

基于视觉长短期记忆网络和双重注意力机制的微表情识别

包永堂,刘福仪,张 鹏

(山东科技大学 计算机科学与工程学院,山东 青岛 266590)

摘 要:针对微表情识别中特征提取不充分、有效特征定位不准确以及传统的卷积神经网络难以有效捕捉和表征微表情动态特征等问题,本研究提出一种基于视觉长短期记忆网络和双重注意力机制的微表情识别模型(ViLDA)。该模型将视觉长短期记忆网络引入到微表情识别中,获取更具信息量的特征和面部运动区域之间的依赖关系;为关注微表情重要的细微肌肉运动特征并去除噪声,设计了中值增强的空间通道双重注意力模块;设计了基于小波变换的纹理增强模块,用于增强面部静态纹理特征;模型融合静态特征和细微肌肉运动特征对微表情进行分类。在公开的微表情数据集 SMIC、CASME II 和 SAMM 上进行实验验证表明,识别准确率分别达到 0.737 8、0.821 1 和 0.703 7,F1 值分别达到 0.731 0、0.820 7 和 0.624 7;与 KPCANet、FRL-DGT、I3D-MOCO、Meta-MMFNet、GTS-GN-AL 和 LFGNet 等当前先进的模型相比,该模型实现了更优异的识别性能。

关键词:微表情识别;视觉长短期记忆网络;纹理特征;双重注意力机制;小波变换

中图分类号:TP391

文献标志码:A

Micro-expression recognition based on vision-LSTM and dual attention mechanism

BAO Yongtang, LIU Fuyi, ZHANG Peng

(College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China)

Abstract: To address the issues of insufficient micro-expression feature extraction, inaccurate motion feature localization, and the limitations of traditional convolutional neural networks in effectively capturing and representing dynamic micro-expression features, this study proposed a micro-expression recognition model based on vision long- and short-term memory (Vision-LSTM) networks and a dual attention mechanism (ViLDA). This model introduced vision-LSTM networks into micro-expression recognition for the first time, enabling the extraction of more informative features and capturing the dependencies among facial motion regions. A median-enhanced spatial-channel dual attention module was designed to focus on subtle and significant micro-expression muscle movements while reducing noise. A wavelet transform-based texture enhancement module was also developed to improve facial static texture features. By integrating static features with subtle muscle motion features, the model achieves classification. Experiments on public micro-expression datasets of SMIC, CASME II, and SAMM yielded recognition accuracies of 0.737 8, 0.821 1, and 0.703 7, respectively, with respective F1 scores of 0.731 0, 0.820 7, and 0.624 7. Compared to state-of-the-art methods, such as KPCANet, FRL-DGT, I3D-MOCO, Meta-MMFNet, GTS-GN-AL, and LFGNet, the proposed model demonstrates superior recognition performance.

收稿日期:2024-10-15

基金项目:教育部人文社会科学研究规划基金项目(25YJAZH005);山东省自然科学基金项目(ZR2021QF017);山东省泰山学者计划项目(tstp20221128);青岛市科技惠民计划项目(25-1-5-smjk-18-nsh);山东科技大学青年教师教学拔尖人才计划项目(BJ20231201)

作者简介:包永堂(1983—),男,山东青州人,副教授,博士,主要从事情感计算、人工智能等研究。

张 鹏(1990—),男,山东日照人,副教授,博士,主要从事计算机视觉和机器学习研究,本文通信作者。

E-mail: pengzhang_skd@sdu.edu.cn

Key words: micro-expression recognition; vision-LSTM; texture features; dual attention mechanism; wavelet transform

微表情识别技术在帮助人类理解情绪方面起着至关重要的作用,并对人机交互具有重要意义^[1]。与宏观表情不同,微表情具有自发性(非人为控制)、快速性(持续时间通常在 1/25~1/5 s)和微妙性(涉及面部区域小)等特点,能够更准确地揭示个体试图隐藏的真实情绪,反映其在特定环境中的真实感受。微表情识别已广泛应用于谎言检测、临床诊断和商业谈判等领域。

在微表情识别研究的初始阶段,主要依赖于手工特征提取,该方法可以保留局部变化区域或去除无关区域。然而,由于人脸微表情具有复杂的运动模式,且不同的微表情与不同的动作单元相关联,导致难以提取更深层次的特征^[2]。近年来,随着深度学习在计算机视觉领域的成功应用^[3],微表情识别技术得以推动发展^[4],识别性能显著提升。例如,Xia 等^[5]利用无参数注意力机制的单个递归神经网络来捕获微表情特征。Lei 等^[6]使用神经网络模型直接学习微表情视频序列中的顶点帧,以获取微表情相关的语义特征。Zhang 等^[7]结合长短期记忆网络(long short-term memory, LSTM),提出一种新的 Transformer 架构,用于同时捕获微表情的短期和长期时空特征。Wei 等^[8]提出一种基于注意力机制的微表情自适应放大网络,以获取放大后的微表情特征;Bao 等^[9]结合自表情重建和原型记忆对比学习技术,学习鲁棒的微表情特征;Wang 等^[10]提出一个新的轻量级网络,并结合注意力机制来提取微表情特征。注意力机制的引入提高了微表情识别的准确率,但在处理运动幅度极小的微表情时,依然存在难以精准定位有效特征的问题。此外,微表情通常集中在面部的局部区域,这些局部区域之间存在依赖关系,传统卷积神经网络是局部感受野,难以有效捕捉这些依赖关系。

针对上述问题,本研究提出一种基于视觉长短期记忆网络和双重注意力机制的微表情识别模型(vision-LSTM and dual attention model, ViLDA)。该模型引入视觉长短期记忆网络(Vision-LSTM)^[11],利用其交替遍历机制和矩阵记忆结构,提取更具信息量的特征,捕捉全局上下文信息,学习不同面部区域之间的依赖关系。此外,提出一种中值增强的空间通道注意力模块(median-enhanced spatial-channel attention module, MESCA),以不同尺度捕捉和融合特征,关注细微肌肉运动特征的同时去除噪声,从而解决有效特征定位不准确的问题。同时,设计一种基于小波变换的纹理增强模块(wavelet transform-based texture enhancement module, WTTE),用于增强微表情序列顶点帧的面部静态纹理特征。最后,融合静态特征和细微肌肉运动特征,进一步提高表征能力。实验结果表明,ViLDA 模型在多个微表情数据集上均取得了优异的性能。

1 ViLDA 模型

1.1 模型总体结构

本研究提出的微表情识别模型 ViLDA 的总体结构如图 1 所示,由数据预处理、特征提取和表情分类三部分组成。图 1 中, MESCA 为中值增强的空间通道注意力模块, WTTE 为基于小波变换的纹理增强模块, FC 表示全连接层, L_{ce} 表示交叉熵损失。首先,选取视频序列中的起始帧和顶点帧作为模型输入,并对两帧进行差分处理。然后,将差分图输入到 Vision-LSTM 网络,依次经过图像块分割、图像块嵌入、位置嵌入和交替的 mLSTM(matrix LSTM)块等操作,以提取更丰富的特征。为关注到细微肌肉运动特征并去除噪声,将提取的特征输入到 MESCA 模块。同时,将微表情的顶点帧输入到 WTTE 模块进行纹理增强,并对增强后的顶点帧进行卷积、ReLU 激活和池化操作,以提取微表情静态纹理特征。最后,将静态特征和细微肌肉运动特征进行融合,并通过一个全连接层进行分类,模型采用交叉熵损失函数进行训练。

1.2 数据预处理

微表情预处理包括起始帧和顶点帧选取、差分图生成。起始帧代表视频序列中微表情运动初始发生的帧,顶点帧代表视频序列中微表情运动强度最高的帧,顶点帧对于获取判别微表情的有效特征至关重要。本研究直接选取 CSAME II^[12]和 SAMM^[13]数据集中标注的起始帧和顶点帧, SMIC^[14]数据集没有标

注起始帧和顶点帧,本研究采用 Bao 等^[9]提出的方法选取。该方法先从给定的一段微表情视频序列中选择起始帧和顶点帧构成图像对,表示为 (X_1, X_2) ,然后计算两帧之间的差分信息,将其作为模型输入之一,另一个输入则直接选取顶点帧 X_2 。

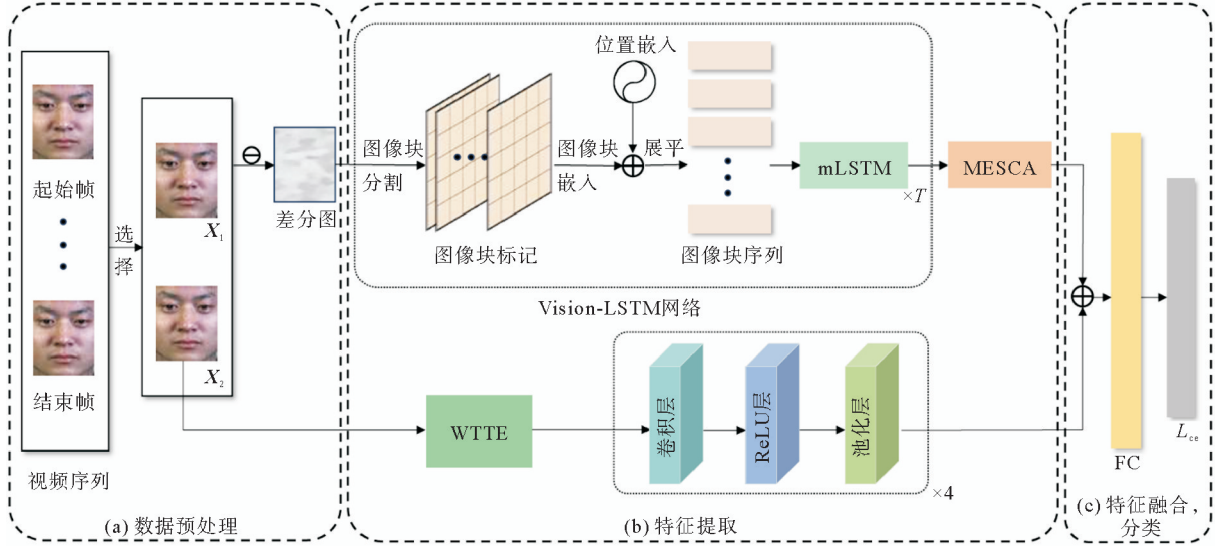


图 1 ViLDA 模型总体结构

Fig. 1 Overall structure of ViLDA model

1.3 Vision-LSTM 网络

传统的卷积神经网络由于其局部感受野的限制,无法捕捉不同面部运动区域间的依赖关系。因此,本研究引入 Vision-LSTM 作为骨干网络,利用其交替遍历机制和矩阵记忆结构,获取更具信息量的特征,从而捕捉不同面部区域间的依赖关系。

$I \in \mathbb{R}^{B \times C \times H \times W}$ 表示由差分信息构成的差分图像,其中 B 为批量大小, C 为通道数, H 为图像高度, W 为图像宽度。微表情的数据样本经过预处理后,得到差分图像的通道数为 3,图像高度和宽度皆为 224 像素。差分图像依次经过图像块分割、图像块嵌入、位置嵌入和交替的 mLSTM 块处理 4 个阶段:

1) 图像块分割。微表情涉及局部的肌肉运动,若采用整张图像作为输入,面部局部区域的微小肌肉运动特征容易被面部其他静止区域掩盖。通过图像块分割, Vision-LSTM 网络可以有效关注到不同面部区域的肌肉运动,从而避免被无关区域干扰。给定输入 $[B, C, H, W]$,采用 16×16 像素的图像块进行分割,则图像被分割为多个不重叠且相邻的 16×16 局部区域,每个图像块的特征维度为 $16 \times 16 \times C$,图像块数量为 $(H/16) \times (W/16)$ 。

2) 图像块嵌入。图像块嵌入将原始像素转换为更高层次的语义特征,使 Vision-LSTM 网络更有效地捕捉微表情的细微变化。为了既能保留足够的判别信息,又避免维度冗余,本研究将嵌入维度 D 设置为 192。在图像块嵌入过程中,每个图像块投影到维度 D ,从而将差分图像的特征维度由 $[B, H/16, W/16, 16 \times 16 \times C]$ 转化为 $[B, H/16, W/16, D]$ 。

3) 位置嵌入。针对微表情多区域协同变化的特点,在差分图像经过图像块分割和图像块嵌入后,为每个图像块特征向量添加一个维度为 $[1, H/16, W/16, D]$ 的可学习向量作为位置信息,使 Vision-LSTM 网络能关联不同区域的信息,从而捕捉微表情的协同肌肉运动特征。

4) 交替的 mLSTM 块。交替的 mLSTM 块采用奇数块从左到右、偶数块从右到左的顺序处理图像块序列,使 Vision-LSTM 网络能够更全面地捕捉微表情差分图像中的长距离依赖与局部特征。这种交替处理增强了不同面部区域的关联性,从而提升微表情识别性能。交替的 mLSTM 块如图 2 所示。其中, P_F 为投影因子, Flip 表示翻转序列操作, SiLU 为激活函数。图像块特征向量 $[B, H/16, W/16, D]$ 在进入 mLSTM 块之前,特征向量中的 $[H/16, W/16]$ 会被转换成一维向量 N ,得到输入的图像块序列 $u \in$

$\mathbf{R}^{B \times N \times D}$ 。图像块序列 $\mathbf{u}$ 首先通过 Norm 层进行归一化处理,然后通过上投影操作($P_F=2$)进行线性变换,将图像块序列的维度由 D 变成 $2D$,接着将该序列拆分为两部分 $\mathbf{u}_1$ 和 $\mathbf{u}_2$ 。 $\mathbf{u}_1$ 经过翻转(Flip)后得到 $\mathbf{u}_t$,并被送往 mLSTM 层。mLSTM Layer 通过线性投影从 $\mathbf{u}_t$ 中计算查询 $\mathbf{q}_t$ (Query)、键 $\mathbf{k}_t$ (Key) 和值 $\mathbf{v}_t$ (Value),用于后续的矩阵记忆更新和特征检索。mLSTM 层采用协方差更新规则维护矩阵记忆 $\mathbf{C}_t$,用于长期存储关键特征,使 Vision-LSTM 网络能够学习长距离依赖。 $\mathbf{C}_t$ 的计算式为:

$$\mathbf{C}_t = f_t \mathbf{C}_{t-1} + i_t \mathbf{v}_t \mathbf{k}_t^T. \quad (1)$$

式中: f_t 是遗忘门,通过控制历史特征的保留来抑制静止区域的特征干扰; i_t 是输入门,通过控制新特征的存入来强化肌肉运动区域的特征。

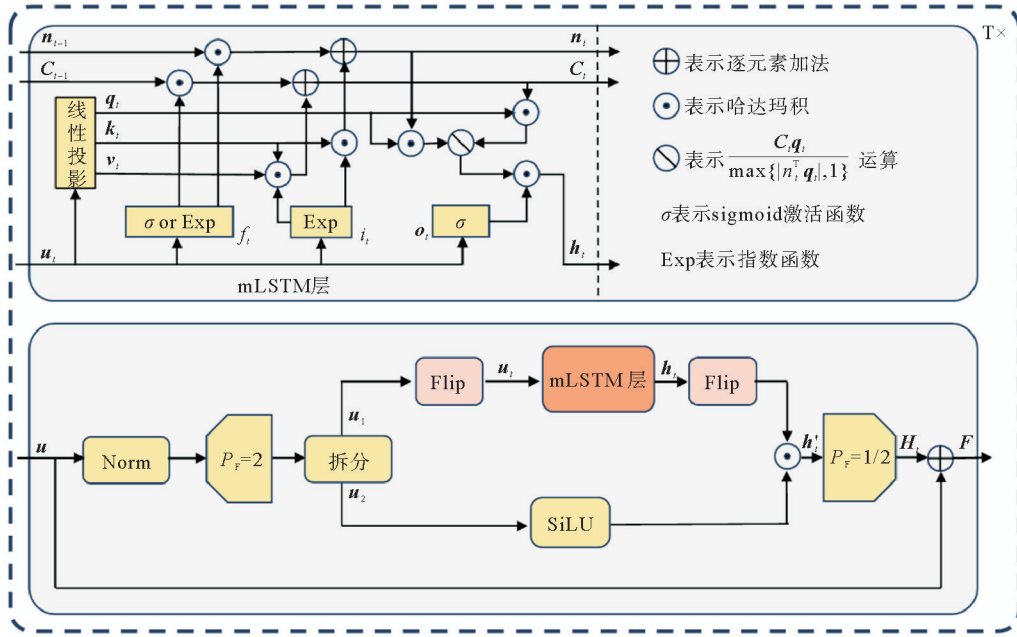


图 2 交替的 mLSTM 块

Fig. 2 Alternating mLSTM block

归一化状态 $\mathbf{n}_t$ 用于追踪记忆更新的缩放因子,以确保矩阵记忆 $\mathbf{C}_t$ 在计算隐藏状态 $\mathbf{h}_t$ 的数值稳定性。计算式为:

$$\mathbf{n}_t = f_t \mathbf{n}_{t-1} + i_t \mathbf{k}_t. \quad (2)$$

隐藏状态 $\mathbf{h}_t$ 是 mLSTM Layer 最终输出的序列,由输出门 $\mathbf{o}_t$ 、矩阵记忆 $\mathbf{C}_t$ 、归一化状态 $\mathbf{n}_t$ 和查询 $\mathbf{q}_t$ 计算得到,计算式为:

$$\mathbf{h}_t = \mathbf{o}_t \odot \frac{\mathbf{C}_t \mathbf{q}_t}{\max\{|\mathbf{n}_t^T \mathbf{q}_t|, 1\}}. \quad (3)$$

式中, $\odot$ 表示元素级相乘(哈达玛积)。

隐藏状态 $\mathbf{h}_t$ 经过再次翻转(Flip)后,首先与经过 SiLU 激活函数处理的图像块序列 $\mathbf{u}_2$ 进行逐元素相乘,生成图像块序列 $\mathbf{h}_t'$;然后,通过下投影操作($P_F=1/2$)的线性变换,将序列维度从 $2D$ 降低到 D ,得到图像块序列 $\mathbf{H}_t$;最后,将图像块序列 $\mathbf{H}_t$ 与初始输入的图像块序列 $\mathbf{u}$ 相加,得到最终的特征 $\mathbf{F}$ 。

1.4 双重注意力机制

为了使模型更精准地关注到细微的肌肉运动特征,从而有效解决微表情识别中特征聚焦不准确的问题,本研究设计了一个中值增强的空间通道注意力模块(MESCA),如图 3 所示。该模块能够以不同尺度捕捉和融合特征,在保留微表情动态特征信息的同时去除噪声。首先,如图 3(a)所示,通道注意力机制通过聚合输入微表情特征 $\mathbf{F}$ 的全局信息,生成通道注意力图,并以此对输入特征的各个通道加权,过程如下。

1) 对输入特征图 $\mathbf{F}$ 进行全局平均池化、全局最大池化和全局中值池化操作,得到 3 个不同的池化结果。

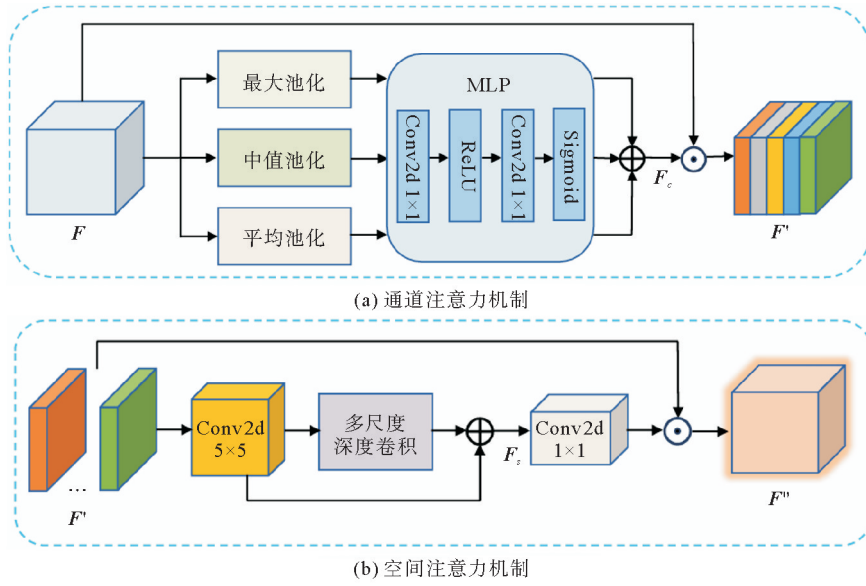


图 3 中值增强的空间通道注意力模块

Fig. 3 Median-enhanced spatial-channel attention module

2) 将每个池化结果输入共享的多层感知器(multilayer perceptron, MLP)中。该 MLP 包含两个 1×1 卷积层、一个 ReLU 激活函数和一个 Sigmoid 激活函数。第一个卷积层将特征维度从 D 降到 D/r (r 为降维比率),第二个卷积层将特征维度恢复到 D 。最后,使用 Sigmoid 激活函数将输出值压缩到 $[0, 1]$ 范围内,得到 3 个注意力图。

3) 将 3 个池化结果的注意力图进行元素级相加,得到最终的通道注意力图 F_c ,

$$F_c = \text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F)) + \text{MLP}(\text{MedianPool}(F)). \quad (4)$$

式中: $\text{AvgPool}()$ 、 $\text{MaxPool}()$ 和 $\text{MedianPool}()$ 分别表示平均池化、最大池化和中值池化, $\text{MLP}()$ 表示多层感知器。

4) 将通道注意力图与原始输入特征图进行元素级相乘,得到加权后的特征图,即:

$$F' = F_c \odot F. \quad (5)$$

然后,如图 3(b)所示,空间注意力机制通过多尺度深度卷积层捕捉输入特征图的空间关系,并生成空间注意力图,其过程如下。

1) 输入的特征图 F' 通过一个 5×5 的深度卷积层提取特征。

2) 将得到的特征通过多尺度深度卷积层进行并行处理,其中异构尺寸的卷积核 (1×7 、 7×1 、 1×11 、 11×1 、 1×21 、 21×1) 分别提取不同感受野下的空间特征。

3) 将前两步得到的输出进行元素级相加融合后得到特征图 F_s 。

4) 将 F_s 通过一个 1×1 卷积层,生成最终的空间注意力图。该注意力图与通道加权后的特征图 F' 进行元素级相乘,得到最终的输出特征图 F'' ,

$$F'' = \text{Conv}1 \times 1(F_s) \odot F'. \quad (6)$$

1.5 纹理增强模块

静态特征对微表情识别同样具有重要作用,因此,本研究设计了一种基于小波变换的纹理增强模块(WTTE),如图 4 所示,旨在增强微表情面部纹理特征。通过提取增强后的微表情纹理特征,并将其与肌肉运动特征相融合,实现动态特征和静态的结合,从而进一步增强微表情表征能力,具体过程如下。

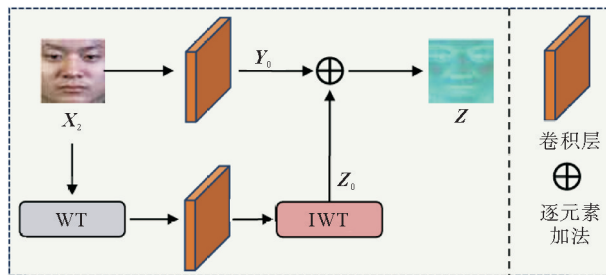


图 4 基于小波变换的纹理增强模块

Fig. 4 Wavelet transform-based texture enhancement module

首先,对输入的顶点帧图像 $\mathbf{X}_2$ 进行卷积操作:

$$\mathbf{Y}_0 = \text{Conv}(\mathbf{W}, \mathbf{X}_2)。(7)$$

式中, $\mathbf{W}$ 是 $k \times k$ 深度可分离卷积核的权重张量。

其次,对顶点帧图像 $\mathbf{X}$ 进行小波变换分解(WT()):

$$[\mathbf{X}_{\text{LH}}, \mathbf{X}_{\text{HL}}, \mathbf{X}_{\text{HH}}] = \text{WT}(\mathbf{X}_2)。(8)$$

式中, $\mathbf{X}_{\text{LH}}$ 、 $\mathbf{X}_{\text{HL}}$ 、 $\mathbf{X}_{\text{HH}}$ 分别表示其水平、垂直、对角高频分量。

然后,对 $\mathbf{X}_{\text{LH}}$ 、 $\mathbf{X}_{\text{HL}}$ 和 $\mathbf{X}_{\text{HH}}$ 分别进行卷积操作(Conv()):

$$\mathbf{Y}_{\text{LH}}, \mathbf{Y}_{\text{HL}}, \mathbf{Y}_{\text{HH}} = \text{Conv}(\mathbf{W}, (\mathbf{X}_{\text{LH}}, \mathbf{X}_{\text{HL}}, \mathbf{X}_{\text{HH}}))。(9)$$

接下来,组合 $\mathbf{Y}_{\text{LH}}$ 、 $\mathbf{Y}_{\text{HL}}$ 和 $\mathbf{Y}_{\text{HH}}$ 进行逆小波变换(IWT()):

$$\mathbf{Z}_0 = \text{IWT}(\mathbf{Y}_{\text{LH}}, \mathbf{Y}_{\text{HL}}, \mathbf{Y}_{\text{HH}})。(10)$$

最后,将 $\mathbf{Y}_0$ 和 $\mathbf{Z}_0$ 相加得到纹理增强的顶点帧:

$$\mathbf{Z} = \mathbf{Y}_0 + \mathbf{Z}_0。(11)$$

2 实验结果与分析

2.1 数据集

为了与现有的模型进行对比,本研究在 SMIC、CASME II 和 SAMM 等 3 个公共基准数据集上进行了实验。本研究采用单数据库实验(single database experiment, SDE)设置,在 CASME II 和 SAMM 数据集上进行 5 类微表情识别,在 SMIC 数据集上进行 3 类微表情识别。表 1 展示了 3 个数据集中用于 SDE 的微表情样本的详细信息。

SMIC 数据集提供了来自 3 种不同类型的记录相机(HS,高速相机;VIS,视觉相机;NIR,近红外相机)的数据子集。HS 能够捕获到微表情的瞬态变化,通常用于微表情实验。这些样本来自 16 名参与者,分辨率为 640×480 ,包含 3 种情绪:积极、消极和惊讶。

CASME II 数据集由相机以每秒 200 帧的速度录制。在严格控制的实验环境中,26 名参与者生成了这些样本,分辨率为 640×480 。此外,CASME II 包含 5 种情绪:高兴、压抑、惊讶、厌恶和其他。所有的样本都包含了起始帧、顶点帧、结束帧以及动作单元的标注。

SAMM 数据集的分辨率为 2040×1088 ,由高速相机以每秒 200 帧的速度录制。SAMM 包含 7 种情绪:高兴、惊讶、愤怒、蔑视、悲伤、恐惧和其他。每位参与者都需要完成一份问卷和自我报告,以准确标记情绪。与 CASME II 类似,SAMM 所有样本都包括了起始帧、顶点帧、结束帧以及动作单元的标注。

2.2 实验设置

实验采用了留一被试法(leave-one-subject-out, LOSO)作为交叉验证方法,评估模型的性能。LOSO 交叉验证的使用规则是:在每次验证时,选取其中一位受试者的全部数据作为测试集,其余受试者的全部数据则作为训练集。

实验环境为 Window 11, GPU 选用 NVIDIA GeForce RTX 4060, Pytorch 版本为 2.0.0;所有数据集的图像大小调整为 224×224 ;优化器使用 Adam,学习率初始化为 0.000 8,权重衰减系数设置为 0.9, batchsize 大小为 4,单个受试者的训练周期设置为 70。

2.3 评估指标

为了评估不同模型在 SDE 上的性能,本研究使用两个指标:准确率和 F1 分数,来衡量单个数据集上的性能。准确率可表示为:

$$A = \frac{P}{N}。(12)$$

表 1 3 个数据集的详细信息

Table 1 Details of three datasets

SMIC		CASME II		SAMM	
情绪	数量	情绪	数量	情绪	数量
消极	71	高兴	32	高兴	26
积极	51	压抑	27	愤怒	57
惊讶	42	惊讶	25	蔑视	11
—	—	厌恶	63	惊讶	15
—	—	其他	99	其他	26

式中: A 为准确率, P 为正确预测结果的总数, N 为测试数据的总数。

F1 分数为每个类别的测试数据分配相同的权重。基于混淆矩阵,计算了每一类微表情的真阳性 T_P^c 、假阳性 F_P^c 和假阴性 F_N^c 。最终的 F1 分数是平均每个类别的 F1 分数,用 F 表示。

$$F_c = \frac{2 \times T_P^c}{2 \times T_P^c + F_P^c + F_N^c}, \quad (13)$$

$$F = \frac{1}{C} \sum_{c=1}^C F_c. \quad (14)$$

式中: F_c 表示类别 c 的 F1 分数, C 表示类别的数量。

2.4 实验结果对比

本研究与其他 9 个先进的深度学习方法模型进行了对比,分别为 SLSTT^[7]、KPCANet^[15]、FRL-DGT^[16]、I3D-MOCO^[17]、Meta-MMFNet^[18]、MFE-Net^[10]、PCB-PCANet+^[19]、GTS-GN-AL^[20] 和 LFGNet^[21],实验结果如表 2 所示。从表中可以看出,与其他模型相比,本模型在准确率、F1 分数上均取得较好的表现,特别是在 CSAME II 数据集上的分类准确率和 F1 分数达到 0.821 1 和 0.820 7,优于其他模型。由于 SMIC 和 SAMM 数据集更小的样本量以及种族多样性,这增加了模型泛化的难度,导致模型在这两个数据集上的识别性能没有达到最优。

表 2 对比实验结果

Table 2 Results of the comparative experiments

模型	SMIC		CASME II		SAMM	
	准确率	F1 分数	准确率	F1 分数	准确率	F1 分数
SLSTT ^[7]	0.737 1	0.724 0	0.758 1	0.753 0	0.723 9	0.640 0
KPCANet ^[15]	0.717 4	0.710 0	0.704 6	0.659 2	0.638 3	0.521 5
FRL-DGT ^[16]	—	—	0.757 0	0.748 0	—	—
I3D-MOCO ^[17]	0.756 1	0.749 2	0.763 0	0.736 6	0.683 8	0.543 6
Meta-MMFNet ^[18]	0.631 3	—	0.809 5	—	—	—
MFE-Net ^[10]	—	—	0.789 3	0.805 3	—	—
PCB-PCANet+ ^[19]	0.646 3	0.646 7	0.556 9	0.557 4	—	—
GTS-GN-AL ^[20]	—	—	0.817 8	0.812 0	0.713 2	0.711 0
LFGNet ^[21]	0.733 5	0.740 1	0.738 8	0.743 9	0.639 5	0.642 2
ViLDA	0.737 8	0.731 0	0.821 1	0.820 7	0.703 7	0.624 7

2.5 混淆矩阵结果

为了更深入地分析分类结果,本研究将 ViLDA 模型在 3 个数据集上的混淆矩阵用表 3 进行呈现。从表 3 中可以看出,在 CASME II 数据集上,ViLDA 模型在各类别的分类上均表现出较高的准确率,尤其是惊讶类别的准确率达到 0.92,因为 ViLDA 模型能更好捕捉到嘴角和下巴的细微运动。在 SAMM 数据集上结果如表 4 所示,在愤怒类别上表现出较好的识别性能。但是,蔑视类别容易被误分类为愤怒和惊讶、高兴容易被误分类为愤怒,这是因为其肌肉运动幅度过小,导致运动容易与其他类别被混淆。在 SMIC 数据集上结果如表 5 所示,模型在积极样本上表现出较好的识别性能。但惊讶样本容易被预测为消极样本。这是因为消极样本是一个复合类别,由蔑视、压抑、厌恶等多种微表情组成,导致惊讶样本容易与其他微表情样本混淆。

2.6 消融实验

为验证各个模块的有效性,本研究在 3 个数据集上进行了消融实验,结果如表 6 所示。可以看出,在骨干网络的基础上添加 WTTE 模块,提取静态特征并与骨干网络提取的动态特征融合后,模型在 SMIC、CASME II 和 SAMM 数据集上的准确率分别提升了 0.050 6、0.032 9 和 0.032 3,F1 分数分别提高了

表 3 CASME II 数据集的混淆矩阵

Table 3 Confusion matrix of the CASME II dataset

真实 标签	预测标签				
	高兴	压抑	厌恶	惊讶	其他
高兴	0.690	0.031	0.031	0	0.250
压抑	0.037	0.630	0	0	0.330
厌恶	0	0	0.830	0	0.170
惊讶	0.040	0	0	0.920	0.040
其他	0.010	0.010	0.081	0.010	0.890

表 4 SAMM 数据集的混淆矩阵

Table 4 Confusion matrix of the SAMM dataset

真实 标签	预测标签				
	愤怒	蔑视	高兴	惊讶	其他
愤怒	0.930	0	0	0.035	0.035
蔑视	0.270	0.270	0.091	0.270	0.091
高兴	0.460	0	0.460	0.038	0.038
惊讶	0.330	0	0	0.600	0.067
其他	0.230	0	0.038	0.038	0.690

0.005 9、0.018 3 和 0.060 1,表明融合微表情的动态特征和静态特征能够有效地提高识别性能。在骨干网络的基础上添加 MESCA 模块后,模型在 SMIC、CASME II 和 SAMM 数据集上的准确率分别提升了 0.060 4、0.034 7 和 0.044 1,F1 分数分别提高了 0.039 1、0.025 3 和 0.077 0,表明该模块有助于模型关注到细微的肌肉运动特征,从而提高微表情识别性能;在骨干网络的基础上同时添加 WTTE 和 MESCA

模块后,模型在 SMIC、CASME II 和 SAMM 数据集上的准确率分别提升了 0.078 7、0.040 6 和 0.667,F1 分数分别提高了 0.072 0、0.055 7 和 0.089 9。这些结果充分证明了 WTTE 和 MESCA 模块对微表情识别任务的有效性,均能提高识别性能,并且微表情动态特征和静态特征融合后,能够进一步增强模型性能。

为探究不同的特征融合策略对模型性能的影响,本研究在 3 个微表情数据集上分别进行了 3 组实验,结果如表 7 所示。可以看出,采用特征相加的融合策略时,尽管在数据集 CASME II 上的识别准确率并非最高,但模型的 F1 分数在所有数据集上均最优。因此,选择相加作为特征融合方式是提升模型性能的可行方案。

表 5 SMIC 数据集的混淆矩阵

Table 5 Confusion matrix of the SMIC dataset

真实标签	预测标签		
	消极	积极	惊讶
消极	0.750	0.140	0.110
积极	0.140	0.800	0.059
惊讶	0.240	0.120	0.640

表 6 消融实验结果

Table 6 Results of ablation experiments

实验	模型			SMIC		CASME II		SAMM	
	Vision-LSTM	MESCA	WTTE	准确率	F1 分数	准确率	F1 分数	准确率	F1 分数
A	✓	—	—	0.659 1	0.659 0	0.780 5	0.765 0	0.637 0	0.534 8
B	✓	—	✓	0.709 7	0.664 9	0.813 4	0.783 3	0.669 3	0.594 9
C	✓	✓	—	0.719 5	0.698 1	0.815 2	0.790 3	0.681 1	0.611 8
D	✓	✓	✓	0.737 8	0.731 0	0.821 1	0.820 7	0.703 7	0.624 7

表 7 不同融合策略的实验结果

Table 7 Experimental results of different fusion strategies

融合策略	SMIC		CASME II		SAMM	
	准确率	F1 分数	准确率	F1 分数	准确率	F1 分数
拼接	0.715 6	0.704 8	0.831 8	0.796 6	0.678 6	0.598 6
相乘	0.707 3	0.695 9	0.792 7	0.734 6	0.681 5	0.579 3
相加	0.737 8	0.731 0	0.821 1	0.820 7	0.703 7	0.624 7

2.7 模型复杂度对比分析

本节对提出的 ViLDA 的计算量 (FLOPs) 和参数量进行对比分析。为了确保实验的公平性和结果的可比性,本研究选择 CASME II 数据集进行对比分析。表 8 比较了 ViLDA 和其他微表情识别模型的计算量、参数量以及准确率。与 SLSTT^[7]、AMAN^[8] 和 AdaMER^[22] 相比, ViLDA 的 FLOPs 分别减少 96.3%、95.8% 和 43.3%,

参数量分别减少 92.6%、19.2% 和 81.2%,但识别准确率却分别提高 6.3%、6.71% 和 4.84%,表明本研究提出的 ViLDA 能够减少推理时间和内存需求,并且能提高识别准确率,更适合于实际应用。

表 8 微表情识别模型复杂度对比

Table 8 Model complexity comparison of micro-expression recognition

模型	FLOPs/M	参数量/M	准确率/%
SLSTT ^[7]	185 566.86	92.61	75.81
AMAN ^[8]	162 981.62	8.39	75.40
AdaMER ^[22]	121 02.50	36.10	77.27
ViLDA	685 6.89	6.78	82.11

3 结论

本研究提出一种基于视觉长短期记忆网络和双重注意力机制的微表情识别模型 (ViLDA)。通过引入 Vision-LSTM 网络,利用其交替遍历机制和矩阵记忆结构,提取更具信息量的特征,捕捉全局上下文信息,学习不同面部区域之间的依赖关系。同时,结合设计的中值增强的空间通道注意力模块 (MESCA),以不同尺度捕捉和融合特征,关注细微肌肉运动特征的同时去除噪声。此外,还使用设计的基于小波变换的纹理增强模块 (WTTE) 对微表情序列的顶点帧进行增强,并提取增强后的面部静态特征。最后,融合静态特征和细微肌肉运动特征,进一步提高模型表征能力,并在 3 个公用数据集上取得了较好的识别准确率和 F1 分数。未来的工作考虑将宏表情的先验知识用于微表情识别,以进一步提高模型的泛化能力。

参考文献:

- [1] BEN X Y, REN Y, ZHANG J P, et al. Video-based facial micro-expression analysis: A survey of datasets, features, and algorithms[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5826-5846.
- [2] ZENG X M, ZHAO X C, ZHONG X Y, et al. A survey of micro-expression recognition methods based on LBP, optical flow and deep learning[J]. Neural Processing Letters, 2023, 55(5): 5995-6026.
- [3] 杨淑琪, 李哲, 刘国强, 等. 深层特征引导的多尺度上下文聚合图像变化检测网络[J]. 山东科技大学学报(自然科学版), 2024, 43(4): 121-128.
YANG Shuqi, LI Zhe, LIU Guoqiang, et al. Deep feature-guided multi-scale context-aggregated image change detection network[J]. Journal of Shandong University of Science and Technology (Natural Science), 2024, 43(4): 121-128.
- [4] LI Y T, WEI J S, LIU Y, et al. Deep learning for micro-expression recognition: A survey[J]. IEEE Transactions on Affective Computing, 2022, 13(4): 2028-2046.
- [5] XIA Z Q, PENG W, KHOR H Q, et al. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 8590-8605.
- [6] LEI L, CHEN T, LI S G, et al. Micro-expression recognition based on facial graph representation learning and facial action unit fusion[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2021: 1571-1580.
- [7] ZHANG L F, HONG X P, ARANDJELOVIC 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.
- [8] WEI M T, ZHENG W M, ZONG Y, et al. A novel micro-expression recognition approach using attention-based magnification adaptive networks[C]//International Conference on Acoustics Speech and Signal Processing. Piscataway: IEEE, 2022: 2420-2424.
- [9] BAO Y T, WU C X, ZHANG P, et al. Boosting micro-expression recognition via self-expression reconstruction and memory contrastive learning[J]. IEEE Transactions on Affective Computing, 2024, 15(4): 2083-2096.
- [10] WANG Y, ZHANG Q Y, SHU X. Micro-expression recognition using a multi-scale feature extraction network with attention mechanisms[J]. Signal Image and Video Processing, 2024, 18(6): 5137-5147.

- [11] ALKIN B, BECK M, POPPEL K, et al. Vision-LSTM: xLSTM as generic vision backbone[C/OL]//The 13th International Conference on Learning Representations, San Diego, 2025. DOI:10.48550/arXiv.2406.04303.
- [12] ZHAO S R, TANG H Y, MAO X L, et al. DFME: A new benchmark for dynamic facial micro-expression recognition[J]. IEEE Transactions on Affective Computing, 2024, 15(3):1371-1386.
- [13] DAVISON A K, LANSLEY C, COSTEN N, et al. SAMM: A spontaneous micro-facial movement dataset[J]. IEEE Transactions on Affective Computing, 2018, 9(1):116-129.
- [14] LI X B, PFISTER T, HUANG X H, et al. A spontaneous micro-expression database: inducement, collection and baseline[C/OL]//The 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Piscataway: IEEE, 2013. DOI:10.1109/FG.2013.6553717.
- [15] FENG W J, XU M L, CHEN Y X, et al. Nonlinear deep subspace network for micro-expression recognition[C/OL]//3rd Workshop on Facial Micro-Expression: Advanced Techniques for Multi-Modal Facial Expression Analysis, New York: Association for Computing Machinery, 2023. DOI:10.1145/3607829.3616444.
- [16] ZHAI Z J, ZHAO J H, LONG C J, et al. Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Piscataway: IEEE, 2023:22086-22095.
- [17] WANG T H, SHANG L. Temporal augmented contrastive learning for micro-expression recognition[J]. Pattern Recognition Letters, 2023, 167:122-131.
- [18] GONG W J, ZHANG Y, WANG W, et al. Meta-MMFNet: Meta-learning-based multi-model fusion network for micro-expression recognition[J]. Association for Computing Machinery, 2023, 20(2):1551-6857.
- [19] WANG S Q, LONG F, YAO J F. Micro-expression Recognition Based on PCB-PCANet+[C]//Neural Information Processing, Singapore: Springer, 2024:179-190.
- [20] WEI J S, PENG W, LU G M, et al. Geometric graph representation with learnable graph structure and adaptive AU constraint for micro-expression recognition[J]. IEEE Transactions on Affective Computing, 2024, 15(3):1343-1357.
- [21] DING X M, LI Y Y, WU Y L, et al. Enhancement-suppression driven lightweight fine-grained micro-expression recognition[J/OL]. Journal of Visual Communication and Image Representation, 2025, 107. DOI:10.1016/j.jvcir.2024.104383.
- [22] 包永堂, 武晨曦, 张鹏, 等. 自适应光流估计驱动的微表情识别[J]. 中国图象图形学报, 2024, 29(10):3060-3073.
BAO Yongtang, WU Chenxi, ZHANG Peng, et al. Adaptive optical flow estimation-driven micro-expression recognition[J]. Journal of Image and Graphics, 2024, 29(10):3060-3073.

(责任编辑:傅 游)