计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900028

结合特征融合和注意力机制的微表情识别方法

李星燃　张立言　姚树婧
南京航空航天大学计算机科学与技术学院　南京 211106
(lixingran@nuaa.edu.cn)

摘　要　微表情指当人们试图隐藏或抑制自己的真实情感时,脸上出现的一种无法控制的肌肉运动。此类情绪面部表情由于具有持续时间短、动作幅度小、难以掩饰和抑制的特点,因此其识别精度受到了制约。为了应对这些挑战,文中提出一种结合特征融合和注意力机制的微表情识别方法,同时考虑了光流特征和人脸特征,通过进一步加入注意力机制来提升识别性能。该网络由3个部分组成:1)提取每个微表情片段中Onset到Apex的光流与光学应变,将垂直光流、水平光流、光学应变输入到一个浅层3DCNN中,以提取光流特征;2)以深度卷积神经网络ResNet-10为迁移模型,加入卷积注意力模块以提取人脸特征;3)将两个特征向量拼接起来进行分类。利用所提方法在3个自发微表情数据集中进行实验,结果表明,所提方法在微表情识别方面优于传统方法和现有深度学习方法。

关键词:微表情识别;特征融合;注意力机制;深度学习;迁移学习
中图法分类号　TP183

Micro-expression Recognition Method Combining Feature Fusion and Attention Mechanism

LI Xing-ran,ZHANG Li-yan and YAO Shu-jing
School of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211106,China

Abstract　Micro-expression refers to an uncontrollable muscle movement on the face when people try to hide or suppress their true emotions. Due to the short duration,small motion range,and difficulty in concealing and restraining,the recognition accuracy of such emotional facial expressions is restricted. In order to cope with these challenges,this paper proposes a novel micro-expression recognition method combining feature fusion and attention mechanism,considering optical flow features and face features,and further adding attention mechanism to improve the recognition performance. The processing steps of this method are as follows:1)Extract the optical flow and optical strain from Onset to Apex in each micro-expression segment,input the vertical optical flow,horizontal optical flow and optical strain into a shallow 3DCNN,and extract the optical flow features. 2)Taking the deep convolution neural network ResNet-10 as the backbone network,the convolution attention module is added to extract face features. 3)Combine the two feature vectors for classification. The experimental results reveal that the proposed method is superior to the traditional methods and existing deep learning methods in micro-expression recognition.

Keywords　Micro-expression recognition,Feature fusion,Attention mechanism,Deep learning,Transfer learning

1　引言

表情是“情绪”的直接表征,当一个人试图隐藏自己的真实情绪时,微表情就会出现[1]。与宏表情相比,它能更准确地反映一系列细微甚至无意识的想法[2]。微表情包含大量关于真实情绪的重要且有效的信息,在心理咨询、司法调查等许多领域都有潜在应用[3-4],因此越来越受到相关学科研究者的关注。与普通面部表情相比,微表情持续时间更短,通常为1/25～1/5 s[5-6],这对于人和机器来说都具有挑战性。由于微表情的这一特点,人类很难用肉眼感知微表情,经过专业培训的人员的识别准确率也不超过50%[5]。近年来,卷积神经网络作为一种有效的深度学习模型取得了巨大成功,人脸微表情的自动识别在计算机视觉领域越来越受到重视。

传统的基于手工特征的微表情分析方法包括时空局部二值模式(LBP)[7]、光流直方图(HOOF)和3D方向梯度直方图(3DHOG)等。但是,这些方法从视频中提取的信息大多是表面的,缺乏用于抽象特征表示的必要信息,而且需要复杂的实验设计和烦琐的参数调整才能获得理想的结果。近年来,基于深度学习的卷积神经网络(CNN)得到普及,被广泛应用于人脸识别[8]、车辆识别[9]等多种计算机视觉领域。一般来说,

到稿日期:2021-09-02　返修日期:2021-10-07
基金项目:国家自然科学基金(61772268);江苏省自然科学基金(BK20190065)
This work was supported by the National Natural Science Foundation of China(61772268) and Natural Science Foundation of Jiangsu Province(BK20190065).
通信作者:张立言(zhangly84@126.com)

基于深度学习的技术被证明在大多数计算机视觉问题上优于手工特征技术。最近，一些基于 CNN 的微表情识别方法[10-13]被提出。但是，基于深度学习的方法由于缺乏训练数据，无法学习对微表情识别有帮助的低层次特征，而且微表情非常复杂，类别划分的界限往往不够明确，因此使用三维神经网络(3DCNN)的特征来理解面部细微的运动可以更好地区分微表情[14]。

Liong 等[15]提出一种浅层三流 3DCNN(STSTNet)，将空间和时间信息嵌入到微表情视频片段中，通过轻量级计算来提取具有判别性的高级特征和微表情的细节。虽然光流相比原始数据具有更高的特征，并被证明对微表情识别有效[16]，但是，目前的研究工作只是单独使用光流特征或者人脸特征进行微表情识别，并未考虑光流特征和人脸特征的融合问题。而且，微表情具有区域依赖性，即往往发生在一些较小但特定的面部区域。例如，图 1 中高兴的微表情涉及动作单元 AU12，这会导致嘴角向两侧拉伸，而悲伤时会抬高内眉 AU1；在表达厌恶时 AU4 和 AU9 同时出现，此时两眉中部降低。因此，识别微表情会更关注嘴角和眼睛区域，对面部其他区域的关注较少[17]，而注意力机制有助于关注特定面部区域。

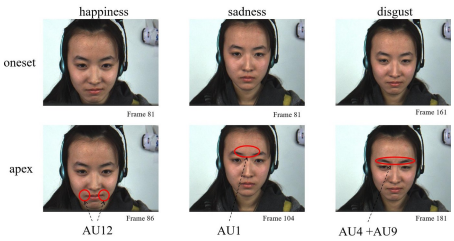

图 1 CASME II 数据集中揭示不同情绪的峰值帧的动作单元

Fig. 1 Different emotions revealed by the action units at apex frame from CASME II datasets

因此，为了解决上述问题，本文提出了一种结合特征融合和注意力机制的微表情识别方法，同时考虑光流特征和人脸特征，通过进一步加入注意力机制来提升识别性能。该网络模型的设计重点体现在如下方面：首先，提取每个微表情片段中 Onset 到 Apex 的光流与光学应变，将垂直光流、水平光流、光学应变输入到一个浅层的 3DCNN 中以提取光流特征；然后，使用深度卷积神经网络 ResNet-10 做迁移学习，引入卷积注意力模块搭建深度学习网络以提取人脸特征；最后，融合两个特征向量进行微表情分类。

本文的主要贡献包括以下几个方面：

(1)引入了结合通道注意力和空间注意力的机制，使网络学习更关注特定的面部区域，提取相对鲁棒的面部特征用于微表情识别。

(2)根据现有知识，本文是首次在特征层次上融合经过 3DCNN 得到的光流特征和加入注意力机制的人脸特征，以构建更具判别性的微表情特征表示的方法。

(3)本文在 3 个自发微表情数据集中进行了实验，结果表明，本文方法在微表情识别方面取得的结果优于其他深度学习方法和传统方法，证明了该方法的有效性。

本文第 2 节介绍了微表情研究的常用方法；第 3 节详细介绍了本文提出的方法和设计原理；第 4 节描述了实验结果并对其进行了分析，然后进行了消融实验；最后总结全文并展望未来。

2 相关工作

近年来，微表情识别受到了越来越多的关注，如何从微表情视频序列中提取微表情特征成为了关键。目前，微表情识别研究大致可以分为两类：1)基于手工特征的方法；2)基于深度学习的方法。早期的微表情识别方法严重依赖于手工特征，如基于局部二值模式(LBP)[7]的算法、光流直方图(HOOF)和 3D 方向梯度直方图(3DHOG)。LBP 是一种基于纹理的特征提取方法，具有判别能力强、表示紧凑和计算复杂度低的特点。对于微表情视频片段，不同帧之间的动态信息至关重要。Pfister 等[18]使用三正交平面局部二值模式(LBP-TOP)作为时空局部纹理描述符来提取动态特征，并通过支持向量机(SVM)、多核学习(MKL)和随机森林(RF)方法来输出识别结果。基于光流的方法[16,19-21]和基于梯度的描述符[22-23]也被广泛应用。基于光流的特征描述符可以推断出不同帧之间的相对运动信息，以便捕获微表情识别的细微肌肉运动。Liu 等[16]将光流引入微表情领域，提出了主方向平均光流(MDMO)，使用 RF 作为分类器，取得了比已有方法更高的准确率。在使用 3DHOG 进行微表情识别的工作中，Chen 等[24]提出了加权 3DHOG 特征，并利用模糊分类对微表情进行识别。

然而，这些方法从视频中提取的信息大多是表面的，缺乏用于抽象特征表示的必要信息[25]，而且需要复杂的实验设计和烦琐的参数调整才能获得理想的结果。近年来，基于深度学习的卷积神经网络(CNN)等方法取得了前所未有的进展，被广泛应用于人脸识别[8]、车辆识别[9]等多种计算机视觉领域。一般来说，基于深度学习的技术被证明在大多数计算机视觉问题上优于手工特征技术。Kim 等[10]最先将 CNN 用于微表情分析，提出一种新的微表情特征表示方法，该方法使用 CNN 对不同时间状态下的空间信息进行编码，这些编码的特征被输入到长短期记忆网络(LSTM)中，用于学习时间特征。Zhao 等[26]使用 4 个卷积层和 3 个池化层来捕获判别性的高级微表情特征，在输入层之后加入 1×1 卷积层，在不增加模型计算负荷的情况下增加输入数据的非线性表达。Wang 等[27]提出一种与残差网络协同的微注意力机制，微注意力由 10 个残差块构成，其微注意力单元是 ResNet 的扩展。为了标识映射而设计的 shortcut 连接可以减少退化问题，通过学习特征图的空间注意力，网络可以聚焦于面部的细微运动。Quang 等[28]试图将 Capsule 引入微表情识别工作中，提出一种简单而有效的微表情识别网络 CapsuleNet。首先识别出所有片段的 Apex 帧；然后仅将 Apex 帧的像素值输入到预训练的 Resnet18 中，得到一系列 28 * 28 的特征图；再将特征图输入到 CapsuleNet 中以提取特征并进行分类。Sun 等[29]提出一种包含 AUs 识别和人脸视图分类的多任务教师网络，然后将学到的知识提炼为微表情识别的浅层学生网络。Peng 等[11]提出的双流卷积神经网络模型(DTSCNN)是第一个针对微表情识别的端到端中型神经网络，该网络提取数据集

每两帧之间的光流，利用两路神经网络学习光流的变化，将SVM分类器的结果用于决策级融合，给出总体识别率。这种直接将光流送到CNN中提取信息的方式并未用到人脸信息。Li等[12]提出利用以灰度帧序列、垂直和水平光学图像为输入数据的三维融合CNN来提取深度学习特征，以进行微表情识别，这种在图像层次进行的融合容易受背景、光照的影响。

基于深度学习的方法提高了识别精度，但由于缺乏训练数据，无法学习对微表情识别有帮助的低层次特征，而且微表情非常复杂，类别划分的界限往往不够明确，因此使用3DCNN的特征来理解细微的运动可以更好地区分微表情[14]。Zhi等[30]采用3DCNN进行自学习特征提取，考虑到效率和计算复杂性，所有卷积核的大小都设置为3×5×5，其中3是时间深度，5×5为空间接受场的大小。Reddy等[25]利用CNN框架中的时空信息，提出了两种针对自发微表情识别的3DCNN模型，即MicroExpSTCNN和MicroExpFuseNet。Wu等[31]提出了一种结合2D和3D卷积神经网络(TSNN)的三信息流来提取表情序列特征。该框架充分利用单个3D核大小和多个3D核组合的优点，提升了微表情的识别性能。Liong等[15]提出一种浅层三流3DCNN(STSTNet)，将空间和时间信息嵌入微表情视频片段中，通过轻量级计算提取具有判别性的高级特征和微表情的细节。该网络在基于每个视频的起始帧和峰值帧计算出的3个光流信息(即水平光流特征、垂直光流特征和光学应变特征)中进行特征提取融合，并利用融合特征对微表情进行识别。这些工作都使用3DCNN同时提取时间-空间域特征，特征信息更加丰富。

3 结合特征融合和注意力机制的方法

本文提出的结合特征融合和注意力机制的网络模型同时考虑了光流特征和人脸特征。此外，尽管可获得的数据较少，但效果并未受到影响。该网络由3个部分组成：1)光流特征分支，提取每个微表情片段中Onset到Apex的光流与光学应变，将垂直光流、水平光流和光学应变输入到一个浅层3DCNN中，以提取光流特征；2)人脸特征分支，以ResNet-10为迁移模型，加入卷积注意力模块以提取人脸特征；3)将两个特征向量拼接起来进行分类。详细的模型框架如图2所示。

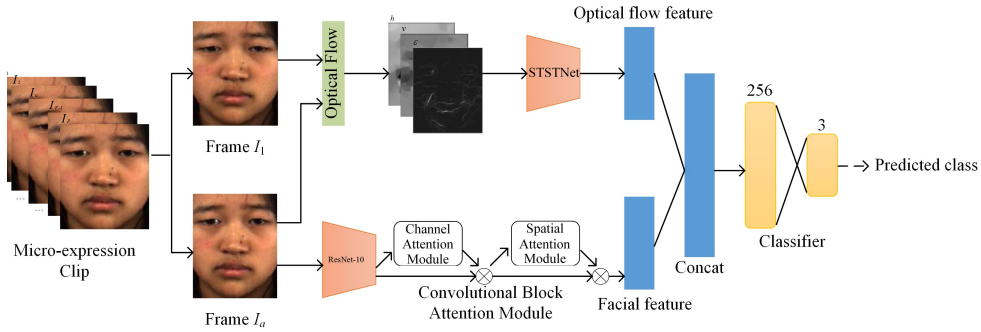

图2 所提模型框架图

Fig. 2 Overall structure of the proposed framework

本文将一个微表情视频片段记为：$S_i=\{I_j \mid i=1,\cdots,n; j=1,\cdots,F\}$，其中$F$为第$i$个序列的总帧数，取自$n$个视频序列的集合。对于每个视频序列，只有一个峰值帧$\boldsymbol{I}_a \in \boldsymbol{I}_1,\cdots,\boldsymbol{I}_F$，它可以位于任何帧的索引处。预测开始的时刻(假设为onset)和峰值帧(apex)的光流向量分别表示为$\boldsymbol{I}_1$和$\boldsymbol{I}_a$。因此，每段分辨率为$X\times Y$的视频只产生一组光流图，表示为：

$$\boldsymbol{O}_i=\{(\boldsymbol{h}_{x,y},\boldsymbol{v}_{x,y}) \mid x=1,\cdots,X; y=1,\cdots,Y\}$$

对于$i\in 1,\cdots,n$，$\boldsymbol{h}_{x,y}$，$\boldsymbol{v}_{x,y}$分别是光流场O_i的水平方向和垂直方向的位移矢量。$\sigma(\cdot)$表示sigmoid激活函数，$\otimes$表示逐元素乘法。

3.1 光流特征分支

考虑到应变模式仅与面部的形变有关，不易受光照条件、面部遮挡等因素影响，在微表情识别任务中有较好的表现，因此给定光流向量，可以推导出光学应变这个特征来描述面部运动模式。对于足够小的面部像素运动，它可以表示面部肌肉组织的形变大小，即：

$$\boldsymbol{\varepsilon}=\frac{1}{2}[\nabla\boldsymbol{u}+(\nabla\boldsymbol{u})^{\mathrm{T}}] \tag{1}$$

其中，$\boldsymbol{u}=[\boldsymbol{h},\boldsymbol{v}]^{\mathrm{T}}$是位移矢量，表示三维空间中面部表情形变导致的位移在二维图像上的投影；∇表示对$\boldsymbol{u}$进行求导。也可以将其展开为矩阵形式：

$$\boldsymbol{\varepsilon}=\begin{bmatrix}\boldsymbol{\varepsilon}_{xx}=\dfrac{\partial\boldsymbol{h}}{\partial x} & \boldsymbol{\varepsilon}_{xy}=\dfrac{1}{2}\left(\dfrac{\partial\boldsymbol{h}}{\partial y}+\dfrac{\partial\boldsymbol{v}}{\partial x}\right)\\ \boldsymbol{\varepsilon}_{yx}=\dfrac{1}{2}\left(\dfrac{\partial\boldsymbol{v}}{\partial x}+\dfrac{\partial\boldsymbol{h}}{\partial y}\right) & \boldsymbol{\varepsilon}_{yy}=\dfrac{\partial\boldsymbol{v}}{\partial y}\end{bmatrix} \tag{2}$$

其中，对角应变分量($\boldsymbol{\varepsilon}_{xx}$，$\boldsymbol{\varepsilon}_{yy}$)是法向应变分量，而($\boldsymbol{\varepsilon}_{xy}$，$\boldsymbol{\varepsilon}_{yx}$)是切应变分量。具体来说，法向应变测量沿特定方向的长度变化，而切应变测量两个角度的变化。由于微表情运动过程中肌肉运动可能包含多个方向，因此每个像素的光学应变大小可以通过法向应变分量和切应变分量的平方和来计算，表达式如下：

$$|\boldsymbol{\varepsilon}_{x,y}|=\sqrt{\boldsymbol{\varepsilon}_{xx}{}^2+\boldsymbol{\varepsilon}_{yy}{}^2+\boldsymbol{\varepsilon}_{xy}{}^2+\boldsymbol{\varepsilon}_{yx}{}^2} \tag{3}$$

将光学应变附加到光流场$\boldsymbol{O}_i$中，每个视频可以推导出一个由3种基于光流表示的三元组$\boldsymbol{\theta}=\{\boldsymbol{h},\boldsymbol{v},\boldsymbol{\varepsilon}\}\in\boldsymbol{R}^3$。

该分支选择STSTNet[15]网络作为骨干网络的原因是，该网络能够通过轻量计算提取具有判别性的高级特征和微表情的细节。按照文献[32]中的建议，使用STSTNet进一步学习光流特征时，将上述描述的光流三元组$\boldsymbol{\theta}$进行重采样之后作为输入传递到网络中。然后，图像经过3个并行流，每个流都由一个卷积层(每个流有不同数量的核)和一个最大池化层

组成。该设计通过在每个流上利用不同数量的 3×3 内核来补充小规模输入数据，从而解决数据不足的问题。此外，利用最大池化操作在消除冗余的同时突出主要特征，最后将输出通道合并形成 3D 光流特征块。具体的网络配置参见文献[15]。

3.2 人脸特征分支

考虑到视频中微表情帧的细微运动的变化，使用峰值帧表示整个微表情序列不仅可以最大程度地减少重复输入帧的冗余，而且可以降低特征学习的计算复杂度。Sun 等[29]的实验结果表明，仅利用微表情峰值帧提供的信息就可以获得较好的识别准确率，且识别效果优于 onset-apex-offset 序列和整个视频。为了提取峰值帧序列的人脸特征，该分支的基础网络采用的是基于注意力机制的 ResNet-10 模型，它是当前应用最为广泛的 CNN 特征提取网络。文献[33]中的实验结果表明，在视频样本的所有面部区域中，“眼＋眉”和“嘴”区域是最有表现力的。换句话说，这些区域贡献了大部分有意义的微表情信息。由于微表情运动具有局部性特点，只出现在人脸的部分区域，大多数面部区域并不存在有效的分类特征，只有少数存在微表情运动的区域才能提供有助于微表情分类的信息，而注意力机制有助于关注特定的面部区域，学习和获得重要特征。因此，在基础架构 ResNet-10 中，通过融合通道注意力和空间注意力来提取人脸微表情特征，通道注意力模型关注各通道间的特征信息，空间注意力模型关注通道内的局部位置信息。借鉴 CBAM 卷积模块的排列方式，将通道注意力模块和空间注意力模块顺序排列。图 3 给出了在 ResNet-10 残差块中集成了 CBAM 的网络模型。

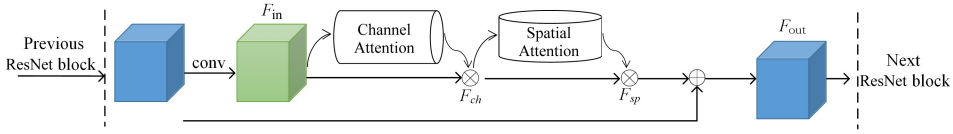

图 3 所提残差注意力模块

Fig. 3 Proposed residual attention module

训练 ResNet-10 进行微表情识别。为了减少深度学习网络在训练微表情数据集时的过拟合问题，考虑引入迁移学习的方法。这种基于迁移学习的方法被证明是在小型数据库上应用深度 CNN 的有效方法。一种通用的迁移方式是在新的数据集上进行参数微调。本文遵循文献[34]中的思想，首先使用 ImageNet 数据集初始化原始残差网络（没有注意力单元）；然后使用 4 个宏表情基准（CK＋[35]、Oulu-CASIA NIR&VIS 面部表情[36]、Jaffe[37]和 MUGFE[38]）进行预训练，有关这些数据集的详细信息将在第 4 节提供；最后，通过微表情数据集（CASMEII[39]，SAMM[40]和 SMIC[41]）对残差网络和注意力单元进行微调。

通道注意力将特征图的每个通道作为特征检测器来关注给定图像中有意义的内容。此外，在共享网络中同时利用平均池化和最大池化，其中前者捕获目标属性范围，后者收集与判别属性特征相关的线索。Woo 等[42]已经证明，采用平均池化和最大池化两个池化操作来聚合通道信息更有利于模型进行注意力推断。令 F_{in} 表示图 2 中卷积层的输出，MLP 表示具有一个隐藏层的多层感知机，则通道注意力模块的输出为：

$$F_{ch}=F_{in}\otimes\sigma(MLP(AvgPool(F_{in}))+MLP(MaxPool(F_{in}))) \tag{4}$$

与通道注意力不同，空间注意力捕捉信息区域。空间注意力也采用了平均池化和最大池化，其输出通过卷积层进行级联卷积，记为 $Conv_F$。空间注意力模块的输出为：

$$F_{sp}=F_{ch}\otimes\sigma(Conv_F([AvgPool(F_{ch}),MaxPool(F_{ch})])) \tag{5}$$

因此，每个残差注意块的输出可以记为 $F_{out}=F_{sp}+F_{pout}$，其中 F_{pout} 表示前一个残差注意块的输出。

3.3 特征融合

在连接光流和人脸这两个特征向量之前，对两个分支的输出进行 L2 归一化处理，可以写成：

$$C_i=\frac{O_i}{\sqrt{\sum_{d=1}^{D}O_i^d}} \tag{6}$$

其中，$O_i^d(i=1,2;d=1,\cdots,512)$ 是每个分支的输出元素，C_i 是归一化输出。归一化后，将来自两个分支的特征向量拼接起来，这个过程中原始信息不会损失，最终进行分类并得到总体识别率。

4 实验

实验验证了本文提出的结合特征融合和注意力机制的微表情识别方法的有效性。首先描述本文使用的数据集及其性能度量方法；然后将本文方法与其他最新方法进行比较与分析；最后针对本文方法进行消融实验，以分析模型的稳定性。

4.1 数据集与性能度量

为了验证本文方法的有效性，实验采用 MEGC2019[43]复合数据集对其进行了验证。复合数据集由 CASME II，SAMM 和 SMIC 这 3 个自发式微表情数据集组合而成，包含来自 68 个对象的 442 个微表情样本。表 1 列出了这 3 个微表情数据集的详细信息。为了使 3 个数据集可以一起使用，在第二届微表情识别大赛中使用了一组共同的简化情感类别，即“Positive”“Negative”和“Surprise”，并将原有的标签重新映射到新的标签空间中，以更好地避免因不同刺激和环境设置所引发的情感类别的模糊性。跨库数据集在增加训练样本的同时也有利于提升算法的泛化能力，使其更适用于真实场景。这里需要提及的是，为了在峰值帧数量受到限制的情况下进行特征提取，需要扩充数据集以增加样本数量，并防止模型过度拟合。因此，实验使用了一种数据增强策略，即以旋转和缩放的方式进行数据扩充，对样本数较少的这类情感进行多次不同角度的旋转，每个训练序列都在[－10°，10°]范围内旋转，在[0.9，1.1]范围内缩放，样本数较多的这类情感保持原样本数，这大大增大了当前数据集，使其适合训练。

表 1 复合微表情数据集总结

Table 1 Summary of composite micro-expression databases

Micro-Expression Datasets		CASME II	SAMM	SMIC
Subjects		24	28	16
Samples		145	133	164
Frames Per sec		200	200	100
Expression Classes	Negative	88	92	70
	Positive	32	26	51
	Surprise	25	15	43
Frame Annotations		onset, offset, apex	onset, offset, apex	onset, offset

为了避免实验的偶然性及网络在学习中可能存在的偏差，本文将来自 3 个数据集的样本放在一起执行留一受试交叉验证(Leave-One-Subject-Out, LOSO)，每次选取一名受试者的样本作为测试集，将其余受试者的样本作为训练集，迭代交叉验证 k 次，其中 k 为微表情数据集中受试者的数量。MEGC2019 复合数据集包含 68 名受试者的微表情样本，因此需要执行 68 轮 LOSO，即 $k=68$。从表 1 中很容易看出，融合后的数据集中“Surprise”:“Positive”:“Negative”大约为 1:1.3:3，数据集还存在严重的数据类别不平衡问题。因此，为了识别 3 个微表情类别，使用非加权平均召回率(Unweighted Average Recall, UAR)和非加权 F1 得分(Unweighted F1-score, UF1)指标来减小潜在的偏差。

UF1 在多分类问题中是一个很好的选择，因为它可以同等强调稀有类。为了计算 UF1，首先计算每个类别 c 的 F1 值。UF1 就是每个类别的 F1 值的平均值。

$$F1_c=\frac{TP_c}{2TP_c+FP_c+FN_c} \tag{7}$$

$$UF1=\frac{1}{C}\sum_c F1_c \tag{8}$$

其中，C 为微表情的标签数量，$F1_c$ 为类 c 分类结果的 $F1$ 值，TP_c，FP_c 和 FN_c 分别为类 c 分类结果中真阳性、假阳性和假阴性的数量。

UAR 也称“平衡准确率”，这是一种代替加权平均召回的更合理的评价标准，因为加权平均召回的预测更偏向于较大类别的结果。

$$UAR=\frac{1}{C}\sum_c Acc_c \tag{9}$$

$$Acc_c=\frac{TP_c}{N_c} \tag{10}$$

其中，N_c 为类 c 的样本数量；Acc_c 为类 c 的准确率，即 c 类中预测正确的数量占类样本总数量的比例；UAR 为所有类别的准确率的平均值。

为了对 3 个微表情数据集进行训练，本文采用了迁移学习的方法。在预训练步骤中，使用 4 种常用的宏表情数据集，即 CK+、Oulu-CASIA NIR&VIS 面部表情、Jaffe 和 MUGFE。CK+[35]包含 123 名受试者的 593 个视频片段，有愤怒、蔑视、厌恶、恐惧、快乐、悲伤和惊讶 7 种情绪类型，每个片段从正常表情帧开始，到峰值帧结束。实验从每个对应的视频片段中选择最后 3 帧，最终获得 852 张图像。Oulu-CASIA NIR&VIS 面部表情数据集[36]包含 80 名年龄在 23～58 岁的受试者的视频，收集了 6 种表情，即快乐、悲伤、惊讶、愤怒、恐惧和厌恶。视频由近红外(NIR)和可见光(VIS)两套成像系统拍摄。实验选择了 VIS 系统在正常室内照明下拍摄的每个视频的最后 3 帧，最终得到 1 200 张图像。Jaffe[37]包含了由 10 位日本女性拍摄的 219 张图片，每个对象都表现出了悲伤、快乐、愤怒、厌恶、惊讶、恐惧和中立 7 种表情。实验选择了对应的带有正确表情标签的图像进行微表情任务分类，最终得到 151 张图像。MUGFE[38]数据集由 86 名受试者的 1 032 个视频片段组成，每个片段包含 50～160 帧，从中性帧开始到结束，峰值帧对应于其中的各个情感类型。实验从峰值帧周围选取 6～10 帧，最终得到 8 228 张图像。

4.2 实验结果与分析

本节将本文方法与基于手工特征的方法[16,21,41]和一些现有深度学习方法[15,28,32,44-45]进行了比较，在复合(Full)、CASME II、SMIC 和 SAMM 数据集上评估了本文方法。不同方法的 $UF1$ 和 UAR 结果如表 2 所列。实验中涉及的比较方法均使用原论文中的实验结果。

表 2 微表情识别性能比较

Table 2 Comparison of micro-expression performance recognition

	Methods	Full		CASME II		SMIC		SAMM	
		*UF*1	*UAR*	*UF*1	*UAR*	*UF*1	*UAR*	*UF*1	*UAR*
Traditional methods	LBP-TOP[41]	0.5882	0.5785	0.7026	0.7429	0.2000	0.5280	0.3954	0.4102
	MDMO[16]	—	—	0.4966	0.5169	0.5845	0.5879	—	—
	Sparse MDMO[21]	—	—	0.6911	0.6695	0.7041	0.7051	—	—
Deep learning methods	ATNet[44]	0.631	0.613	0.798	0.775	0.553	0.543	0.496	0.482
	CapsuleNet[28]	0.6520	0.6506	0.7068	0.7018	0.5820	0.5877	0.6209	0.5989
	OFF-ApexNet[32]	0.7196	0.7096	**0.8764**	**0.8681**	0.6817	0.6695	0.5409	0.5392
	Dual-Inception[45]	0.7322	0.7278	0.8621	0.8560	0.6645	0.6726	0.5868	0.5663
	STSTNet[15]	0.7353	0.7605	0.8382	0.8686	0.6801	0.7013	0.6588	0.6810
	Proposed	**0.7673**	**0.7779**	0.8537	0.8563	**0.7042**	**0.7118**	**0.6631**	**0.6858**

从表 2 中可以清楚地看出，基于深度学习的特征提取优于传统的手工特征提取。本文提出的模型在复合数据集上获得的 $UF1$ 和 UAR 性能最好，分别为 0.7673 和 0.7779。此外，本文方法在 3 个单独的数据集(CASME II, SMIC 和 SAMM)上实现了比手工方法更好的性能提升。本文模型没有像 LBP-TOP 方法一样使用微表情序列中的所有帧来提取特征，而是仅将每个微表情的单个峰值帧作为输入数据。但是，与 LBP-TOP 方法相比，它的 $UF1$ 和 UAR 仍分别高出 17.91%和 19.94%。通过进一步分析发现，使用所提模型时，CASME II 中的 $UF1$ 和 UAR 在 3 个数据集中是最高的，主要原因是该数据集中的样本量最多。这表明，与 CASME II 数据集相比，SAMM 和 SMIC 数据集的识别更具挑战性。

对于 SAMM 数据集，这很可能是较高的类不平衡问题以及参与者的年龄和种族差异较大导致的。对于 SMIC 数据集，人脸区域分辨率和帧率的降低是性能下降的主要原因。在以后的工作中可以考虑采用 GAN(Generative Adversarial Network)等数据增强方法来增加数据集的数量和类别间样本的平衡性。

为了验证上述分析，并进一步表明该实验结果是从完整的复合数据集中获得的，图 4 给出了本文模型在复合数据集上的分类结果混淆矩阵。可以看出，排除正确的预测样本，大部分其他的积极和惊讶样本都被预测为消极类别，这主要是由于任务类出现类不平衡，消极类别在数据集中占主导地位。

图 4 复合数据集的混淆矩阵

Fig. 4 Confusion matrix of composite dataset

本文进一步利用类激活映射(CAM)来探究加入了注意力模块的网络模型的聚焦区域。图 5 给出了一些加入了 3.2 节中设计的注意力模块后的算法的可视化示例，特征图显示在右边，红色矩形表示信息最丰富的区域。

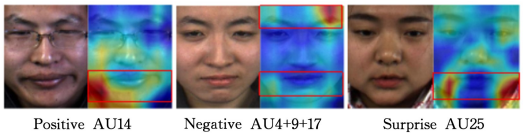

图 5 加入注意力模块的网络特征图可视化(电子版为彩色)

Fig. 5 Visualization of network feature map with attention module

对于样例，微表情“积极”情绪主要发生在受试者的嘴巴区域，“消极”情绪主要发生在受试者的眉毛上方以及嘴巴区域。可以看出，本文提出的网络模型可以提取人脸信息区域的情绪信息且网络的关注区域符合预期，很好地关注到了“眼＋眉”和“嘴”区域。结果表明，注意力机制有助于关注特定面部区域，提高了微表情特征的识别率。

4.3 消融实验

为了验证本文模型加入注意力机制后提取的人脸特征的有效性以及光流特征和人脸特征融合方法的有效性，使用留一交叉验证法处理从视频序列中提取的图像帧，设计了 4 种方法的对比实验。

实验 1 单独的光流特征，使用 STSTNet 网络作为基础架构，提取具有判别性的光流特征。

实验 2 单独的人脸特征，以 ResNet-10 为迁移模型提取人脸特征。

实验 3 加入了注意力机制的人脸特征，以 ResNet-10 为迁移模型，加入卷积注意力模块提取人脸特征。

实验 4 将光流特征和加入注意力机制的人脸特征进行融合。

4 种方法的对比实验结果如表 3 所列。

表 3 使用留一交叉验证法的对比实验结果

Table 3 Comparative experiment results using leave-one-out cross-validation method

方法		实验 1	实验 2	实验 3	实验 4
		光流特征	人脸特征	人脸特征＋attention	光流特征＋人脸特征＋attention
CASME II	*UF*1	0.8382	0.798	0.8002	0.8537
	UAR	0.8686	0.775	0.7779	0.8563
SMIC	*UF*1	0.6801	0.553	0.5583	0.7042
	UAR	0.7013	0.543	0.5439	0.7118
SAMM	*UF*1	0.6588	0.496	0.5090	0.6631
	UAR	0.6810	0.482	0.4835	0.6858

与实验 2 相比，结合注意力机制的人脸特征即实验 3 效果更好。两种网络的主要区别在于注意力机制，证明了利用注意力机制进行微表情学习的有效性。在未来也可以实现该模型光流特征分支的注意力机制。实验 1 相比实验 3 取得了更好的结果，这意味着光流相比原始数据在卷积神经网络模块上具有更好的特征表达能力，对微表情微小的面部肌肉运动有更强的抽取能力。在所有方法中，实验 4 即将光流特征和加入注意力机制的人脸特征融合的结果最优(*UF*1 和 *UAR* 分别达到了 0.7673 和 0.7779)，与使用单一的光流特征或者人脸特征相比有更加突出的表现，证明了特征融合对微表情识别的重要性。

结束语 本文提出了一种结合特征融合和注意力机制的方法来识别微表情。首先，将基于深度卷积神经网络提取的光流特征和人脸特征进行融合，以构建更具判别性的微表情特征表示。其次，引入了通道注意力与空间注意力相结合的机制，以关注特定的面部区域，提取相对鲁棒的面部特征进行微表情识别。该方法既能提取人脸的整体特征，又能集中提取人脸特定面部区域的关键特征。最后，在 CASME II，SMIC 和 SAMM 数据集上评估了本文方法。实验结果表明，该模型的 *UAR* 和 *UF*1 性能优于其他深度学习方法和传统方法，验证了本文方法的有效性。

本文提供了一种有效的微表情识别方法，其仍有进一步提升的空间。例如，现有数据集样本稀缺是目前微表情识别技术发展的最大瓶颈。因此，未来的工作可以考虑通过生成对抗网络(GANs)等数据增强方式来生成大量伪微表情样本，以减小类别分布严重不均衡对识别准确率的影响。此外，微表情在时域上表现稀疏，并不是每一帧都具有能够代表当前微表情种类的信息量，寻找一个潜在的、局部的、代表当前微表情种类的帧组合可能会进一步提高识别精度。

参 考 文 献

[1] EKMAN P. Emotions Revealed：Understanding Faces and Feelings[M]. Weidenfeld & Nicolson，2004.

[2] TAKALKAR M，XU M，WU Q，et al. A survey：facial micro-ex-

pression recognition [J]. Multimedia Tools and Applications, 2018, 77(15): 1-25.

[3] EKMA P. Lie Catching and Microexpressions[M]// The Philosophy of Deception. 2009: 118-133.

[4] FRANK M, HERBASZ M, SINUK K, et al. I see how you feel: Training laypeople and professionals to recognize fleeting emotions[C]// The Annual Meeting of the International Communication Association. New York, 2009: 1-35.

[5] HOUSE C, MEYER R. Preprocessing and descriptor features for facial micro-expression recognition [OL]. [2016-10-15]. https://web. stanford. edu/class/ee368/Project_Spring_1415/Reports/House_Meyer. pdf.

[6] ZHANG M, FU Q, CHEN Y H, et al. Emotional Context Influences Micro-Expression Recognition [J]. PLoS ONE, 2014, 9(4): 1-7.

[7] OJALA T, PIETIKÄINEN M, HARWOOD D. A Comparative Study of Texture Measures with Classification Based on Feature Distributions [J]. Pattern Recognition, 1996, 29(1): 51-59.

[8] JI C M, SONG T C. Sparse Representation-Based Classification Under Optimization Forms for Face Recognition[J]. Journal of Chongqing University of Technology (Natural Science), 2020, 34(2): 120-126.

[9] RAO W J, GU Y H, ZHU T T, et al. Intelligent license plate recognition method in complex environment [J]. Journal of Chongqing University of Technology (Natural Science), 2021, 35(3): 119-127.

[10] KIM D H, BADDAR W J, RO Y M. Micro-expression recognition with expression-state constrained spatio-temporal feature representations[C]// Proceedings of the 24th ACM International Conference on Multimedia. 2016: 382-386.

[11] PENG M, WANG C Y, CHEN T, et al. Dual Temporal Scale Convolutional Neural Network for Micro-Expression Recognition [J]. Frontiers in Psychology, 2017, 8: 1-12.

[12] LI J, WANG Y, SEE J, et al. Micro-expression recognition based on 3D flow convolutional neural network[J]. Pattern Analysis and Applications, 2019, 22(4): 1331-1339.

[13] LIANG Z Y, HE J L, SUN Y. An evolutionary method of three-dimensional convolutional neural networks for automatic recognition of micro expressions [J]. Computer Science, 2020, 47(8): 227-232.

[14] MERGHANI W, DAVISON A K, YAP M H. A review on facial micro-expressions analysis: datasets, features and metrics[J]. arXiv: 1805. 02397, 2018.

[15] LIONG S T, GAN Y S, SEE J, et al. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition [C]// IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019: 1-5.

[16] LIU Y J, ZHANG J K, YAN W J, et al. A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition [J]. IEEE Transactions on Affective Computing, 2016, 7(4): 299-310.

[17] CHEN B, ZHANG Z, LIU N, et al. Spatiotemporal Convolutional Neural Network with Convolutional Block Attention Module for Micro-Expression Recognition [J]. Information (Switzerland), 2020, 11(8): 380.

[18] PFISTER T, LI X, ZHAO G, et al. Recognising spontaneous facial micro-expressions[C]// International Conference on Computer Vision. IEEE, 2011: 1449-1456.

[19] XU F, ZHANG J, WANG J Z. Microexpression identification and categorization using a facial dynamics map[J]. IEEE Transactions on Affective Computing, 2017, 8(2): 254-267.

[20] LIONG S T, SEE J, WONG K S, et al. Less is more: Micro-expression recognition from video using apex frame[J]. Signal Processing: Image Communication, 2018, 62: 82-92.

[21] LIU Y J, LI B J, LAI Y K. Sparse MDMO: Learning a discriminative feature for micro-expression recognition[J]. IEEE Transactions on Affective Computing, 2018, 12(1): 254-261.

[22] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2005, 1: 886-893.

[23] POLIKOVSKY S, KAMEDA Y, OHTA Y. Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor[C]// International Conference on Crime Detection & Prevention. IET, 2010.

[24] CHEN M, MA H T, LI J, et al. Emotion recognition using fixed length micro-expressions sequence and weighting method[C]// IEEE International Conference on Real-time Computing and Robotics (RCAR). IEEE, 2016: 427-430.

[25] REDDY S, KARRI S T, DUBEY S R, et al. Spontaneous Facial Micro-Expression Recognition using 3D Spatiotemporal Convolutional Neural Networks[C]// International Joint Conference on Neural Networks (IJCNN). IEEE, 2019: 1-8.

[26] ZHAO Y, XU J. A convolutional neural network for compound micro-expression recognition[J]. Sensors, 2019, 19(24): 5553.

[27] WANG C, PENG M, BI T, et al. Micro-attention for micro-expression recognition[J]. Neurocomputing, 2020, 410: 354-362.

[28] QUANG N V, CHUN J, TOKUYAMA T. CapsuleNet for Micro-Expression Recognition[C]// IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019: 1-7.

[29] SUN B, CAO S, LI D, et al. Dynamic Micro-Expression Recognition Using Knowledge Distillation[C]// IEEE Transactions on Affective Computing. 2020.

[30] ZHI R, XU H, WAN M, et al. Combining 3D convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition[J]. IEICE Transactions on Information and Systems, 2019, 102(5): 1054-1064.

[31] WU C, GUO F. TSNN: Three-Stream Combin-ing 2D and 3D Convolutional Neural Network for Micro-Expression Recognition[J]. IEEJ Transactions on Electrical and Electronic Engineering, 2021, 16(1): 98-107.

[32] GAN Y S, LIONG S T, YAU W C, et al. OFF-ApexNet on micro-expression recognition system[J]. Signal Processing: Image

Communication,2019,74:129-139.

[33] LIONG S T,SEE J,WONG K S,et al. Automatic Apex Frame Spotting in Micro-expression Database[C]// IAPR Asian Conference on Pattern Recognition. IEEE,2015:665-669.

[34] PENG M,WU Z,ZHANG Z,et al. From macro to micro expression recognition:Deep learning on small datasets using transfer learning[C] // IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE,2018:657-661.

[35] 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]// Computer Vision & Pattern Recognition Workshops. IEEE,2010:94-101.

[36] ZHAO G,HUANG X,TAINI M,et al. Facial expression recognition from near-infrared videos[J]. Image and Vision Computing,2011,29(9):607-619.

[37] 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. IEEE,1998:200-205.

[38] AIFANTI N,PAPACHRISTOU C,DELOPOU-LOS A. The MUG facial expression database[C]// International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10. IEEE,2010:1-4.

[39] YAN W J,LI X,WANG S J,et al. CASME II:an improved spontaneous micro-expression database and the baseline evaluation[J]. PLoS one,2014,9(1):1-8.

[40] DAVISON A K,LANSLEY C,COSTEN N,et al. SAMM:A Spontaneous Micro-Facial Movement Dataset[J]. IEEE Transactions on Affective Computing,2018,9(99):116-129.

[41] LI X,PFISTER T,HUANG X,et al. A Spontaneous Micro-expression Database:Inducement,collection and baseline[C]// IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). IEEE,2013:1-6.

[42] WOO S,PARK J,LEE J Y,et al. CBAM:Convolutional Block Attention Module[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018:3-19.

[43] SEE J,YAP M H,LI J,et al. MEGC 2019—the second facial micro-expressions grand challenge[C] // IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE,2019:1-5.

[44] PENG M,WANG C,BI T,et al. A novel apex-time network for cross-dataset micro-expression recognition[C]// 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE,2019:1-6.

[45] ZHOU L,MAO Q,XUE L. Dual-inception network for cross-database micro-expression recognition[C]// IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE,2019:1-5.

LI Xing-ran, born in 1998, postgraduate. Her main research interests include micro-expression recognition and deep learning.

ZHANG Li-yan, born in 1984, Ph. D, professor, is a member of China Computer Federation. Her main research interests include multimedia analysis, computer vision and deep learning.

(责任编辑:喻藜)