉验证+旋转	图像

与 CK+数据集实验的结果相比, 分类准确率与其他方法普遍下降不同, 本文方法反而有所提升. 原因归纳为以下 2 点:

(1) Oulu-CASIA 的面部区域相较 CK+更小, 而且分辨率更低, 给多数方法带来分类上的困难. 本文方法采用数据生成策略, 加大了特征提取网络的训练数据规模, 有助于提升网络的泛化性能;

(2) Oulu-CASIA 数据集在每个身份的面部表情图像上无缺失, 而 CK+有大量的缺失, 不平衡的数据集会造成深度学习网络在学习样本分布上的不均衡. 由于本文方法基于图像生成策略, 故原始数据标签越完整, 身份非独立性越强, starGAN 生成器的生成效果越好.

本文方法在 Oulu-CASIA 数据集的混淆矩阵如表 4 所示. 通过斜对角线上标签分类准确率可见, 本文方法在该数据集上的各表情分类准确率均超过 99.651 2%. 在惊奇表情上的识别效果最好,

表 4 本文方法在 Oulu-CASIA 数据集上的混淆矩阵

	An	Di	Fe	Ha	Sa	Su
An	99.852 3	0.036 9	0.323 0	0	0.078 6	0
Di	0.071 3	99.651 2	0.099 0	0.114 4	0.064 1	0
Fe	0	0.041 7	99.807 4	0.074 0	0	0.769 0
Ha	0	0	0.041 5	99.920 3	0	0.038 2
Sa	0.161 2	0	0	0.079 4	99.759 4	0
Su	0	0	0	0.041 7	0	99.958 3

达到了 99.950 0%以上. 与标签非均衡的 CK+数据集相比, 本文方法在标签均衡的 Oulu-CASIA 数据集上的分类准确率更高.

3.5 特征提取网络结构研究

针对第 2.2 节含有 5 层卷积层的特征提取网络的结构, 本节给出了其他层数的结果. CNN4 表示采用 4 层卷积层, CNN5 表示采用 5 层卷积层, 以此类推. 由表 5 可见, 网络层数的增加会给实验结果带来细微的变化, 但是总体变化量不大; 此外卷积层过多, 会极大地减少训练模型所需的时间. 对比 CNN6 和 CNN5 的实验结果发现, 增加模型参数对该特征提取网络性能无持续提升.

表 5 不同深度特征提取网络实验结果

卷积层	准确率/%		参数数目
	CK+	Oulu-CASIA	
CNN4	98.288 0	98.989 9	4 601 793
CNN5	98.653 2	99.824 8	6 962 113
CNN6	98.347 6	99.179 3	9 322 433
CNN7	99.093 7	99.611 1	11 682 753

3.6 CK+与 Oulu-CASIA 数据集混合实验

为验证本文方法在不同数据集上的迁移性与有效性, 对 CK+与 Oulu-CASIA 数据集进行混合实验. 由第 3.1 节可知, CK+数据集在 FER 的标签种类为 7, 比 Oulu-CASIA 多出蔑视标签, 并且在图 6 中能明显分辨数据集存在较大差异. 本实验采用剔除蔑视标签的 CK+数据集并混合 Oulu-CASIA 数据集, 进行 10 折交叉验证, 结果取平均值.

实验结果如表 6 所示, CK+(6)表示剔除蔑视标签的 6 种 FER 实验结果, CK+&Oulu-CASIA 表示 CK+(6)与 Oulu-CASIA 混合数据集实验结果, 本文(WMSE)表示不采用 MSE 损失函数训练特征提取网络. 从每列结果对比可以发现, Oulu-CASIA 与 CK+混合对本文方法的 FER 上没有不良影响, 在无 MSE 损失函数训练的模型上比原有的 CK+(6)提升了 0.369 3%. 实验表明, 在更复杂的混合面部表情数据集中, 本文方法仍有较好的分类效果与积极的现实意义.

表 6 CK+与 Oulu-CASIA 混合实验结果

数据集	准确率/%		图像数/幅
	本文(WMSE)	本文	
CK+(6)	94.298 2	94.625 0	848
Oulu-CASIA	99.806 9	99.824 8	1 440
CK+&Oulu-CASIA	94.667 5	94.625 0	2 288

4 结 语

为解决身份信息对面部表情分类任务的不良影响, 本文提出了一种基于身份感知的 FER 方法, 首先通过 starGAN 生成图像, 然后采用特征提取网络进行特征提取, 最后采用基于 MML 的深度网络进行二分类投票, 得到表情预测结果. 本文方法具有以下贡献性: 将深度度量学习与面部图像合成技术相结合, 完成 FER 分类任务; 将原来的图像多分类任务转化为采用图像生成技术进行同身份下差异比较的二分类任务, 避免了身份差异对网络训练的干扰; 实验结果上, 本文方法在标签数目均衡与不均衡的数据集上均有比现有多数方法更高的准确率, 特征提取网络结构实验证明了其在 FER 任务中具有普遍有效性. 其不足在于, 本文只重点考虑了提高实验环境下收集的数据集分类准确率, 没有考虑涵盖范围更广且具有身份独立性的自然环境下收集的数据集. 未来工作将更多地集中在完善和研究非实验环境中的表情分类的任务.

参考文献(References):

- [1] Corneanu C A, Simón M O, Cohn J F, *et al.* Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(8): 1548-1568
- [2] Sariyanidi E, Gunes H, Cavallaro A. Automatic analysis of facial affect: a survey of registration, representation, and recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(6): 1113-1133
- [3] Soleymani M, Asghari-Esfeden S, Fu Y, *et al.* Analysis of EEG signals and facial expressions for continuous emotion detection[J]. *IEEE Transactions on Affective Computing*, 2016, 7(1): 17-28
- [4] Meng Z B, Liu P, Cai J, *et al.* Identity-aware convolutional neural networks for facial expression recognition[C] // *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition*. Los Alamitos: IEEE Computer Society Press, 2017: 558-565
- [5] Yang H Y, Zhang Z, Yin L J. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks[C] // *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition*. Los Alamitos: IEEE Computer Society Press, 2018: 294-301
- [6] Choi Y, Choi M, Kim M, *et al.* StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C] // *Proceedings of the IEEE/CVF Conference on Com-*

- puter Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 8789-8797
- [7] De Maesschalck R, Jouan-Rimbaud D, Massart D L. The Mahalanobis distance[J]. *Chemometrics and Intelligent Laboratory Systems*, 2000, 50(1): 1-18
  - [8] Kanade T, Cohn J F, Tian Y L. Comprehensive database for facial expression analysis[C] // *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*. Los Alamitos: IEEE Computer Society Press, 2000: 46-53
  - [9] 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] // *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2010: 94-101
  - [10] Zhao G Y, Huang X H, Taini M, *et al.* Facial expression recognition from near-infrared videos[J]. *Image and Vision Computing*, 2011, 29(9): 607-619
  - [11] Hu G S, Liu L, Yuan Y, *et al.* Deep multi-task learning to recognise subtle facial expressions of mental states[C] // *Proceedings of the 15th European Conference on Computer Vision*. Heidelberg: Springer, 2018: 106-123
  - [12] Zhao X Y, Liang X D, Liu L Q, *et al.* Peak-piloted deep network for facial expression recognition[C] // *Proceedings of the 14th European Conference on Computer Vision*. Heidelberg: Springer, 2016: 425-442
  - [13] Ding H, Zhou S K, Chellappa R. FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition[C] // *Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition*. Los Alamitos: IEEE Computer Society Press, 2017: 118-126
  - [14] Fabian Benitez-Quiroz C, Srinivasan R, Martinez A M. Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild[C] // *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2016: 5562-5570
  - [15] Peng G Z, Wang S F. Weakly supervised facial action unit recognition through adversarial training[C] // *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2018: 2188-2196
  - [16] Yang H Y, Ciftci U, Yin L J. Facial expression recognition by de-expression residue learning[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2018: 2168-2177
  - [17] Mirza M, Osindero S. Conditional generative adversarial nets[OL]. [2020-06-30]. <https://arxiv.org/abs/1411.1784>
  - [18] Kacem A, Daoudi M, Amor B B, *et al.* A novel space-time representation on the positive semidefinite cone for facial expression recognition[C] // *Proceedings of the IEEE International Conference on Computer Vision*. Los Alamitos: IEEE Computer Society Press, 2017: 3199-3208
  - [19] Jung H, Lee S, Yim J, *et al.* Joint fine-tuning in deep neural networks for facial expression recognition[C] // *Proceedings of the IEEE International Conference on Computer Vision*. Los Alamitos: IEEE Computer Society Press, 2015: 2983-2991
  - [20] Sariyanidi E, Gunes H, Cavallaro A. Robust registration of dynamic facial sequences[J]. *IEEE Transactions on Image Processing*, 2017, 26(4): 1708-1722
  - [21] Ryu B, Rivera A R, Kim J, *et al.* Local directional ternary pattern for facial expression recognition[J]. *IEEE Transactions on Image Processing*, 2017, 26(12): 6006-6018
  - [22] Xie S Y, Hu H F. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks[J]. *IEEE Transactions on Multimedia*, 2019, 21(1): 211-220
  - [23] Goodfellow I, Pouget-Abadie J, Mirza M, *et al.* Generative adversarial nets[C] // *Proceedings of the 27th International Conference on Neural Information Processing Systems*. New York: ACM Press, 2014: 2672-2680
  - [24] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs[C] // *Proceedings of the 34th International Conference on Machine Learning*. New York: ACM Press, 2017: 2642-2651.
  - [25] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN[OL]. [2020-06-30]. <https://arxiv.org/abs/1701.07875>
  - [26] Zhu J Y, Park T, Isola P, *et al.* Unpaired image-to-image translation using cycle-consistent adversarial networks[C] // *Proceedings of the IEEE International Conference on Computer Vision*. Los Alamitos: IEEE Computer Society Press, 2017: 2242-2251
  - [27] Gulrajani I, Ahmed F, Arjovsky M, *et al.* Improved training of Wasserstein GANs[C] // *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Cambridge: MIT Press, 2017: 5769-5779
  - [28] Isola P, Zhu J Y, Zhou T H, *et al.* Image-to-image translation with conditional adversarial networks[OL]. [2020-06-30]. <https://arxiv.org/abs/1611.07004>
  - [29] Ioffe S, Szegedy C. Batch Normalization: accelerating deep network training by reducing internal covariate shift[OL]. [2020-06-30]. <https://arxiv.org/abs/1502.03167>
  - [30] Alam M, Vidyaratne L S, Iftikharuddin K M. Sparse simultaneous recurrent deep learning for robust facial expression recognition[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(10): 4905-4916
  - [31] Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2015: 4353-4361
  - [32] Zhang K P, Zhang Z P, Li Z F, *et al.* Joint face detection and alignment using multitask cascaded convolutional networks[J]. *IEEE Signal Processing Letters*, 2016, 23(10): 1499-1503
  - [33] Kumawat S, Verma M, Raman S. LBVCNN: local binary volume convolutional neural network for facial expression recognition from image sequences[OL]. [2020-06-30]. <https://arxiv.org/abs/1904.07647v1>