小型微型计算机系统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2024-0303
2025年8月 第8期
Vol.46 No.8 2025

结合多区域特征和特征融合的微表情识别

曹春萍,张 迪

(上海理工大学 光电信息与计算机工程学院,上海 200093)
E-mail:3244279275@qq.com

摘 要: 微表情运动微弱、短暂和局部化的特点,使得难以从微表情视频序列中相关的局部区域中提取有效特征,进而导致准确识别微表情变得十分困难.针对上述问题,基于残差网络和长短期记忆网络,提出一种结合多区域特征提取模块(Multi-region Feature Extraction Module,MFEM)和多层特征融合模块(Multi-level Feature Fusion Module,MFFM)的微表情识别方法.首先,对微表情视频序列采用欧拉视频放大算法实现运动增强得到灰度序列,并结合TV-L1光流法的光流序列作为输入.有效特征提取阶段中,利用MFEM模块提取多个相关的局部区域中的显著特征,增强网络提取有效特征的能力;通过MFFM模块减少信息丢失,产生更综合的特征,提高模型学习微表情特征的能力;然后进行时序建模并分类.在casme2和samm数据集上进行实验,准确率分别达到84.959%、74.265%,UF1分别为0.855和0.604,优于现有方法.

关 键 词: 微表情识别;残差网络;多区域特征提取;多层特征融合
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2025)08-1986-07

Micro-expression Recognition Combining Multi-region Feature and Feature Fusion

CAO Chunping,ZHANG Di
(School of Optical-Electrical & Computer Engineering,University of Shanghai for Science & Technology,Shanghai 200093,China)

Abstract: The weak,short and localized movements of micro-expression make it difficult to extract efficient features from relevant and local regions in the video sequence of micro-expression,thus making it difficult to accurately identify micro-expressions. For the issue,based on residual networks and long short-term memory networks,a micro-expression recognition method incorporating Multi-region Feature Extraction Module(MFEM)and Multi-level Feature Fusion Module (MFFM)is proposed. Firstly,the micro-expression video sequences are motion-enhanced using Euler video magnification to obtain grayscale sequences,which are then concatenated with optical sequences using TV-L1 method as input. In efficient feature extraction stage,the MFEM module extracts salient features from multiple relevant and local regions,improves the network's capacity to capture efficient features. The MFFM module reduces the miss of the information,produces more comprehensive features,improves the model's capability to learn micro-expression features. Then,temporal modeling and classification are performed. Experiments are performed on casme2 and samm datasets,the proposed method performers better than the existing methods,and the accuracy reaches 84.959% and 74.265%,respectively,and the UF1 is 0.855 and 0.604.
Keywords: micro-expression recognition;residual networks;multi-region feature extraction;multi-level feature fusion

0 引 言

微表情是瞬间发生且难以被发现察觉的微小脸部肌肉变化,但能够揭示人类内心的最真实的情感状态,为此在司法[1]、医疗健康[2]、公共安全[3]等领域中得到广泛的应用,因此对微表情的识别的研究也逐渐备受关注.但微表情运动发生时间不长,一般不到0.2秒[4];肌肉运动强度微弱,仅凭肉眼难以察觉;且往往只发生在面部的局部区域[5].由于这些特点,使得难以从微表情视频序列中提取有用的可判别特征,进而导致发现并准确识别微表情情感类别变得十分困难.因此对于怎么从微表情视频序列中相关的面部局部区域中提取出可用于区分微表情情感类别的有效特征的研究具有必要性.

早期的传统机器学习方法主要依靠提取手工特征进行分类,往往只能提取表面信息,难以提取深层次信息,如LBP-TOP[6],Bi-WOOF[7],识别效果不佳.目前则广泛使用深度学习方法,对于微表情视频序列中的空间信息一般使用卷积神经网络(Convolutional neural network,CNN)提取特征,而对于序列间的时间顺序关系则考虑采用长短期记忆网络(Long Short-Term Memory Networks,LSTM)处理,这类方法的识别准确率普遍优于传统方法[8].同时,考虑到微表情只在一些小的局部的面部区域中表达,Wang[9]等人将微注意机制与残差网络(Residual Networks,ResNet)相结合,提出了Micro-attention方法,该方法加强了对微表情最显著的局部区域的关注并提取有效特征进行分类,从而获得更高的识别效果.这类方法可以帮助网络关注能够表达微表情的最显著的局部区域,并利用网络最后一层中提取的有效显著特征进行分类,缓解了微表情的局部性问题,并获得不错的识别准确率.然而,

收稿日期:2024-06-06 收修改稿日期:2024-07-05 基金项目:国家重点研发计划项目(2021YFF0600605)资助. 作者简介:**曹春萍**,女,1968年生,硕士,副教授,CCF会员,研究方向为图像处理、机器视觉、数据挖掘等;**张 迪**,女,2001年生,硕士研究生,研究方向为图像分类、机器视觉、表情识别.

上述这类方法通常只能关注微表情中最显著的局部区域的特征,而忽视其他次要区域中的特征,存在无法实现对微表情视频序列中多个相关的面部局部区域提取有效显著特征的问题.此外,这类深度学习方法直接利用最后的深层特征进行分类识别,这样会忽略浅层中提取的微表情的细微运动和局部信息的特征,导致信息丢失,进而影响微表情识别的准确率.

因此,针对上述问题,本文在 ResNet34 和 LSTM 级联的网络模型的基础上,提出一种结合多区域特征和特征融合的微表情识别网络模型.具体来说,在主干网络中引入一个多区域特征提取模块 MFEM,通过注意力机制使模型不仅能够提取微表情运动的最显著局部区域的显著特征,还能够提取其他次要区域中容易被忽视但同样重要的次级显著特征,实现对多个相关的面部局部区域的有效显著特征提取;此外设计一个多层特征融合模块 MFFM,充分利用深度学习模型多层结构捕获的不同语义的特征,将主干网络中不同阶段学习的深层和浅层特征进行交互并融合,减少信息丢失,产生更综合、更具表现力的微表情特征表示,最后实现微表情识别准确率的提升.

本文的主要贡献有:

1)提出多区域特征提取模块 MFEM.关注显著的局部区域特征和抑制区域的次级特征,以此充分提取微表情视频序列中相关面部中的多个局部区域的可用于区分微表情的有效特征.

2)提出多层特征融合模块 MFFM.利用多层特征,减少细微信息丢失的风险,丰富特征间的信息多样性,增强了网络对微表情特征的表示能力.

3)在 ResNet34 + LSTM 级联网络模型的基础上,引入前面两个模块,提出本文的微表情识别网络模型.在常用数据库 casme2[10] 和 samm[11] 进行实验,实现了优于现有方法的效果.

1　相关工作

早期方法是提取手工特征,然后用机器学习的方法分类,提取的信息较浅,且往往依赖专业人员的领域知识,主要分为两类.基于局部二值模式的这类方法基于像素值,因此损失的信息量较少,可以提取纹理特征,但这类方法计算量大,且准确率不高,应用于现实场景效果不好;基于光流[12]的方法能够检测出序列中前后帧间的微小变化,获取运动信息,这类方法基于几何特征,特征维度不高,但需满足光流假设,易受外界环境影响,在复杂场景中无法避免[13].

Patel[14]等人的使用深度学习方法的研究虽然未能获得很好的微表情识别效果,但为后续研究提供了思路.Kim[15]等人首次提出分别用 CNN 和 LSTM 提取空间和时间特征进行识别的微表情识别网络模型,该方法通过 CNN 提高学习到的微表情空间特征的类别可分性,通过 LSTM 编码不同状态下的时间特征并生成潜在的微表情表示,最后取得了良好效果.Khor[16]等人提出了通过通道拼接增加空间信息、通过深度特征拼接增加时间信息的增强长期递归卷积网络 ELRCN,结合了 VGG16 和 LSTM,但该方法十分依赖数据处理.Reddy[17]等人使用 3D CNN 网络提取空间特征并兼顾运动信息进行微表情识别,但该方法参数多,计算量大,复杂度高.Gan[18]等人提出了将 CNN 与峰值帧相结合的 OFF-ApexNet 模型,尽管该方法仅利用微表情序列的峰值帧作为输入,可以显著减少输入数据的冗余,但也会导致微表情的时序信息被忽略,从而影响准确率.Liong[19]等人提出结合了时空信息、可以提取更具差异性的高级特征的浅层三流三维卷积神经网络 STST-Net,取得不错的效果.Zhang[20]等人提出了一种时空转换器的架构,以光流矩阵作为网络输入,用 Transformer 处理空间模式信息,用 LSTM 分析时间维度,得到结合时空信息的特征用于分类识别,取得了不错的性能.上述深度学习方法利用深层或浅层网络提取特征,并利用提取到的最后一层的特征进行分类识别,相比传统方法特征提取能力更好,识别结果也更好,都在一定程度上提高了微表情识别的准确率.但上述深度学习方法没有考虑微表情是通过面部局部区域的瞬间微弱的运动来表达的,存在精度不足、计算复杂冗余、难以从相关的局部区域中提取有效特征等问题.

微表情识别中,注意力机制可以引导模型关注关键的显著区域,抑制不相关的面部区域和干扰噪声.因此,多种注意力机制被提出以缓解微表情局部性问题.比如,Li[21]等人提出一种利用局部和全局信息学习有区别的微表情表示、并抑制不相关区域的微表情识别架构.Yao[22]等人提出一个结合通道注意力机制学习每个不同面部区域特征的通道权值以提取有效特征的三流网络模型.Yang[23]等人提出一个使用 3 种注意力机制对 3 种特征进行处理以辅助模型分类的 MERTA 网络.Su[24]等人提出一种使用组件感知注意力模块凸显相关微表情区域的 KFC-MER 方法.Chen[25]等人提出 AM3F-FlowNet 网络,该网络利用空间注意提取更稳定的特征,利用通道注意选择光流特征,获得了目前较优的微表情识别准确率.Zhao[26]等人提出一种通过卷积注意模块聚焦重要信息并抑制不相关信息的注意力引导的浅层三流卷积神经网络 AT-SCNN,提高了准确率.

上述方法均使用了注意力机制,加强了对微表情运动局部区域的关注,抑制对没有可以用于帮助微表情识别的有效特征的面部区域的关注,并取得了不错的性能,表明采用注意力机制的深度学习方法能够有效提取微表情运动的局部区域中的特征并促进微表情分类.然而,在现有研究中,这类使用注意力机制的方法常常只关注微表情中最显著的局部区域并提取特征用于识别微表情,而对于一些发生在相近的局部面部区域且有重叠区域的运动单元(Action Unit,AU),在区分时容易发生混淆,从而影响识别结果;并且被忽视的其他次要区域中也可能覆盖一些 AU,这些区域中包含一些可用于识别微表情的微小特征,忽视这些特征对准确识别微表情存在一定影响.因此,目前这类方法虽然可以加强对微表情局部区域的关注并提取有效显著特征进行识别,但还是无法实现从相关的多个局部区域中充分提取有效显著特征用于准确识别微表情.另外,深度学习模型的多层结构可以捕获视频中每一帧微表情图像的不同级别的特征,从低级的纹理和边缘到更高级的语义信息,但现有的一些深度学习方法仅利用提取的微表情视频序列中图像的深层特征进行分类,未考虑同时利用深层和浅层特征,可能会忽略浅层中提取的微表情的细微运动和局部信息的特征,造成细微特征的丢失,从而影响识别

效果. 因此,本文基于上述考虑,对现有模型无法充分提取微表情视频序列中相关面部的多个局部区域的有效显著特征以及没有充分利用网络中提取的多层特征表示微表情的问题展开研究,以期望取得更好的结果.

2 方 法

2.1 网络模型

由于残差网络可以更好的学习提取深层特征,且计算复杂度更低,因此本文使用残差网络提取微表情视频序列中的空间信息. 对于微表情视频序列中的时间顺序关系,则使用LSTM 建模得到微表情特征表示. 本文的微表情识别模型以ResNet34 和 LSTM 级联的网络模型作为基础框架,具体如图1 所示.

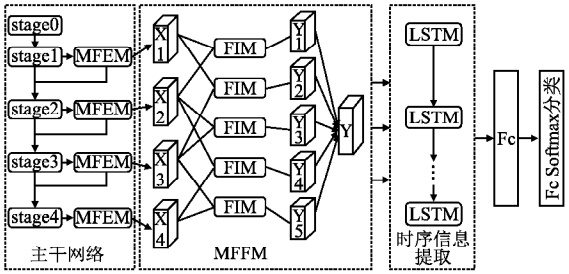


图 1 网络结构

Fig. 1 Network structure

该模型的主干网络 ResNet34 包含 5 个主要阶段,本文在后面 4 个阶段中引入多区域特征提取模块 MFEM,学习每个阶段中显著区域和次级显著区域中的特征,图中 X_1、X_2、X_3、X_4 即为每个阶段学习到的多区域的显著特征. 对于第 1 阶段的特征 X_1,一方面将其与该阶段的原始特征级联增加特征复用性作为下一阶段的输入,用于学习更高级的特征;另一方面作为多层特征融合模块 MFFM 的输入. 中间两个阶段采用相似的处理,最后一个阶段的特征则直接输入到下一模块. MFFM 模块通过特征交互模块(Feature Interaction Module, FIM)对来自不阶段的多区域显著特征进行处理,扩展了特征多样性. 对于不同阶段的特征尺寸不匹配的问题,为保证拼接维度,需要将上一阶段的特征图依靠上采样操作,实现尺寸的统一. 接着,进行交互特征的融合,充分利用了各层的特征,减少了细微特征的丢失,产生了更综合、更具有表现力的特征表示. 之后将输出的空间特征构建特征集合,并送入 LSTM 模块对时序数据进行处理. 随后输出的结果通过全连接层(Fully-Connection,FC)进行映射处理,最后利用 softmax 函数得到识别的预测结果.

2.1.1 ResNet34 残差模块

图 2(a)是残差块结构,相比与传统网络,在第 2 层激活函数前增加了一个跳跃连接,激活函数的输入由 $H(x) = F(x)$ 变为 $H(x) = F(x) + x$. 在深度神经网络中,使用残差结构可以减少一些不必要的反向传播更新,加快网络收敛.

本文的主干网络 ResNet34 由 16 个类似于图 2(b)的残差块结构组成,其中 stage1-4 各个阶段分别包含 3、4、6、3 个该残差结构,用于提取特征.

2.1.2 多区域特征提取模块 MFEM

微表情中的一些 AU 的发生区域相近且有重合,使用注意力机制可以关注到这些区域,但由于它们相似且不易区分,

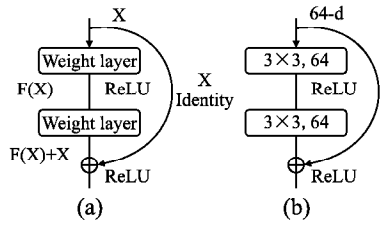


图 2 残差块结构

Fig. 2 Residual block structure

容易发生混淆导致误判,进而影响识别结果. 因此需要学习更多潜在区域的信息,充分提取目标中的相关面部的局部区域中的有效显著特征,更好的区分识别微表情. 参考 CBAM[27] (Convolutional Block Attention Module)的设计思路,本文的多区域特征提取模块采用注意力机制实现关注多个微表情发生的局部区域. 该模块由两部分内容组成,并采用并行架构来防止在串行架构中可能会发生的相互干扰,如图 3(a)所示.

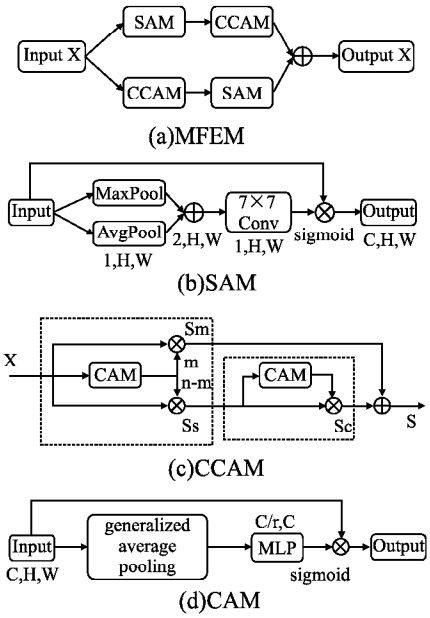


图 3 多区域特征提取模块模块 MFEM

Fig. 3 Multi-region feature extraction module

对于第 1 路分支,其中空间注意(Spatial Attention Module,SAM)如图 3(b). 首先进行池化处理,输出为:

$$f_1 = \varphi_1(X) \in R^{1\times H\times W} \tag{1}$$

$$f_2 = \varphi_2(X) \in R^{1\times H\times W} \tag{2}$$

式中 $X \in R^{C\times H\times W}$ 表示输入,φ_1 和 φ_2 分别表示全局最大和平均池化,f_1,f_2 表示经过池化处理后的结果. 然后通道拼接并使用卷积和激活函数处理. 计算公式为:

$$g_s = \sigma(\tau(f_1 \oplus f_2)) \in R^{1\times H\times W} \tag{3}$$

式中 g_s 表示得到的空间注意权重,$\oplus$表示拼接,τ 表示 7×7 的卷积变换函数,σ 表示 sigmoid 激活函数.

之后计算空间注意特征,输入到下一模块,计算公式为:

$$X_s = X \odot g_s \in R^{C\times H\times W} \tag{4}$$

式中 X_s 表示 SAM 的输出,⊙表示矩阵点乘积运算.

互补通道注意机制(Complementary Channel attention module,CCAM)子模块如图 3(c)所示,由主要区域特征提取和次要区域特征提取两部分组成实现,其中通道注意机制(Channel attention module,CAM)如图 3(d).首先进行广义平均池化[28]处理,计算公式为:

$$\varphi_3 = \left(\frac{1}{X_s(k)}\sum_{x\in X_s(k)} x^{p(k)}\right)^{\frac{1}{p(k)}} \tag{5}$$

式中 k 表示通道,当 $p(k)=1$ 时,等价于 φ_2,$p(k)=\infty$ 时,等价于 φ_1.

输出为:

$$f_3 = \varphi_3(X_s) \in R^{C\times 1\times 1} \tag{6}$$

式中 f_3 表示经过广义平均池化处理后的结果.然后经过 MLP 处理,每层神经元个数分别为 C/r、C,r 代表缩放率;再经过激活函数,计算公式为:

$$g_c^1 = \sigma(W_1(W_0 f_3)) \in R^{C\times 1\times 1} \tag{7}$$

式中 g_c^1 表示得到的通道注意权重,W_1,W_0 是 MLP 中的超参数.

然后得到经过 CCAM 模块处理的主要区域的显著特征 X_c^1,并输入到下一模块提取次级区域的显著特征,计算公式为:

$$X_c^1 = X_s \odot g_c^1 \in R^{C\times H\times W} \tag{8}$$

上述中得到了主要显著通道注意权重 g_c^1,同时也可以计算得到抑制部分的通道权重 g_c^2,计算公式为:

$$g_c^2 = g - g_c^1 \in R^{C\times 1\times 1} \tag{9}$$

式中 $g \in R^{C\times 1\times 1}$ 元素全为 1,g_c^2 表示抑制部分的通道权重.

而抑制的特征 X_c^2 可以表示为:

$$X_c^2 = X_s \odot g_c^2 \in R^{C\times H\times W} \tag{10}$$

X_c^1 和 X_c^2 是互补特征,X_c^2 中也包含了有效特征,可以从 X_c^2 中再次提取显著特征.与主要显著区域特征提取类似,将 X_c^2 作为输入,再次经过 CAM 子模块,计算公式为:

$$X_c^{2s} = X_c^2 \odot \sigma(W_3(W_2(\varphi_3(X_c^2)))) \in R^{C\times H\times W} \tag{11}$$

式中 X_c^{2s} 表示抑制部分的显著特征.

经过上述步骤,第 1 路分支经过 SAM 和 CCAM 模块提取目标特征,包括 X_c^1 和 X_c^{2s}.同理,第 2 分支先经过 CCAM,再经过 SAM 模块提取目标特征.最后通过加和得到多区域特征表示.

2.1.3　多层特征融合模块 MFFM

对于微表情识别,在深度学习模型中,浅层信息中包含细微的运动信息和局部特征等,深层信息包含抽象的高级的全局特征等,为了减少细微特征丢失,丰富特征多样性,本文通过多层特征融合模块 MFFM,有效融合了各层次特征,可以更好地表示微表情.MFFM 模块首先利用通过 FIM 对来自不同层次的特征交互学习,旨在增强特征的表征能力;并将各层次的特征进行有效融合,以产生更综合、更有表现力的特征表示,充分了利用各层特征,减少信息丢失.FIM 模块如图 4 所示.

$X_1 \in R^{C_1\times H_1\times W_1}$,$X_2 \in R^{C_2\times H_2\times W_2}$ 表示不同阶段提取的深层和浅层特征.由于它们的尺寸不匹配,无法直接进行交互,因此需先对 X_2 进行上采样以保证拼接维度,然后再交互,计算公式为:

$$Y = \eta(\tau(X_1 \oplus \mu(\tau(X_2)))) \in R^{(C_1+C_2)\times H_1\times W_1} \tag{12}$$

式中 Y 表示 X_1 和 X_2 的交互特征,η 表示批量归一化,μ 表示上采样操作.

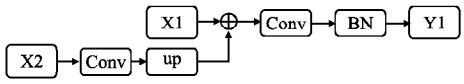


图 4　特征交互模块 FIM

Fig. 4　Feature interaction module

根据式(12)可以得到不同阶段之间的交互特征 Y_1,Y_2,Y_3,Y_4,Y_5,最后经过全局平均池化并拼接成特征 F.计算公式为:

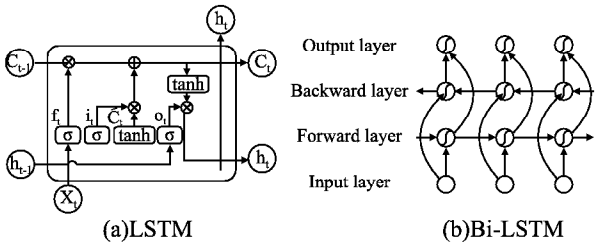
$$F = \varphi_2(Y_1) \oplus \varphi_2(Y_2) \oplus \varphi_2(Y_3) \oplus \varphi_2(Y_4) \oplus \varphi_2(Y_5) \tag{13}$$

为了避免损失,使得时域信息处理更准确,提取的特征 F 无需使用全连接层降维,直接转换成一个独立的向量,并输入到 LSTM 模块.

2.1.4　LSTM 模块

图 5(a)是 LSTM 的神经元结构,其中细胞状态单元 C_t(cell)是关键部分,该网络通过对 C_t 状态单元进行少量线性操作,可以实现了长时期的记忆保留.在 LSTM 网络中,遗忘门 f_t 决定信息丢弃、输入门 i_t 决定新信息的加入、输出门 o_t 则决定 C_t 状态单元的输出.

微表情视频序列中微表情是相关联的,每一帧图像与前后相邻帧都具有依赖关系,因此本文采用双向 LSTM 结构,如图 5(b)所示.



(a)LSTM　　(b)Bi-LSTM

图 5　LSTM 结构

Fig. 5　Long short-term memory networks structure

对上述网络提取的空间特征 F 构建输入到 LSTM 模块的特征集合;然后送入双向 LSTM 模块中训练并输出;最后经过 dropout 层、FC 层和 softmax 函数得到微表情识别的分类结果.

2.2　损失函数

损失函数可以用于描述模型输出结果和实际结果的差异,值越小即表示输出越接近实际结果.本文的损失函数具体计算如下:

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\, log(p_{i,j}) \tag{14}$$

上式中 L_{ce} 表示损失值,N 表示样本数,M 表示分类数,$y_{i,j}$ 表示实际结果,$p_{i,j}$ 表示预测的输出结果.

3　实验结果及分析

3.1　实验数据集及相关介绍

3.1.1　实验数据集

由于微表情数据的采集不易、制作耗时长且标注尤为困

难,导致样本稀少,目前的研究大多在两个常用的公开可申请的数据库 casme2、samm 上进行,因此本文仅在上述两个数据库上进行实验. casme2 数据库由 7 类常见微表情情感组成,

表 1 数据库信息

Table 1 Information of databases

	casme2	samm
样本数量	255	159
受试者	26	29
分辨率	640 ×480	960 ×650
帧率	200	200
开心(happiness)	32	26
惊讶(surprise)	25	15
厌恶(disgust)	63	9
悲伤(sadness)	7	6
恐惧(fear)	2	8
生气(anger)	–	57
压抑(repression)	27	–
轻蔑(contempt)	–	12
其他(others)	99	26

本文未使用样本数仅为 2 和 7 例的恐惧和悲伤两类情感. samm 数据库包含 8 类情感,本文仅考虑数量相对较多的前 5 类. 表 1 所示为数据库详细信息.

3.1.2 预处理

本文对微表情视频序列进行了人脸矫正配准及裁剪以减少无关信息和干扰. 在起始帧到峰值帧之间采用等间距采样的方法固定输入帧长度并减少干扰帧冗余;使用 EVM[29] 算法放大微表情肌肉运动幅度得到灰度图像序列,两种方法的结合可以大幅度增强微表情运动强度. 参考 Liong[30] 等人的比较结果,选择 TV-L1[31] 处理并得到包含运动信息的光流特征序列. 将两种图像序列通过通道拼接作为输入,可以丰富输入信息. 另外,训练时采用镜像翻转、随机角度旋转,缩放操作进行数据增强以防过拟合,并缓解样本量少且类别严重不均匀的问题.

3.1.3 评估方式

本文验证方法使用留一受试法,即训练时每个受试人的样本数据均有机会作为一次验证集,可以防止数据泄露. 针对样本类别严重不均匀问题,选择以下评价指标.

准确率(Accuracy),其计算公式如下:

$$Accuracy=\frac{\sum_{c=1}^{C}\sum_{k=1}^{K}TP_c^{(k)}}{\sum_{c=1}^{C}\sum_{k=1}^{K}TP_c^{(k)}+\sum_{c=1}^{C}\sum_{k=1}^{K}FP_c^{(k)}} \tag{15}$$

式中 C 表示类别数,K 表示留一受试法的折数,即受试人数,$TP_c^{(k)}$,$FP_c^{(k)}$ 分别表示第 k 折第 c 类的真正例和假正例的数量.

未加权 $F1$ 分数(UF1),可以平等的突出罕见的类别,而不会受类别数量影响. 计算公式如下:

$$UF1=\frac{1}{C}\sum_{c=1}^{C}\frac{2\times\sum_{k=1}^{K}TP_c^{(k)}}{2\times\sum_{k=1}^{K}TP_c^{(k)}+\sum_{k=1}^{K}FP_c^{(k)}+\sum_{k=1}^{K}FN_c^{(k)}} \tag{16}$$

式中 $FN_c^{(k)}$ 表示第 k 折第 c 类的假阴性的数量.

未加权平均召回率(UAR),在评估不均匀数据时能够发挥着重要作用,有助于减轻加权平均召回率存在对较大类别偏向影响的问题. 计算公式如下:

$$UAR=\frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{N_c} \tag{17}$$

式中 N_c 是类别 c 的数量.

3.1.4 实验环境

实验硬件环境采用 NVIDIA V100(8GB)GPU,软件环境为 PyCharm IDE,使用 Python 3.8 语言,以 PyTorch 框架搭建神经网络模型. 训练参数方面,采用 SGD 作为优化器;动量因子设置为(0.9,0.99);学习率设置为 le-2,并进行动态调整,使用余弦退火策略;dropout 层中的丢弃率设定为 0.3,在全连接层中则设置为 0.5;使用交叉熵损失函数;在 casme2 数据集上进行 50 轮的训练,在 samm 数据集上进行 80 轮的训练.

3.2 实验结果分析

表 2 列出了本文方法与两个传统方法以及一系列前沿深度学习方法的实验结果对比. 表中所示方法在预处理细节上有所差异,但使用一致的评价标准及验证方法,仍然可以进行有效比较. 这种对比方式在微表情识别研究中常见且被认可. 为保证结果对比的合理性,表中所示方法均是使用相同数据集和分类类别数,采用一致的评价标准,并使用“留一受试法”进行交叉验证的相关方法,以确保实验基准、评价标准和验证方法的一致性. 在表中,前两项代表采用手工特征的传统方法,而其余则是利用深度学习技术的方法. 根据表 2 的数据,本文方法在两个数据集上均表现出最佳的识别效果,准确率分别为 84.959% 和 74.265%,F1 值分别为 0.855、0.604. 与传统的 LBP 方法相比较,casme2 数据集上准确率提升了 38.499%,F1 值提升了 43.1%;与效果最好的光流方法 Bi-WOOF 相比,准确率分别提升了 26.059%、14.465%,F1 值分别提升了 24.5%、1.3%. 相比于目前先进的方法 AM3F-FlowNet,本文方法在准确率上分别提升 0.439%、8.082%,F1 值分别提升了 2.62%、6.3%;相比于 SLSTT 方法,准确率也分别提升了 9.153% 和 1.877%;综上表明了本文方法的有效性.

表 2 与其他方法的比较

Table 2 Comparison with the other methods

方 法	casme2		samm	
	Acc	F1	Acc	F1
LBP-TOP[6]	46.46	0.424	–	–
Bi-WOOF[7](2018)	58.9	0.610	59.8	0.591
CNN + LSTM[15](2016)	60.98	–	–	–
ELRCN[16](2018)	52.44	0.500	–	–
MERTA[23](2019)	60.54	–	–	–
Micro-attention[9](2020)	65.90	0.53	48.50	0.402
KFC-MER[24](2021)	72.76	0.7375	63.24	0.5709
SLSTT[20](2022)	75.806	0.753	72.388	0.640
AM3F-FlowNet[25](2023)	84.52	0.8288	66.18	0.5410
本文	84.959	0.855	74.265	0.604

注:实验结果数据均引自相关文献,– 表示无相关数据.

图 6 是本文方法的具体实验分类结果. 图 6(a)是 casme2 数据集上的分类混淆矩阵,一般来说,“surprise”的面部肌肉运动幅度可能相对较大,使用注意机制可以增加对变化显著区域的关注,提取到有效特征,从而获得较好的识别效果;但“repression”的运动幅度更微弱,一般的方法难以从相关局部区域中提取到可用于区分该类别的有效特征,因此现有的方

法大多难以准确识别出“repression”类别的微表情,主要原因可能是微表情有较多的相似 AU,且一些 AU 的运动部位相近并有重合部位,识别时容易发生混淆,进而误识为其他类别. 比如与“repression”相关的 AU 包括 AU15(发生在嘴角区域)、AU17(发生在下巴区域),“surprise”的发生部位也包含嘴角区域,并且与嘴角相关的 AU 还有很多,因此在区分时容易发生混淆;此外,对于 AU17 发生的区域,一般的方法对此区域的关注很少,很容易忽略该区域,也会导致识别率较低. 但由图可知,本文的方法不仅能够很好地识别出“surprise”类别,而且对于“repression”类别也有很好的识别效果. 如图 6(a)所示,本文提出的模型在“surprise”和“repression”微表情上实现了最高识别准确率,分别为 96% 和 93%,这说明了本文提出的改进不仅能够关注到那些变化显著的局部区域,也能关注运动较微弱的区域,实现对多区域有效特征的提取,减少细微特征的丢失,进而达到较好的性能.

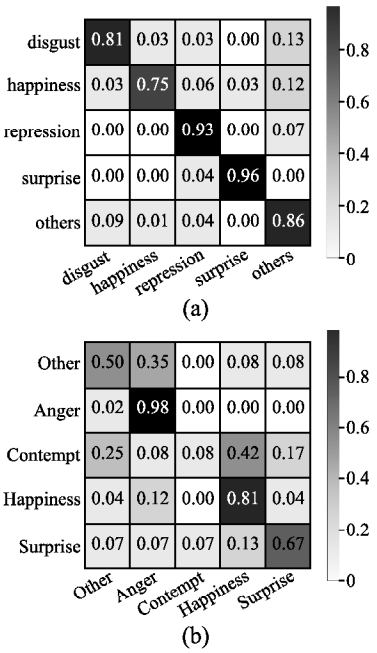


图 6　Casme2、samm 的混淆矩阵

Fig. 6　Confusion matrix of casme2 and samm

samm 数据集相比于 casme2 数据集样本数量更少,而样本量对深度学习模型至关重要,因此本文模型在该数据集的识别效果相对劣于 casme2 数据集,但仍然优于目前的主流方法. 由图 6(b)所示,“anger”微表情的准确率分别为 98%,这是因为数据集中该类微表情占比最高,因此能够很好地学习到相关特征. “happiness”微表情的准确率为 81%,取得不错的效果,这都是依赖于改进模块中使用注意力机制使得模型学习多区域特征、多层特征融合模块使得模型可以利用深层浅层特征更好地表示微表情. 而“contempt”的识别效果较差,这是因为该类别样本数仅为 12 例,因此训练时提取相关特征较为困难,且与其相关的运动单元为 AU12、AU14,因此容易误识为“happiness”,对其识别准确率还有待提升.

3.3　消融实验

为了评估本文提出的多区域特征提取模块 MFEM 和多层特征融合模块 MFFM 在微表情识别任务中的效用,在 casme2 数据集上进行了消融实验,使用准确率、UF1 和 UAR 指标,采用相同的数据、环境及配置,结果如表 3 所示,其中 base 代表本文所使用的未引入改进模块的 ResNet34 + LSTM 级联的基础网络模型.

表 3　消融实验

Table 3　Ablation experiments

方　法	casme2		
	Acc	UF1	UAR
base	73.171	0.718	0.703
base + MFEM	78.455	0.761	0.755
base + MFFM	78.455	0.750	0.741
base + MFEM + MFFM	84.959	0.855	0.861

由表 3 数据可知,当引入了多区域特征提取模块 MFEM 后,准确率提升了 5.284%,UF1、UAR 分别提升了 4.3%、5.2%,主要原因在于该模块中的注意力机制可以引导模型关注多个可以表达微表情的相关的面部局部区域,有利于微表情特征提取. 引入多层特征融合模块 MFFM 后,准确率提升了 5.284%,UF1、UAR 分别提升了 3.2%、3.8%,主要原因在于浅层阶段中提取的特征涵盖了微表情的一些细节信息和局部信息,深层阶段中提取的特征则更多地表现为高级语义特征,而同时利用这些特征可以减少细微信息的丢失,增强了特征的多样性,产生更具表现力的特征,能够更好地表示微表情. 同时引入 MFEM 和 MFFM 模块后,准确率提升了 11.788%,UF1、UAR 分别提升了 13.7%、15.8%,说明两个模块结合使用可以充分发挥作用,证明了本文提出的改进可以有效提升微表情识别的性能.

4　总　结

针对由于表微表情自身运动强度微弱、局部性等特点引起的对微表情视频序列中相关的面部局部区域进行有效特征提取十分困难的问题,本文以运动放大的灰度图像序列与光流图像序列代替原始微表情视频序列作为输入,并在残差和长短期记忆级联的基础网络上提出了结合多区域特征提取和多层特征融合的微表情识别网络模型. 在多区域特征提取模块中引入注意力机制加强了对微表情视频序列中显著的局部区域和次级显著的局部区域的关注,从而提升了网络模型对多个相关区域中显著特征的提取能力. 同时,通过特征融合对网络中提取的各个阶段的深层和浅层特征进行交互和融合,降低了细微特征的损失,产生了更综合、更具表现力的特征表示,从而增强了网络模型对微表情特征的学习表示能力. 根据 casme2 和 samm 数据集上的实验结果显示,本文方法实现了较优的效果,为改进模块的有效性提供了有力支持.

之后的研究将考虑以下两个方面:1)微表情数据集样本稀缺,且类别严重不均匀,这极大影响了微表情识别的性能,因此提出有效的数据增强方法及如何解决决策边界偏移问题值得探索. 2)目前的微表情识别方法仅依靠单一的视频序列图像信息,未来可以考虑多模态信息融合的方法进行微表情识别,以期达到更好的效果.

References:

[1] Vrij A, Mann S. Who killed my relative? police officers' ability to

detect real-life high-stake lies[J]. Psychology, Crime Law, 2001, 7(2): 119-132.

[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] Sharon W. Airport security: intent to deceive? [J]. Nature, 2010, 465(7297): 412-415.

[4] Yan W, 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.

[5] Ekman P, Friesen V W. Nonverbal leakage and clues to deception [J]. Psychiatry, 2016, 32(1): 88-106.

[6] Pfister T, Li X, Zhao G, et al. Recognising spontaneous facial micro-expressions[C]//International Conference on Computer Vision, IEEE, 2011: 1449-1456.

[7] Liong S, See J, Wong K, et al. Less is more: micro-expression recognition from video using apex frame[J]. Signal Processing: Image Communication, 2018, 62: 82-92, doi: 10.1016/j.image.2017.11.006.

[8] ZHANG R, HE N. A survey of micro-expression recognition methods[J]. Computer Engineering and Applications, 2021, 57(1): 38-47.

[9] Wang C, Peng M, Bi T, et al. Micro-attention for micro-expression recognition [J]. Neurocomputing, 2020, 410: 354-362, doi: 10.1016/j.neucom.2020.06.005.

[10] 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.

[11] 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.

[12] 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.

[13] YU M, ZHONG Y X, WANG Y. A survey of facial micro-expression analysis methods[J]. Computer Engineering, 2023, 49(2): 1-14.

[14] Patel D, Hong X, Zhao G. Selective deep features for micro-expression recognition[C]//International Conference on Pattern Recognition, IEEE, 2017: 2258-2263.

[15] Kim D H, Baddar W J, Ro Y M. Micro-expression recognition with expression-state constrained spatio-temporal feature representations [C]//Proceedings of the ACM on Multimedia Conference, New York: ACM Press, 2016: 382-386.

[16] Khor H Q, See J, Phan R C W, et al. Enriched long-term recurrent convolutional network for facial micro-expression recognition [C]//Proceedings of The 13th IEEE International Conference on Automatic Face and Gesture Recognition, 2018: 667-674.

[17] Reddy T P S, Karri T S, Dubey R S, et al. Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks[J]. CoRR, 2019, abs/1904.01390, doi: 10.1109/IJCNN.2019.8852419.

[18] Y S G, Sze Teng L, Wei Chuen Y, et al. Off-apexnet on micro-expression recognition system[J]. Signal Processing: Image Communication, 2019, 74: 129-139, doi: 10.48550/arXiv.1805.08699.

[19] Liong S T, Gan S Y, See J, et al. A shallow triple stream three-dimensional cnn (STSTNet) for micro-expression recognition[C]//14th IEEE International Conference on Automatic Face & Gesture Recognition, 2019: 1-5.

[20] Zhang L F, Hong X P, Member, 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.

[21] Li Yante, Huang Xianghua, Zhao Guoying. Joint local and global information learning with single apex frame detection for micro-expression recognition[J]. IEEE Trans on Image Processing, 2020, 30: 249-263, doi: 10.1109/TIP.2020.3035042.

[22] Yao Liansheng, Xiao Xiao, Cao Ranlei, et al. Three stream 3d cnn with se block for micro-expression r ecognition[C]//Proceedings of The International Conference on Computer Engineering and Application, 2020: 439-443.

[23] Yang B, Cheng J, Yang Y, et al. MERTA: micro-expression recognition with ternary attentions[J]. Multimedia Tools and Applications, 2019, 80(11): 1-16.

[24] Su Y, Zhang J, Liu J, et al. Key facial components guided micro-expression recognition based on first & second order motion[C]//IEEE International Conference on Multimedia and Expo, 2021: 1-6.

[25] Chenghao F, Wenzhong Y, Danny C, et al. AM3F FlowNet: attention-based multi-scale multi-branch flow network [J]. Entropy, 2023, 25(7), doi: 10.3390/E25071064.

[26] ZHAO M H, DONG S S, HU J, et al. Attention guided three-stream convolutional neural network for microexpression recognition[J]. Journal of Image and Graphics, 2024, 29(1): 111-122.

[27] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision, 2018: 3-19.

[28] Filip R, Giorgos T, Ondrej C. Fine-tuning cnn image retrieval with no human annotation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(7): 1655-1668.

[29] Wu H, Rubinstein M, Shih E, et al. Eulerian video magnification for revealing subtle changes in the world [J]. ACM Transactions on Graphics, 2012, 31(4): 1-8.

[30] Liong S T, Gan Y S, Zheng D N, et al. Evaluation of the spatio-temporal features and gan for micro-expression recognition system [J]. Journal of Signal Processing Systems, 2020, 92(1): 705-725.

[31] Zach C, Pock T, Bischof H. A duality based approach for realtime tv-l1 optical flow[C]//Proceedings of the 29th DAGM Conference on Pattern Recognition, 2007: 214-223.

附中文参考文献：

[8] 张 人,何 宁. 微表情识别研究综述[J]. 计算机工程与应用, 2021, 57(1): 38-47.

[13] 于 明,钟元想,王 岩. 人脸微表情分析方法综述[J]. 计算机工程, 2023, 49(2): 1-14.

[26] 赵明华,董爽爽,胡 静,等. 注意力引导的三流卷积神经网络用于微表情识别[J]. 中国图象图形学报, 2024, 29(1): 111-122.