DOI:10.13878/j.cnki.jnuist.20221118001

冉瑞生[1]　石凯[1]　江小鹏[1]　王宁[1]

基于双注意力 CrossViT 的微表情识别方法

摘要

微表情是人们试图隐藏自己真实情绪时不由自主泄露出来的面部表情,是近年来情感计算领域的热点研究领域.微表情是一种细微的面部运动,难以捕捉其细微变化的特征.本文基于交叉注意力多尺度 ViT(CrossViT)在图像分类领域的优异性能以及能够捕捉细微特征信息的能力,将 CrossViT 作为主干网络,对网络中的交叉注意力机制进行改进,提出了 DA 模块(Dual Attention)以扩展传统交叉注意力机制,确定注意力结果之间的相关性,从而提升了微表情识别精度.本网络从三个光流特征(即光学应变、水平和垂直光流场)中学习,这些特征是由每个微表情序列的起始帧和峰值帧计算得出,最后通过 Softmax 进行微表情分类.在微表情融合数据集上,UF1 和 UAR 分别达到了 0.727 5 和 0.727 2,识别精度优于微表情领域的主流算法,验证了本文提出网络的有效性.

关键词

微表情识别;CrossViT;交叉注意力机制;光流特征

中图分类号 TP391.4

文献标志码 A

收稿日期 2022-11-18

资助项目 重庆市技术创新与应用发展专项面上项目(cstc2020jscx-msxmX0190);重庆市教委科学技术研究重点项目(KJZD-K202100505)

作者简介

冉瑞生,男,博士,教授,主要从事机器学习、计算机视觉等方面的研究.rshran@cqnu.edu.cn

1 重庆师范大学 计算机与信息科学学院,重庆,401331

0 引言

微表情是人们试图隐藏自己真实情绪时不由自主泄露出来的面部表情,即使是专业演员也很难伪装.除了日常生活中普通的面部表情,在某些情况下,情绪也会以微表情的形式表现出来.与普通的面部表情相比,微表情的持续时间仅有 1/25~1/3 s[1],并且参与的肌肉运动强度很微弱[2].因此,微表情可以被视为推断人类情绪的可靠线索之一,这使得它们在司法系统、刑侦审讯和临床诊断中得到广泛应用.

由于微表情识别的广泛应用性,近年来,研究者开展了大量的研究.这些研究主要分为基于传统机器学习的方法和基于深度学习的方法.在传统机器学习方法中,特征提取是影响算法性能的关键.局部二值模式(LBP)[3]是一种特征提取算法,它根据当前像素值对相邻像素进行阈值处理,有效地描述了图像纹理特征.此后,针对微表情识别任务还提出了多种 LBP 算法,如三正交局部二值模式(LBP-TOP)[4]和六交叉点局部二值模式(LBP-SIP)[5].Huang 等[6]提出一种积分投影方法,将形状属性与时空纹理特征相结合,实现了微表情识别的判别时空局部二元模式(SLBP).此外,还存在两个时空描述符:主方向平均光流(MDMO)[7]和人脸动态图(FDM)[8].Liu 等[9]进一步将 MDMO 纳入经典的图正则化稀疏编码中,生成了稀疏 MDMO 特征.马浩原等[10]提出平均光流直方图(MHOOF),提取相邻两帧间感兴趣区域的 HOOF 特征以检测峰值帧,将峰值帧和起始帧的 MHOOF 特征用于微表情识别.Liong 等[11]提出了双加权定向光流(Bi-WOOF)特征描述符,将光流幅值和光学应变大小作为权值,生成人脸区域各块的方向直方图进行微表情识别.

传统方法需要繁琐的手工特征设计,而且微表情识别的准确率低.考虑到深度学习在面部表情识别中取得的良好表现,研究人员开始试图将深度学习应用于微表情的识别任务.Quang 等[12]首次将胶囊网络(CapNet)[13]应用于微表情识别模型中,该模型设计简单,所需的训练数据很少,并且具有很强的鲁棒性.Lai 等[14]则通过在 VGG 网络中添加残差连接,增加网络深度的同时也缓解了梯度消失的问题,在该研究中还使用了空洞卷积替换传统卷积,扩大感受野的同时也能够捕捉多尺度的上下文信息.Wang 等[15]在 ResNet 网络上进行改进,在网络中添加微注意力提升模型对面部区域的关注,从而提升识别的精度.Liong 等[16]提出一种利用光流特征进行微表情检测和识别的

方法,它可以更好地表现精细、微妙的面部运动.在此基础上,Liong 等[17]进一步提出了浅三流三维 CNN(STSTNet),并利用光流特征训练网络.这些研究表明,由于微表情数据集样本数量小,浅层神经网络更适合于微表情识别任务.此外,Verma 等[18]也试图通过递增的方式提取更显著的表情特征,来捕捉面部区域每个表情的微观层面特征.Khor 等[19]引入长期循环卷积网络(ELRCN)模型用于微表情识别,该模型通过结合深度空间特征学习模块和时间特征学习模块对微表情特征进行编码.

目前主流的微表情识别算法一般是采用卷积网络提取特征.Zhao 等[20]提出 6 层 CNN 网络进行特征提取.Khor 等[21]提出一个轻量级的双流浅层网络,其网络整体由 CNN 组成.Zhi 等[22]将 CNN 与 LSTM 串联起来,直接处理不同时长的微表情序列.

Transformer 是一种主要基于自注意力机制的深度神经网络,最初应用于自然语言处理领域.受到 Transformer 强大的表示能力的启发,研究人员开始提出将 Transformer 扩展到计算机视觉任务.Ma 等[23]首次将 Transformer 架构应用到表情识别中,在该网络中首先使用 ResNet18 提取输入图像的特征图,最后再放入多层 Transformer 编码器中进行分类.Zhang 等[24]提出 SLSTT 网络,该网络结构将微表情序列光流特征送入到 Transformer 编码器中,通过 LSTM 架构对时间和空间特征融合后进行分类.刘忠洋等[25]基于注意力机制进行多尺度特征融合,证明了多尺度特征融合在图像分类上的有效性.

Chen 等[26]提出一种双分支的 Transformer 分别提取不同尺度特征以及基于 CrossAttention 的融合机制融合不同分支的特征.对于视觉 Transformer,通过改进自注意力机制能够有效提升网络的性能.Huang 等[27]扩展了传统的自注意力机制,以确定注意力结果和查询结果的相关性.杨春霞等[28]提出的基于 BERT 与注意力机制融合的模型,表明 Transformer 架构在关于情感分析任务中有较好的表现.受上述文献启发,本文对于注意力机制进行了改进,以提升微表情识别精度.

Huang 等[27]研究表明,Transformer 编码器中的自注意力机制所提取的特征中包含了一些冗余和无用的特征信息,在微表情领域,这些冗余和无用的特征信息不利于后续的微表情识别任务.另外,由于微表情是一种面部运动幅度很低的情感表达,传统的卷积神经网络难以捕捉到这些细微的特征.而人们最近提出的多尺度网络较传统卷积网络能够捕捉更加细微的特征信息[25-26],以获得更加丰富的特征信息用于微表情识别.基于此,本文将交叉注意力多尺度 ViT(CrossViT)网络进行改进并应用到微表情识别上,实验表明提出的方法取得了较好的识别效果.本文的贡献有如下几点:

1) 本文所提出的模型较早地将 CrossViT 网络应用到微表情领域,证明了其在微表情识别上的有效性;

2) 本文对 CrossViT 网络中原有的注意力机制进行了改进,提出了 DA(Dual Attention)模块,该模块扩展了传统交叉注意力机制,确定注意结果和查询之间的相关性,以保留网络中有用的特征信息,从而有效提升了网络的识别性能;

3) 本文所提出的模型在 CASME Ⅱ、SMIC 和 SAMM 三个数据集上均取得了良好的识别性能,验证了本文模型在微表情识别上的有效性.

1 相关工作

1.1 光流特征

微表情识别的早期研究方法主要是基于手工特征的传统机器学习方法.这些手工特征是利用设计好的特征提取算子提取对应的特征,并将特征送入 SVM 等分类器进行微表情分类.手工特征提取的方法可以分为两种:第一种是基于表观特征的方法,该方法考虑到图像的像素之间的关系并进行相应特征的提取,可以得到微表情序列的动态纹理信息,如 LBP[3]、LBP-TOP[4]等;第二种是基于几何特征的方法,该方法考虑到图像局部特征区域和特征点的位移和形变,进行相应的特征提取.光流特征是一种基于几何特征的特征提取方法,其基于光流的特征描述符推断不同帧之间的相对运动,能为微表情识别捕获微表情连续帧之间的时间特征.

光流特征中的光流是指空间运动物体在观察成像平面上的像素运动的瞬时速度.其特征提取是利用图像序列中像素在时间域上的变化以及相邻帧之间的相关性来找到上一帧跟当前帧之间存在的对应关系,从而计算相邻帧之间的运动信息,通过 TVL1 光流法可以计算出微表情序列中起始帧和峰值帧之间的水平和垂直光流矢量.光流应变代表的是人脸运动变化强度,能够作为加权方案,以突出每个光流的重要,从而减少了小强度的光流噪声.每个像素点的光流应变可以通过计算水平和垂直的光流矢量的

平方和进行计算.

1.2 CrossViT

CrossViT 是将两个不同分支的图像标记,通过交叉注意力进行类标记融合.CrossViT 的整体网络架构如图 1 所示.该网络主要是由 K 个多尺度 Transformer 编码器(图 1 中黄色区域)组成,将光流特征图送入到 2 个不同尺度的分支中.

1) La 分支对粗粒度的特征块进行操作,在该分支中,将原始光流特征图作为特征块输入,然后将特征块扁平成一维向量后通过投影函数得到更大的嵌入向量,并经过 M 个 Transformer 编码器进行特征提取;

2) Sm 分支对细粒度的特征块进行操作,在该分支中,输入光流特征图被划分成 4 个大小相同的特征块,将每个特征块扁平成一维向量后通过投影函数得到更小的嵌入向量,并经过 N 个 Transfomer 编码器进行特征提取,其中 $M>N$.

经过 2 个分支 Transformer 编码器提取的粗粒度与细粒度的特征信息送入到交叉注意力(CrossAttention)模块进行 L 次信息交互以获得更丰富的类标记,最后将 2 个分支的类标记拼接后输出,将得到的特征信息输入到分类器.

交叉注意力是 CrossViT 的重要模块,它是在自注意力的基础上提出的一种注意力机制.交叉注意力最早是用于 Transformer 的编码器和解码器的连接.不同于自注意力,交叉注意力能够将不同尺度、不同模态的特征进行关联,提升模型的性能.近年来,CrossViT 在多尺度图像分类上取得成功,证明了交叉注意力机制可以处理不同形式的内容,并且能够融合不同尺度的数据.

考虑到微表情是细微的面部运动,将光流特征图划分为不同尺度的图像特征图,大尺度的图像特征图表示了微表情高层的语义信息,小尺度的特征图表示微表情低层的细节信息,将不同尺度的特征信息融合,有利于获得更丰富的特征表示.网络中的视觉 Transformer 相较于卷积神经网络具有全局感受野,并且其中的查询、键和值都依赖于输入数据,因此其具有自适应权重聚合的特性,能够获得更好的特征表示.

2 双注意力 CrossViT

2.1 双注意力模块

CrossViT 网络的最大特点是利用交叉注意力模块实现不同尺度的特征交互,获得更有代表性的特征信息.然而 CrossViT 原有的交叉注意力模块虽然能够将不同尺度的信息进行交互,但是交互后保留

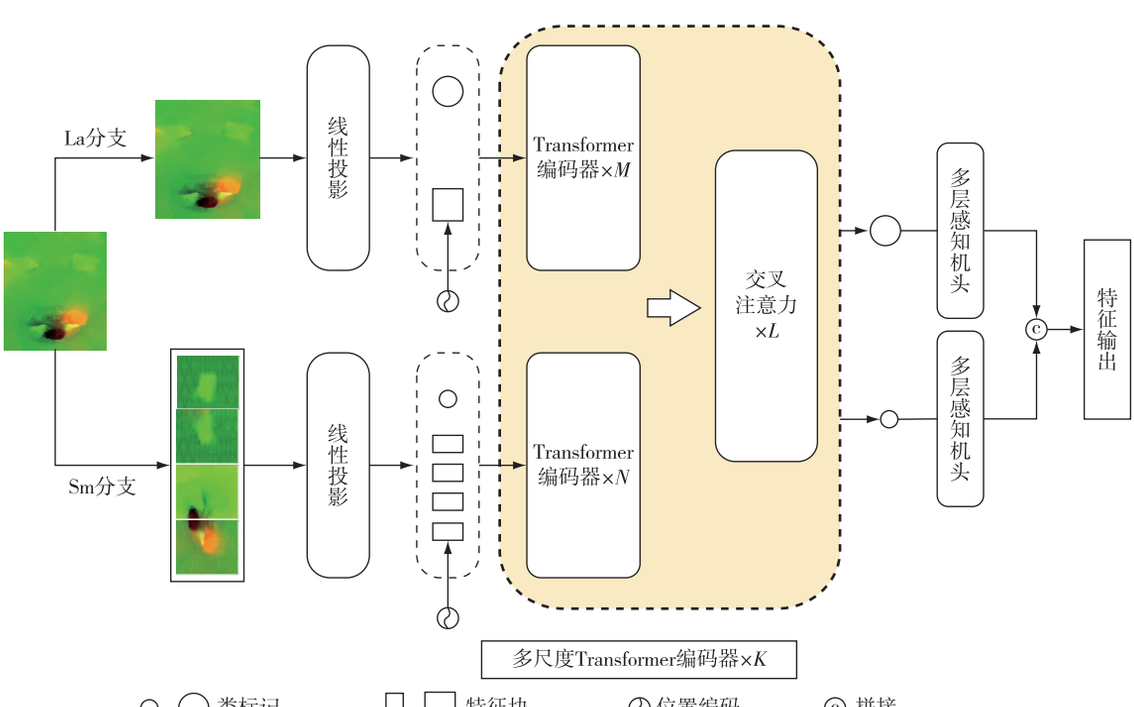

图 1 CrossViT 网络架构

Fig. 1 CrossViT network architecture

了很多无用的特征信息,不利于之后的图像分类任务.

本文对 CrossViT 的交叉注意力模块进行了改进,在该模块中额外增加一个注意力机制,改进的模块能够过滤掉无用的注意力结果,只保留有用的注意力结果,从而提升了微表情识别的精度.提出的双注意力模块(DA)如图 2:用不同大小的方块表示不同尺度的特征块,用不同大小的圆圈表示粗粒度和细粒度的类标记.由 Tranformer 编码器输出的粗粒度的特征块和类标记和细粒度的特征块作为 DA 模块的输入,具体来说,La 分支首先将自身的类标记与 Sm 分支的特征块连接在一起,如下式所示:

$$\boldsymbol{x}'^{1}=[f^{1}(\boldsymbol{x}^{1}_{\mathrm{cls}})\,\|\,\boldsymbol{x}^{\mathrm{s}}_{\mathrm{patch}}], \tag{1}$$

式(1)中,$f^{1}(\cdot)$ 是维度对齐的投影函数,$\boldsymbol{x}^{1}_{\mathrm{cls}}$是 La 分支的类标记向量,$\boldsymbol{x}^{\mathrm{s}}_{\mathrm{patch}}$是 Sm 分支的特征块向量,将 $\boldsymbol{x}'^{1}$ 和 $\boldsymbol{x}^{1}_{\mathrm{cls}}$之间执行 DA 注意力操作,其中类标记作为唯一的查询.在数学上,查询向量($\boldsymbol{q}$)、键向量($\boldsymbol{k}$) 和值向量($\boldsymbol{v}$) 及注意结果 A 的计算方式为

$$\boldsymbol{q}=\boldsymbol{x}'^{1}_{\mathrm{cls}}\boldsymbol{W}_{q},\quad \boldsymbol{k}=\boldsymbol{x}'^{1}\boldsymbol{W}_{k},\quad \boldsymbol{v}=\boldsymbol{x}'^{1}\boldsymbol{W}_{v}, \tag{2}$$

$$A=\mathrm{Softmax}(\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}/\sqrt{D/h}), \tag{3}$$

式(2)中,$\boldsymbol{W}_{q},\boldsymbol{W}_{k},\boldsymbol{W}_{v}\in\mathbf{R}^{C\times(C/h)}$ 是可学习的参数,D 表示嵌入的维度,h 是注意力头的数量.由于本模块只在查询中使用类标记,因此DA注意力操作中的注意结果 A 的计算和内存复杂度是线性的,使得整个操作过程更加高效.本文与注意力机制一样,使用多个注意力头,并且将其表示为多头双注意力模块(MDA).而为了扩展注意力机制,过滤掉无用的注意力结果,$\boldsymbol{q}$ 向量经过线性变换和 Sigmoid 函数激活之后得到指导向量 $\boldsymbol{q}'$,该向量对注意力结果 A 进行指导,过滤掉无用的注意力结果后得到新的值向量 $\boldsymbol{v}'$.

$$\boldsymbol{q}'=\sigma(\boldsymbol{W}'_{q}\boldsymbol{q}),\quad \boldsymbol{v}'=\boldsymbol{W}_{A}\boldsymbol{A}, \tag{4}$$

$$DA(\boldsymbol{x}'^{l})=\boldsymbol{q}'\boldsymbol{v}', \tag{5}$$

式(4)中,$\boldsymbol{W}'_{q}$ 和 $\boldsymbol{W}_{A}$ 是可学习的参数,σ 表示 Sigmoid 函数.而最终模型输出 $\boldsymbol{R}^{1}$ 的定义如下:

$$y^{1}_{\mathrm{cls}}=f^{1}(\boldsymbol{x}^{1}_{\mathrm{cls}})+\mathrm{MDA}(\mathrm{LN}([f^{1}(\boldsymbol{x}^{1}_{\mathrm{cls}})\,\|\,\boldsymbol{x}^{\mathrm{s}}_{\mathrm{patch}}])), \tag{6}$$

$$\boldsymbol{R}^{1}=[g^{1}(\boldsymbol{y}^{1}_{\mathrm{cls}})\,\|\,\boldsymbol{x}^{1}_{\mathrm{patch}}], \tag{7}$$

式(6)、(7)中,$f^{1}(\cdot)$ 和 $g^{1}(\cdot)$ 是维度对齐的投影函数和反投影函数.

DA 模块能够对来自 La 分支和 Sm 分支的信息进行充分融合,并且能够对注意力结果进行筛选只保留有用的特征信息,以进行后续的下游任务.如图 3 所示,本文将图 1 中的多尺度 Transformer 编码器

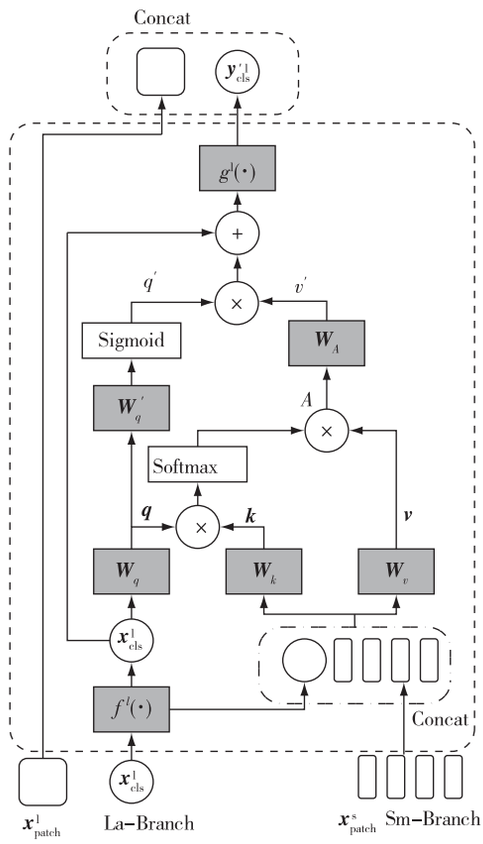

图 2　DA 模块

Fig. 2　Dual attention module

(黄色区域)中的交叉注意力模块替换成 DA 模块,从而得到双注意力多尺度 Transformer 编码器.

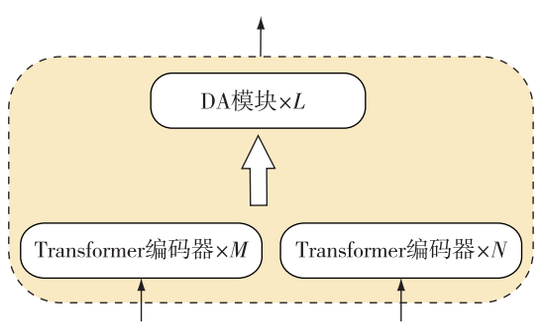

图 3　双注意力多尺度 Transformer 编码器

Fig. 3　Dual attention multi-scale Transformer encoder

2.2　双注意力 CrossViT

本文在保持原 CrossViT 其他结构没有改变的基础上,将多尺度 Transformer 编码器里面的交叉注意力结构(CrossAttention)直接替换为本文提出的 DA 模块,以进行微表情识别.双注意力模块(DA)对于注意力结果能够进行有效的筛选,只保留有用的注意力结果.考虑到 CrossViT 的交叉注意力模块会对

多尺度的特征信息进行交互,将会产生多个注意力结果.为了保留有用的注意力结果,本文将 CrossViT 中的交叉注意力模块替换为双注意力模块,从而提出了双注意力 CrossViT,该网络有效提升了微表情识别的精度.

3 基于双注意力 CrossViT 的微表情识别

本文将上述提出的双注意力 CrossViT 架构用于微表情识别任务.由于双注意力 CrossViT 机构能将输入光流图划分为不同的尺度,多尺度的特征信息交互能够得到更丰富的特征表示,并且能够对信息交互的过程无用的特征信息进行筛选,从而能够得到更具有代表性的特征信息,以提高最终微表情识别的精度.基于双注意力 CrossViT 的微表情识别架构如图 4 所示,该架构总体分为 3 个部分:

1)对微表情数据集中的微表情序列进行预处理,预处理包括微表情原始样本序列的人脸裁剪和人脸对齐,并通过峰值帧定位算法定位出微表情序列中的峰值帧;

2)对微表情样本的起始帧和峰值帧进行光流计算,将得到的水平、垂直光流矢量及光流应变进行融合得到光流特征图;

3)将光流特征图输入到双注意力 CrossViT 网络中进行特征提取,之后通过 Softmax 进行微表情分类.

4 实验结果与分析

本文所有实验均在一台安装了 Ubuntu 18.04.4 操作系统的服务器上进行,CPU 的型号为 Intel(R) Core(TM) i7-10700 CPU @ 2.90 GHz,内存为 16 GB,GPU 的型号为 NVIDIA 3090,显存大小为 24 GB,CUDA 版本为 10.0.实验数据设置如下:批处理大小为 256,最大轮次数为 800,学习率取值为 0.000 5,使用 Adam 优化器来优化模型.

4.1 数据集

本文使用 3 个独立数据集验证网络的分类性能,分别是 CASME Ⅱ[29]、SMIC[30]和 SAMM[31].3 个独立数据集分类种类不一致,为了消除其种类不一致所造成的误差,需要对 3 个数据集进行预处理.

首先针对 CASME Ⅱ数据集,其一共有 5 种情感类型,因此本文不使用其标签类型为‘others’的样本数据,将‘Repression’和‘Disgust’样本标签统一划分为标签‘Negative’.然后针对 SAMM 数据集,其一共有 8 种情感类型,本文将‘Fear’、‘Anger’、‘Disgust’、‘Sadness’以及‘Contempt’统一划分为标签‘Negative’,并且不使用标签为‘Other’的微表情样本.

为了验证数据集在不同性质数据集上的泛化能力,本文还使用了 3 个独立数据集抽样选择的融合数据集.本文使用 MEGG(The Second Facial Micro-Expression Grand Challenge)提出的融合数据集划分规则,将从 3 个数据集抽样出一定比例的数据集样本,其样本抽样的情况如表 1 所示.

表 1 融合数据集抽样

Table 1 Sampling of fusion databases

数据集	受试者	类型标签			
		消极	积极	惊讶	总计
CASME Ⅱ	24	88	32	25	145
SAMM	28	92	26	15	133
SMIC	16	70	51	43	164
Full	68	250	109	83	442

4.2 评估指标

本文实验采用留一受试交叉验证(Leave One

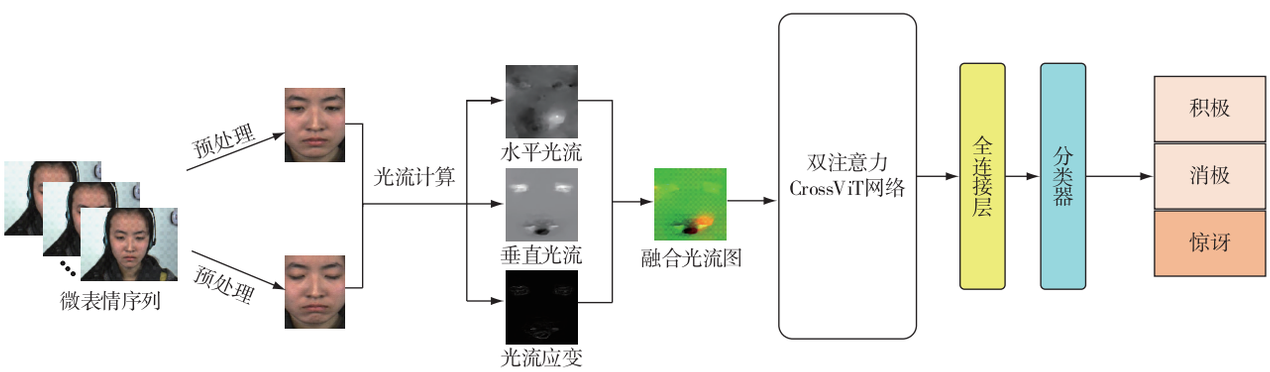

图 4 基于双注意力 CrossViT 的微表情识别架构

Fig. 4 Micro-expression recognition architecture based on dual attention CrossViT

Subject Out,LOSO)对融合数据集进行验证.LOSO 是微表情识别领域中最广泛使用的评估方法[12,17,31-33],该方法能够对数据集进行公平验证,防止小样本数据集出现过拟合的问题.每轮的 LOSO 将选择其中一名受试者的样本作为测试集,其余受试者的样本作为训练集,得到该受试者的 UF1(Unweighted F1-Score)和 UAR(Unweighted Average Recall).经过 k 轮的交叉验证后,LOSO 的最终实验结果的计算公式为 $a=\frac{1}{k}\sum_{j=1}^{k}a_j$,其中 k 为受试者数量,a_j 为第 j 轮的 UF1 和 UAR.

由于 3 个独立微表情数据集以及融合数据集都存在数据严重不平衡的问题,可以通过 UF1 和 UAR 这两个指标减少类不平衡的偏差,从而更好地衡量算法的性能.其中,UF1 通过计算每个类别 F1-Score 的平均值确定,其计算公式如下:

$$\mathrm{F1}_i=\frac{2\mathrm{TP}_i}{2\mathrm{TP}_i+\mathrm{FP}_i+\mathrm{FN}_i},\tag{8}$$

$$\mathrm{UF1}=\frac{1}{C}\sum_{i=1}^{C}\mathrm{F1}_i,\tag{9}$$

其中:TP_i,FP_i 和 FN_i 分别是微表情数据集中类别 i 的真阳性、假阳性和假阴性的数量;C 是当前微表情数据的类别数量.UAR 通过计算每个类别的平均准确率除以类数进行确定,其计算公式如下:

$$\mathrm{ACC}_i=\frac{\mathrm{TP}_i}{N_i},\tag{10}$$

$$\mathrm{UAR}=\frac{1}{C}\sum_{i=1}^{C}\mathrm{ACC}_i,\tag{11}$$

其中,N_i 为类别标签 i 的样本总数.

4.3 对比实验

为了验证双注意力 CrossViT 网络在微表情识别上的有效性,本文在 3 个独立微表情数据集及融合数据集上进行了广泛的实验,并且和目前微表情领域的其他主流方法进行对比.本文的基准方法为 LBP-TOP,其他方法(ATNET、CapsuleNet、SA-AT、STSTNet)为主流的深度学习方法.从表 2 可以看出,在融合数据集上,双注意力 CrossViT 网络的 UF1 指标达到了 0.727 5,UAR 指标达到了 0.727 2,比基准方法分别提高了 0.139 5 和 0.148 7.在 CASME Ⅱ、SMIC、SAMM 数据集上,相比于基准方法,本文网络的性能均有明显的提升.对比实验中主流的深度学习方法,改进的 CrossViT 网络也有明显的优势.在融合数据集和 SMIC、CASME Ⅱ数据集上,双注意力 CrossViT 网络相较于其他方法达到了最好的性能,在 SAMM 数据集上,也取得了较好的性能表现,仅次于 CapsuleNet.由于视觉 Transformer 需要大量的数据样本才能取得最好的性能,而 SMIC 和 SAMM 数据集样本数量较少,所以相较于其他数据集性能表现不太优秀.

为了更好地对本文所提方法的有效性进行分析,在融合数据集和 3 个独立数据集上构造混淆矩阵,如图 5 所示.可以看出所提方法在 CASME Ⅱ数据集的分类性能相较于其他 3 个数据集,其分类性能最好.由于 SMIC 数据集未标记实际峰值帧位置,因此峰值帧定位算法定位出的峰值帧位置与其实际峰值帧位置存在误差,并且 SMIC 数据集也存在分辨率低、帧数低的问题,因此其分类表现在 4 个数据集中是最差的.另外对于 SAMM 数据集,其数据集类别不平衡,消极类别在整个数据集中的样本数量最大,导致所提方法对于‘消极’标签的分类性能相较于其他两个分类标签表现更好.在融合数据集中,‘消极’标签的分类结果也是表现最好的.

为了验证实验参数设置对于算法稳定性的影响,本文在 3 个独立微表情数据集上通过设置不同的迭代次数进行实验,如图 6 所示,迭代次数设置在 200 到 800 之间时,算法的 UF1 和 UAR 指标曲线较为平稳,而且实验结果中两个指标的值高于表 2 中

表 2 对比实验结果

Table 2 Comparative experiment results

方法	Full		CASME Ⅱ		SMIC		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP[31]	0.588 0	0.578 5	0.702 6	0.742 9	0.200 0	0.528 0	0.395 4	0.410 2
ATNET[32]	0.631 0	0.613 0	0.798 0	0.755 0	0.553 0	0.543 0	0.496 0	0.482 0
SA-AT[33]	0.593 6	0.595 8	0.760 7	0.755 2	0.551 2	0.546 3	0.447 6	0.486 8
CapsuleNet[12]	0.652 0	0.650 6	0.706 8	0.701 8	0.582 0	0.587 7	**0.620 9**	**0.598 9**
STSTNet[17]	0.720 9	0.725 0	0.836 8	0.837 9	0.542 1	0.542 9	0.512 2	0.496 3
本文方法	**0.727 5**	**0.727 2**	**0.863 9**	**0.866 4**	**0.699 2**	**0.698 2**	0.583 9	0.586 9

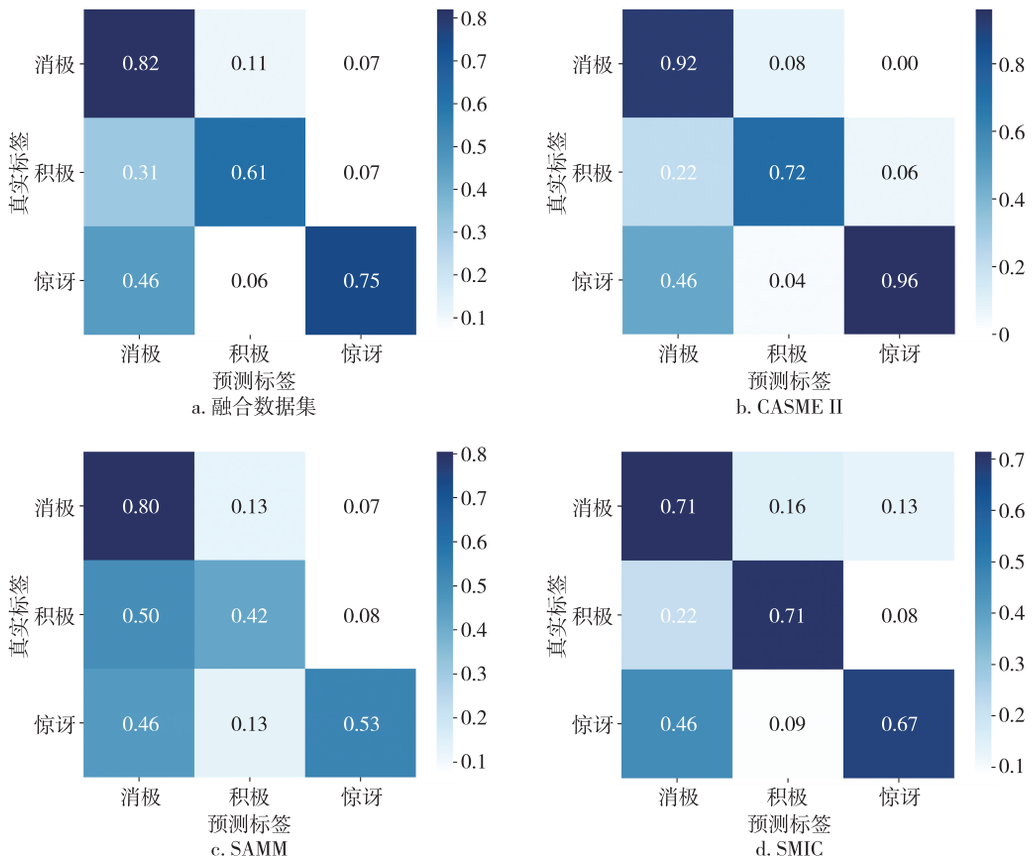

图 5　混淆矩阵

Fig. 5　Confusion matrices

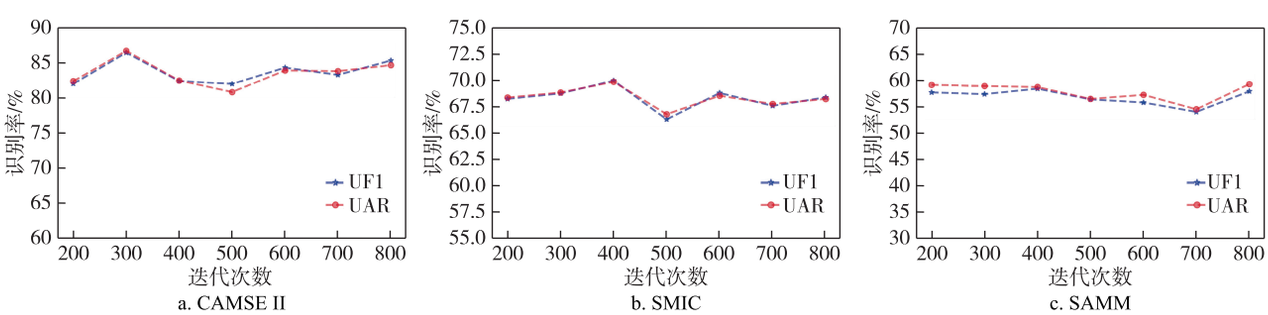

图 6　迭代次数对网络识别率的影响

Fig. 6　Influence of iteration times on network recognition rate

的主流算法的指标.

由于 CrossViT 的 La 分支主要用于提取特征信息,Sm 分支只是提供补充信息,因此 2 个分支中的不同分块数量会影响算法识别性能. 为了验证 CrossViT 2 个分支的不同分块数量对于算法识别性能的影响,本文在 3 个独立数据集上进行实验,实验结果如表 3 所示. 从表 3 可知,不同的分块情形,CAMSE Ⅱ数据集的 UF1 和 UAR 指标分布在 0.8 左右,SMIC 数据集的 UF1 和 UAR 指标分布在 0.6 左右,SAMM 数据集的 UF1 和 UAR 指标分布在 0.5 左右,3 个独立数据集中,当 La 分支分块数量为 1 且 Sm 分支分支分块数量为 4 时,其实验结果最好.

4.4　消融实验

为了验证所提出的 DA 模块在 CrossViT 网络中的有效性,本文进行了针对 DA 模块的消融实验. 实验细节是使用 CrossViT 原有的交叉注意力模块和本文所提出的 DA 模块进行比较,分析 2 个模块在融合数据集上的识别精度. 实验结果如表 4 所示,本文

表 3　分块数量实验结果

Table 3　Experimental results of number of blocks

La 分支分块数量	Sm 分支分块数量	CASME Ⅱ		SMIC		SAMM	
		UF1	UAR	UF1	UAR	UF1	UAR
1	1	0.807 6	0.801 8	0.668 4	0.666 1	0.550 4	0.568 8
1	4	**0.863 9**	**0.866 4**	**0.699 2**	**0.698 2**	**0.583 9**	**0.586 9**
1	16	0.829 1	0.818 9	0.692 1	0.688 1	0.557 5	0.561 1
4	4	0.790 5	0.797 2	0.623 8	0.616 7	0.518 2	0.502 9
4	16	0.783 1	0.789 7	0.632 8	0.633 3	0.536 1	0.531 6

提出的 DA 模块能够保留最有用的注意力结果,最终获得更好的特征表示.在融合数据集上,DA 模块的 UF1 的指标为 0.727 5,UAR 的指标为 0.727 2,性能较 CrossViT 原有的交叉注意力模块均有提高.

表 4　消融实验结果

Table 4　Ablation experiment results

注意力模块	UF1	UAR
双注意力模块	0.727 5	0.727 2
交叉注意力模块	0.714 5	0.706 5

5　结束语

本文基于 CrossViT 为主干网络,对网络中的交叉注意力模块进行改进,提出了双注意力模块,实现了不同尺度的光流图像的特征融合并保留了有用的注意力结果,有效地提升了微表情识别的准确率,并且将起始帧到峰值帧的水平和垂直光流矢量及光流应变融合为光流特征图,通过多尺度的特征提取进行微表情分类.本文在 3 个独立数据集上和融合数据集上使用 LOSO 交叉验证法验证模型,实验结果表明,本文的方法在识别性能上相较于目前的主流深度学习方法都有了较为明显的提升.

未来的工作重点可以从以下方面进行提升:当前的微表情数据集规模小,应该使用 GAN、迁移学习等网络进一步提升微表情数据集的规模;当前的数据集样本数量不平衡,应该进一步提升在样本类别不平衡情况下的网络识别精度.

参考文献

References

[1] Yan W J,Wu Q,Liang J,et al.How fast are the leaked facial expressions:the duration of micro-expressions[J].Journal of Nonverbal Behavior,2013,37(4):217-230

[2] Porter S,Brinke T L.Reading between the lies:identifying concealed and falsified emotions in universal facial expressions[J].Psychological Science,2008,19(5):508-514

[3] Ahonen T,Hadid A,Pietikäinen M.Face description with local binary patterns:application to face recognition[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2006,28(12):2037-2041

[4] Zhao G Y,Pietikäinen M.Dynamic texture recognition using local binary patterns with an application to facial expressions[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2007,29(6):915-928

[5] Wang Y D,See J,Phan R C W,et al.LBP with six intersection points:reducing redundant information in LBP-TOP for micro-expression recognition[M]//Computer Vision—ACCV 2014.Cham:Springer International Publishing,2015:525-537

[6] Huang X H,Wang S J,Liu X,et al.Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition[J].IEEE Transactions on Affective Computing,2019,10(1):32-47

[7] 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

[8] Xu F,Zhang J P,Wang J Z.Microexpression identification and categorization using a facial dynamics map[J].IEEE Transactions on Affective Computing,2017,8(2):254-267

[9] 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,2021,12(1):254-261

[10] 马浩原,安高云,阮秋琦.平均光流方向直方图描述的微表情识别[J].信号处理,2018,34(3):279-288
MA Haoyuan,AN Gaoyun,RUAN Qiuqi.Mean histogram of oriented optical flow feature for micro-expression recognition[J].Journal of Signal Processing,2018,34(3):279-288

[11] Liong S T,See J,Phan R C W,et al.Less is more:micro-expression recognition from video using apex frame[J].Signal Processing:Image Communication,2018,62:82-92

[12] Quang N V,Chun J,Tokuyama T.CapsuleNet for micro-expression recognition[C]//2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019).May 14-18,2019,Lille,France.IEEE,2019:1-7

[13] Sabour S,Frosst N,Hinton G E.Dynamic routing between

capsules[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York:ACM,2017:3859-3869

[14] Lai Z Y,Chen R H,Jia J L,et al.Real-time micro-expression recognition based on ResNet and atrous convolutions [J]. Journal of Ambient Intelligence and Humanized Computing,2020,8:1-12

[15] Wang C Y,Peng M,Bi T,et al.Micro-attention for micro-expression recognition[J]. Neurocomputing, 2020, 410: 354-362

[16] Liong S T,See J,Phan C W,et al.Spontaneous subtle expression detection and recognition based on facial strain [J].Signal Processing:Image Communication,2016,47: 170-182

[17] Liong S T, Gan Y S, See J, et al. Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition [C]//2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). May 14-18, 2019, Lille, France. IEEE, 2019:1-5

[18] Verma M, Vipparthi S K, Singh G, et al. LEARNet: dynamic imaging network for micro expression recognition [J].IEEE Transactions on Image Processing,2019,29: 1618-1627

[19] Khor H Q,See J,Phan R C W,et al.Enriched long-term recurrent convolutional network for facial micro-expression recognition [C]//2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018).May 15-19,2018,Xi'an,China. IEEE,2018:667-674

[20] Zhao Y,Xu J C.Compound micro-expression recognition system[C]//2020 International Conference on Intelligent Transportation,Big Data & Smart City (ICITBS).January 11-12,2020,Vientiane,Laos.IEEE,2020:728-733

[21] Khor H Q,See J,Liong S T,et al.Dual-stream shallow networks for facial micro-expression recognition[C]// 2019 IEEE International Conference on Image Processing (ICIP). September 22-25, 2019, Taipei, China. IEEE, 2019:36-40

[22] Zhi R C,Liu M Y,Xu H R,et al.Facial micro-expression recognition using enhanced temporal feature-wise model [M]//Communications in Computer and Information Science.Singapore:Springer Singapore,2019:301-311

[23] Ma F,Sun B,Li S.Robust facial expression recognition with convolutional visual transformers[J].arXiv e-print, 2021,arXiv:2103.16854

[24] Zhang L F,Hong X P,Arandjelović O,et al.Short and long range relation based spatio-temporal transformer for micro-expression recognition[J].IEEE Transactions on Affective Computing,2022,13(4):1973-1985

[25] 刘忠洋,周杰,陆加新,等.基于注意力机制的多尺度特征融合图像去雨方法[J/OL].南京信息工程大学学报(自然科学版):1-11[2022-11-11].http://kns.cnki.net/kcms/detail/32.1801.N.20221013.1428.004.html
LIU Zhongyang, ZHOU Jie, LU Jiaxin, et al. Multi-scale feature fusion image rain removal algorithm based on attention mechanism [J]. Journal of Nanjing University of Information Science & Technology (Natural Science Edition):1-11[2022-11-11].http://kns.cnki.net/kcms/detail/32.1801.N.20221013.1428.004.html

[26] Chen C F R,Fan Q F,Panda R.CrossViT:cross-attention multi-scale vision transformer for image classification [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). October 10-17, 2021, Montreal,QC,Canada.IEEE,2022:347-356

[27] Huang L,Wang W M,Chen J,et al.Attention on attention for image captioning[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV).October 27-November 2, 2019, Seoul, Korea (South). IEEE, 2020: 4633-4642

[28] 杨春霞,韩煜,陈启岗,等.基于BERT与注意力机制的方面级隐式情感分析模型[J/OL].南京信息工程大学学报(自然科学版):1-12[2022-11-11].http://kns.cnki.net/kcms/detail/32.1801.N.20221109.1915.002.html
YANG Chunxia, HAN Yu, CHEN Qigang, et al. Aspect-based implicit sentiment analysis model based on BERT and attention mechanism [J]. Journal of Nanjing University of Information Science & Technology (Natural Science Edition):1-12[2022-11-11].http://kns.cnki.net/kcms/detail/32.1801.N.20221109.1915.002.html

[29] Yan W J, Li X B, Wang S J, et al. CASME Ⅱ: an improved spontaneous micro-expression database and the baseline evaluation[J].PLoS One,2014,9(1):e86041

[30] Li X B,Pfister T,Huang X H,et al.A spontaneous micro-expression database:inducement,collection and baseline [C]//2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG).April 22-26,2013,Shanghai,China.IEEE,2013: 1-6

[31] Davison A K,Lansley C,Costen N,et al.SAMM:a spontaneous micro-facial movement dataset[J].IEEE Transactions on Affective Computing,2016,9(1):116-129

[32] Peng M, Wang C Y, 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).September 3-6,2019,Cambridge,UK.IEEE,2019:1-6

[33] Zhou L,Mao Q R,Xue L Y.Cross-database micro-expression recognition:a style aggregated and attention transfer approach[C]//2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW).July 8-12, 2019,Shanghai,China.IEEE,2019:102-107

Micro-expression recognition based on dual attention CrossViT

RAN Ruisheng[1] SHI Kai[1] JIANG Xiaopeng[1] WANG Ning[1]
1 College of Computer and Information Science,Chongqing Normal University,Chongqing 401331,China

Abstract Micro-expression is the facial expression that people reveal involuntarily when they try to hide their true emotions,which is a hot spot in research of affective computing in recent years.Micro-expression is a subtle facial movement thus is difficult to recognize.Considering its excellent performance in image classification and ability to capture subtle feature information,the cross-attention multiscale ViT (CrossViT) is used as the backbone network to improve the cross-attention mechanism in the network,and the Dual Attention (DA) module is proposed to extend traditional cross-attention mechanism to determine the correlation between attention results,thus improve the micro-expression recognition accuracy.The proposed network learns from three optical flow features (optical strain,horizontal and vertical optical flow fields),which are calculated from the starting frame and peak frame of each micro-expression sequence,and classifies the micro-expression by Softmax.Experiments on the micro-expression fusion dataset show that the proposed network reaches 0.727 5 and 0.727 2 in UF1 and UAR,respectively,which is more accurate than the mainstream micro-expression recognition algorithms,verifying the effectiveness of the dual attention CrossViT based network.

Key words micro-expression recognition;CrossViT;cross attention mechanism;optical flow feature