Research on Depression Screening Method Based on Large Language Model Combined with Visual Sensing*

ZHANG Siyong[1], HE Tianxiang[1], ZHANG Jiyuan[1], XU Ruiji[1], MAO Keji[1], QIAN Xiaofeng[2*]
(1.College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou Zhejiang 310023, China;
2.Zhijiang College, Zhejiang University of Technology, Shaoxing Zhejiang 312030, China)

Abstract: Currently, depression screening faces dual challenges, i.e., insufficient computing power of visual sensing devices, which is lead to weak spatio-temporal feature extraction capabilities, and the difficulty in deploying complex models, resulting in the absence of key spatio-temporal features for screening and thereby affecting the accuracy of existing solutions. Based on this, a depression screening method is propose, which integrates large language models with visual sensing. Facial expression sequences is acquired through visual sensing devices, large language models are introduced to generate structured text semantic features of facial expressions, video spatio-temporal features are combined to perform cross-attention alignment, and a spatio-temporal semantic-aware weight matrix is constructed. A dual-branch collaborative fusion strategy is designed to achieve feature weight allocation through weighted fusion and channel integration. Finally, residual learning is utilitied to enhance the network's expression ability. On the AVEC2013 and 2014 datasets, the RMSE and MAE reach 6.92/5.45 and 6.89/5.14 respectively, effectively enhancing the dynamic feature capture ability and fusion effect, providing a high robust visual sensing solution for depression screening.

Key words: visual sensing; depression screening; spatio-temporal semantics; feature fusion; large language models
EEACC: 7230 **doi**: 10.3969/j.issn.1004-1699.250174

大语言模型结合视觉传感的抑郁症筛查方法研究*

张嗣勇[1], 何天祥[1], 张继远[1], 徐瑞吉[1], 毛科技[1], 钱晓峰[2*]
(1.浙江工业大学计算机科学与技术学院, 浙江 杭州 310032; 2.浙江工业大学之江学院, 浙江 绍兴 312030)

摘 要: 目前, 抑郁症筛查面临视觉传感设备算力不足导致时空特征提取能力弱与复杂模型难以部署的双重挑战, 导致筛查关键要素时空特征的缺失, 进而影响现有方案的准确率。基于此, 提出一种融合大语言模型与视觉传感的抑郁症筛查方法, 通过视觉传感设备获取人脸表情序列, 引入大语言模型生成人脸表情结构化文本语义特征, 结合视频时空特征进行交叉注意力对齐, 构建时空语义敏感权重矩阵, 设计双分支协同融合策略, 通过加权融合与通道整合实现特征权重分配, 最后结合残差学习增强网络表达能力。在 AVEC2013 和 2014 数据集上 RMSE 和 MAE 分别达 6.92/5.45 与 6.89/5.14, 有效增强了动态特征捕捉能力和融合效果, 为抑郁症筛查提供了高鲁棒性的视觉传感解决方案。

关键词: 视觉传感; 抑郁症筛查; 时空语义; 特征融合; 大语言模型
中图分类号: TN393 **文献标识码**: A **文章编号**: 1004-1699(2026)09-2076-07

抑郁症是一个在全球范围内普遍存在的精神健康难题, 其极高的发病率给个人、家庭甚至整个社会都带来了极为深远的负面影响[1]。世界卫生组织的最新报告显示, 抑郁症已经成为全球最为高发的精神障碍之一, 影响着超过三亿人的生活[2]。而且相关研究表明, 早期进行筛查和诊断治疗能够帮助抑郁症患者有效控制病情的发展, 从而避免患者出现自残、自杀等危险行为[3]。

近年来, 计算机视觉技术凭借其客观性与高效性, 成为抑郁症辅助诊断领域的研究热点。Uddin 等[4]提出一种基于双流深度时空网络的自动化识别方法, 空间分支提取面部静态特征, 时间分支利用体积局部方向数(Volume Local Directional Number, VLDN)提取动态特征描述符, 再通过构建的多层双向长短期记忆网络, 引入时间中值池化对时空特征的时间片段进行中值聚合来实现抑郁症严重程度的量化评估。Melo 等[5]提出最大化与差分网络, 通过非 3D 卷积架构有效表征抑郁评估相关的面部表情变化。该网络通过最大化块捕捉平滑面部动态, 结合差分块编码突发表情变化, 实现多尺度时间信息

项目来源: 浙江省教育厅一般科研项目(Y202455757); 国家自然科学基金项目(62072410)
收稿日期: 2025-04-07 **修改日期**: 2025-04-24

融合。Niu 等[6]提出 Depressioner 模型,通过分析面部动态特征预测抑郁等级。利用图卷积嵌入模块实现多通道特征优势互补,并利用多尺度向量化模块融合扩张卷积与注意力机制,对各通道进行多尺度时间维度特征提取,沿通道维度聚合生成抑郁相关特征的向量化表示。Shang 等[7]提出融合面部局部四元数特征与全局深度卷积神经网络的抑郁症识别模型,首次将四元数编码引入面部抑郁识别领域。该模型通过四元数编码算法保留颜色通道相关性并提取局部纹理特征,构建局部深度特征与四元数特征融合模块,同时采用深度卷积神经网络提取全局特征,最终通过特征融合实现抑郁评分预测。于明等[8]提出基于残差注意力网络与金字塔扩张 ConvLSTM 的抑郁症识别算法,通过残差注意力网络提取加权空间特征,结合金字塔扩张策略构建 ConvLSTM 模块,捕捉多尺度时序动态变化。该方法通过时空特征融合机制,有效挖掘人脸图像在时间序列上的表情变化规律,最终通过深度神经网络实现抑郁量表得分的回归预测。Pan 等[9]提出时空注意力抑郁识别网络,通过捕捉全局与局部时空信息提升特征提取能力及抑郁识别相关性。通过时空注意力机制生成时空注意力向量以捕捉特征的全局与局部时空关系,并在时空注意力机制中引入注意力向量级融合策略,实现时空域信息的有效整合。

现有研究对于抑郁症筛查虽能够较好的捕捉面部特征,但受限于视频单模态分析的局限性,仅分析静态图像或单帧视频难以表征表情的时间动态变化。针对上述问题,本文提出一种融合大语言模型与视觉传感的抑郁症筛查方法。如图 1 所示,首先利用大语言模型生成人脸表情的结构化描述文本,提取文本语义特征;其次将视频时空特征与文本语义特征对齐校准,通过交叉注意力机制生成时空语义敏感的权重矩阵;随后采用双分支协同融合策略,分别通过加权融合分支和通道整合分支实现特征的权重分配与维度压缩;最终通过残差学习模块增强网络表达能力。

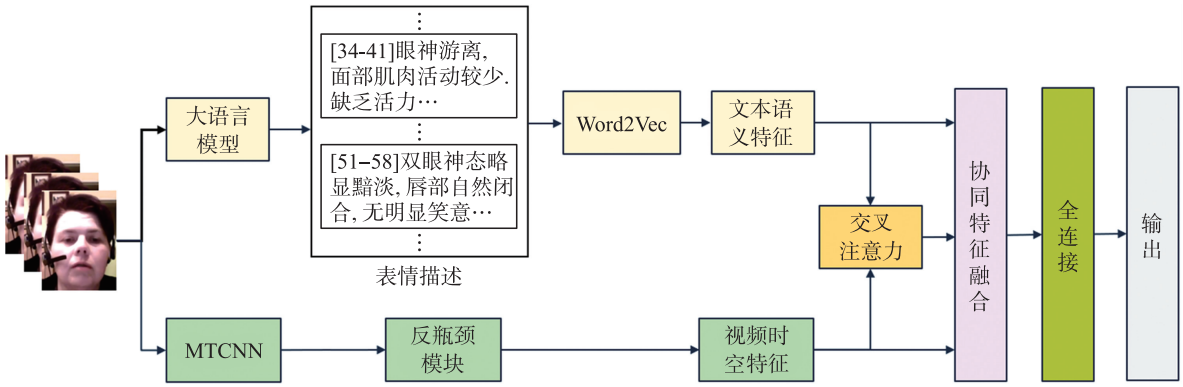


图 1　STSCNet 网络结构

1　基于边缘设备的时空语义网络

1.1　视频特征

用 MTCNN[10]算法提取人脸关键点信息,随后通过特征提取模块获取人脸特征。考虑到抑郁症识别对人脸特征的要求,特征提取过程需更加注重对人脸高维特征的捕捉与分析。现有的 ResNet[11]架构通常采用瓶颈结构,即利用一个卷积核先压缩通道数再扩展,以此降低计算复杂度,同时保留关键特征信息。在处理高维数据时,这种结构确实展现出一定优势。然而,抑郁症筛查任务重,这种瓶颈结构可能并不完全适用。因为它更多地依赖于 RGB 图像的低维特征,而在人脸特征提取的情况下,这些低维特征的可用性会大幅降低。

相比之下,如图 2 所示,Liu 等[12]提出的反瓶颈结构更适合处理此类问题。它包含 3 个卷积核,其中第一个 1×1 卷积核对通道数进行扩展,中间 3×3 卷积核负责提取特征,最后一个 1×1 卷积核则对通道数进行压缩。在特征提取中引入反瓶颈结构,能够在抑郁症筛查任务中更好地提取人脸特征,尤其是在面部信息部分缺失的情况下,有助于网络在人脸存在遮挡的情况下捕捉更多的上下文和细节信息。

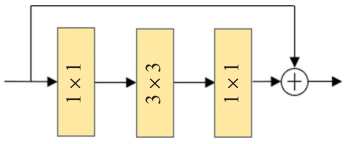


图 2　反瓶颈结构

1.2　语义特征

为进一步模拟临床医生的诊断偏好,本小节引入了语义引导的时空注意力模块。该模块构建的基础,是大语言模型所生成的针对视频中人脸表情的

客观描述文本，图 3 展示了 Phi3[13] 生成的人脸视频描述文本，这些文本带有时间标签和表情状态描述，格式为[起始时间-结束时间]，例如“[1-3]，表情正常；[4-7]，出现轻微皱眉”，能够直观展现出人脸表情随时间维度的动态变化，让后续模型更好地学习抑郁相关的特征。

时间戳	人脸描述
337.500-340.800	眉头轻微收缩，双眼上视，眼球转动频率低，颧肌松驰，苹果肌位置略下移，嘴唇闭合时嘴角轻微下倾。
355.300-358.600	眉头收缩持续且深度增加，眼睑下垂导致眼裂变窄，眨眼频率降低，面部肌肉松驰感加剧，苹果肌位置进一步下移，嘴唇微张。
360.400-363.700	眉头持续收缩，皱纹明显，眼睑下垂状态更显著，眼球上视但焦距分散，面部肌肉保持松驰，苹果肌下移趋势延续，嘴唇微闭。
378.200-381.500	前额皱纹收缩，眼睑下垂使眼裂更窄，眼球凝视上方，面部肌肉松驰感加剧，苹果肌位置进一步下移，嘴唇微张。
386.400-394.200	眉头轻微收缩，前额现细微皱纹，双眼半睁，眼睑松驰，眼球转动幅度小，苹果肌位置上移，嘴唇呈微张状态，面部肌肉整体松驰。
396.200-301.600	眉头持续轻缩，前额皱纹加深，双眼睁开但目光涣散，眼睑松驰，嘴唇保持微张，面部肌肉无大幅动作，延续松驰状态。
402.100-405.600	眉头轻微收缩，双眼睁开，眼睑状态自然，眼球无明显转动，目光呈平视状态：嘴唇微张，面部肌肉整体处于放松状态。
408.200-411.500	眉头保持轻微收缩趋势，前额无明显皱纹：双眼维持睁开状态，眼球转动频率低：嘴唇持续微张，面部肌肉松驰感贯穿。
412.100-415.400	眉头轻微收缩状态延续，双眼平视，眼睑无明显紧绷或下垂：嘴唇保持微张形态，面部肌肉无显著紧绷或舒展动作。
416.300-419.600	眉头保持舒展，双眼平视，眼睑自然：嘴唇微张状态下，出现短暂抿唇动作，面部肌肉伴随细微收缩与放松的交替变化。

图 3 大语言模型生成的视频人脸描述文本

在得到文本内容后，需要将文本转换为词特征向量，Chen 等[14] 研究表明，运用预训练的词向量模型有助于提升文本分类的准确程度，同时能够降低词向量的维度。例如 Word2Vec[15]、GloVe[16]、BERT[17] 等都是常见的词向量模型。考虑到 BERT 参数量较大、计算复杂，在边缘设备或计算资源有限时难以高效运行，Glove 虽然在语义表示上有优势，但在捕捉动态变化方面弱于 Word2Vec，而抑郁症检测需对表情语义的动态变化敏感，因此本章选取 Word2Vec 作为词向量的预训练模型。

1.3 时空语义交叉注意力

在处理视频和语义信息时，视频数据具有丰富的时空信息，包含的是不同时间帧画面的空间特征，语义信息则是对视频的客观描述，二者在形式上存在差异，直接融合难以实现信息有效融合，因此本章引入交叉注意力将时空特征和语义特征映射到同一交互空间，再计算注意力权重。对于时空压缩特征 V_c 和扩展后的语义特征 S' 通过两个不同的 1×1×1 卷积 Conv_q 和 Conv_k，分别做特征投影得到查询投影 Q_v 和键投影 K_s，如式(1)和式(2)所示：

$$Q_v=\mathrm{Conv}_q(V_c) \tag{1}$$

$$K_s=\mathrm{Conv}_k(S') \tag{2}$$

得到查询投影 Q_v 和键投影 K_s 后，进行注意力权重计算，得到权重注意力矩阵 **Attn**，将查询投影 Q_v 和键投影 K_s 做逐元素相乘，为了防止逐元素相乘数值过大导致梯度消失，引入缩放因子 $\sqrt{D}$ 对相乘后的结果进行缩放，最后通过 Sigmoid 函数做归一化处理，如式(3)所示：

$$\mathbf{Attn}=\sigma\left(\frac{Q_v\odot K_s}{\sqrt{D}}\right) \tag{3}$$

1.4 协同式特征融合

得到注意力权重矩阵后，需要将视频特征和语义特征做特征融合，简单的特征融合如拼接或相乘，虽然能够快速整合不同的特征信息，但是在处理视频和语义这种非同源的特征时，简单拼接就显现出局限性，无法有效地建立起二者之间的联系。因此本文采用协同式特征融合，整体结构如图 4 所示，协同式特征融合包含加权融合分支和通道整合分支，

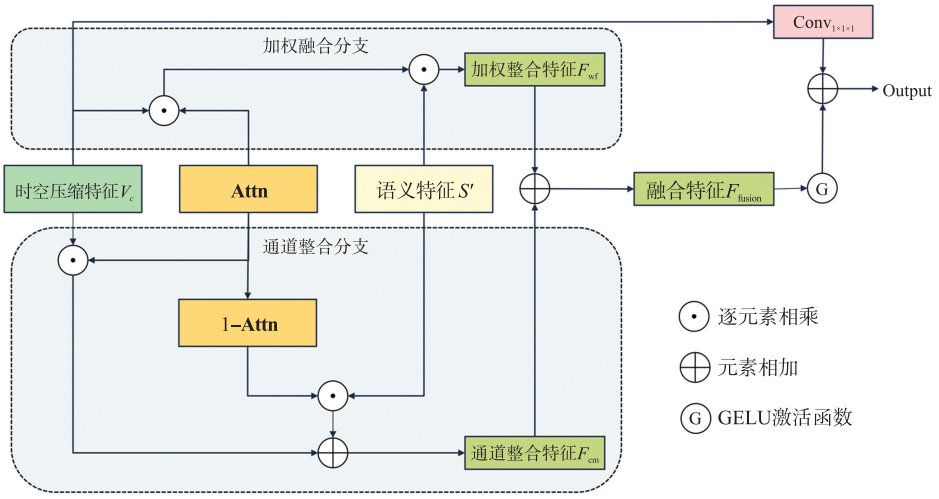


图 4 协同式特征融合结构

加权融合分支通过强化局部视频特征与语义的关联,利用交叉注意力实现特征交互,让模型聚焦于重要的局部信息;通道整合分支则在全局层面上保留高注意力区域视频特征的同时,在低注意力区域补充语义信息,实现了非同源特征的深度融合与信息互补。

加权融合分支将时空压缩特征 V_c、权重注意力矩阵 **Attn** 和语义特征 S'逐元素相乘,得到加权融合后的结果 F_{wf},如式(4)所示:

$$F_{\mathrm{wf}} = V_c \odot \mathbf{Attn} \odot S' \tag{4}$$

通道整合分支中,先分别进行两个操作:一是将时空压缩特征 V_c 与权重注意力矩阵 **Attn** 逐元素相乘;二是把语义特征 S'与反注意力权重矩阵 **1−Attn** 逐元素。然后将这两个相乘的结果拼接在一起得到通道整合后的结果 F_{cm},如式(5)所示:

$$F_{\mathrm{cm}} = \mathrm{Concat}[V_c \odot \mathbf{Attn}, S' \odot (1-\mathbf{Attn})] \tag{5}$$

将两个分支得到的结果拼接得到最终协同式特征融合后的结果 F_{fusion},如式(6)所示:

$$F_{\mathrm{fusion}} = \mathrm{Concat}(F_{\mathrm{wf}}, F_{\mathrm{cm}}) \tag{6}$$

为进一步提升特征质量,在得到 F_{fusion}后,引入残差增强模块。通过 1×1×1 的卷积 $\mathrm{Conv}_{\mathrm{res}}$对原始时空压缩特征 V_c 降维得到 V_{res},再将双路径特征融合的结果 F_{fusion}经 3×3×3 卷积 Conv_f 后,送入激活函数 GELU,最后再和 V_{res} 相加得到最终输出 V_{out},如式(7)所示:

$$V_{\mathrm{out}} = \mathrm{Conv}_{\mathrm{res}}(V_c) + \mathrm{GELU}[\mathrm{Conv}_f(F_{\mathrm{fusion}})] \tag{7}$$

2　实验评估

2.1　实验设置

针对视频采集场景下的隐私保护与实时性需求,本文采用一种边缘−云端协同的人脸特征提取方法。边缘设备优先执行本地原始数据处理,完成人脸特征的前层计算,通过张量形式加密传输中间结果至云端。该设计在保证数据隐私性的同时,有效缓解边缘算力瓶颈问题。

部署设备本文选择 NVIDIA Jetson NX 作为边缘设备,其集成了 6 核 ARM 处理器和 384 个 NVIDIA CUDA 核心,具备强大的并行处理能力。实验中采用 1080P 工业摄像头进行视频数据采集,以上设备如图 5 所示。云端服务器配备 Intel(R) Core(TM) i7-12700KF 处理器、GTX4090Ti 显卡和 128 GB 内存。

2.2　参数设置与评价指标

模型框架使用 PyTorch,在 RTX 4090Ti 显卡上完成训练。训练时采用 AdamW 优化器,共进行 250 个 epoch 训练。学习率设置为 1e−4,最小学习率为 1e−6。权重衰减值设为 0.05,batchsize 设置为 32。

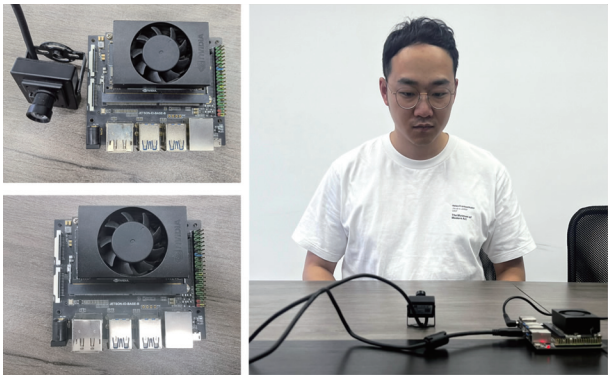


图 5　Jeston NX,摄像头以及实验场景设置

在 AVEC2013[18] 和 AVEC2014[19] 中选用均方根误差(Root Mean Square Error,RMSE)和平均绝对误差(Mean Absolute Error,MAE)作为评估指标。RMSE 和 MAE 的计算公式如式(8)和式(9)所示:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{L=1}^{N}(a_L - \hat{a}_L)^2} \tag{8}$$

$$\mathrm{MAE} = \sqrt{\frac{1}{N}\sum_{L=1}^{N}|a_L - \hat{a}_L|} \tag{9}$$

式中:N 表示测试样本数量,a_L 表示测试样本标签值,$\hat{a}_L$ 代表测试样本识别值。

此外,引入皮尔逊相关系数(Pearson Correlation Coefficient,PCC)衡量模型在测试集上真实值与预测值之间的线性相关程度,计算公式为如式(10)所示:

$$\mathrm{PCC} = \frac{\mu_{y\hat{y}} - \mu_y \times \mu_{\hat{y}}}{\sqrt{\mu_{y^2} - (\mu_y)^2} \times \sqrt{\mu_{\hat{y}^2} - (\mu_{\hat{y}})^2}} \tag{10}$$

式中:$\mu_{y\hat{y}}$表示真实值 y 与预测值 $\hat{y}$ 乘积的均值,μ_y 表示真实值 y 的均值,$\mu_{\hat{y}}$表示预测值 $\hat{y}$ 的均值,μ_{y^2}表示真实值 y 平方后取值的均值,$\mu_{\hat{y}^2}$表示预测值 $\hat{y}$ 平方后取值的均值。

2.3　消融实验

由表 1 可知,Phi3 在 AVEC2013/2014 数据集上的 RMSE 为 5.28/5.57,综合推理延迟和内存占用最优。其优势源于大参数量支持下的精细化人脸表情分析能力,可精准捕捉微表情强度变化等细节特征,生成与视频内容高度契合的语义特征,从而有效降低 STSCNet 预测误差。LLaMA3.2 在 AVEC2013/2014 的 RMSE 为 5.62/5.98,虽数值较高但其推理延迟仅 85 ms/frame 且内存占用为 2.1 GB,展现出边缘设备部署的优势。Gemma3 在 AVEC2013/2014 的 RMSE 为 5.35/5.62,推理延迟 105 ms/frame、内存

占用 3.2 GB,综合性能处于中间水平。

表 2 给出了不同注意力机制对最后识别结果的影响。数据显示,交叉注意力机制性能显著优于其他机制:在 AVEC2013 数据集上,其 RMSE 较自注意力、通道注意力、多头注意力分别降低 0.92、1.4、1.06,MAE 分别降低 0.71、0.98、0.75;在 AVEC2014 数据集上,RMSE 分别降低 0.79、1.47、2.29,MAE 分别降低 0.15、0.28、1.67。这是由于交叉注意力机制通过明确建立视频和文本的跨模态关联,提升了模型处理多模态特征的能力。而自注意力机制只能处理单模态序列特征关联,缺少跨模态交互设计,导致预测误差累积。通道注意力仅优化单模态通道,无法实现跨模态深层信息融合,难以整合视频与音频的互补价值。多头注意力虽尝试多视角提取特征,但因融合复杂度高且缺乏跨模态对齐机制,在 AVEC2014 这种复杂数据集中易导致特征混乱,放大预测误差。

表 1 不同大语言模型的消融实验结果表

模型名称	参数量	推理延迟/(ms/frame)	内存占用/GB	RMSE	
				AVEC2013	AVEC2014
LLaMA3.2[20]	3 B	85	2.1	5.62	5.98
Phi3[13]	**3.8 B**	**138**	**4.8**	**5.28**	**5.57**
Gemma3[21]	4 B	105	3.2	5.35	5.62

表 2 不同注意力机制的消融实验结果表

方法	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
自注意力	7.85	6.16	7.30	5.95
通道注意力	8.32	6.38	7.98	6.08
多头注意力	7.98	6.20	8.80	7.47
交叉注意力	**6.92**	**5.45**	**6.51**	**5.80**

2.4 性能对比

为了进一步验证 STSCNet 模型的有效性,本小节在 AVEC2013 和 AVEC2014 数据集上分别进行了测试,并与现有方法进行了比较,如表 3 与表 4 所示。

表 3 STSCNet 与其他先进模型在 AVEC2013 数据集的实验对照

方法	模态	AVEC2013	
		RMSE	MAE
Uddin 等[4]	V	8.93	7.04
Melo 等[5]	V	7.55	6.24
Niu 等[6]	V	8.13	6.28
Shang 等[7]	V	8.20	6.38
于明等[8]	V	8.85	6.79
Pan 等[9]	V	7.26	5.97
Niu 等[22]	V+A	8.16	6.14
Meng 等[23]	V+A	10.96	8.72
Pan 等[9]	V+A	7.99	6.23
STSCNet	V+T	**6.92**	**5.45**

Niu 等[22]提出基于时空注意力网络与多模态特征融合的抑郁症预测方法,通过整合视听模态时空动态信息构建预测模型。该方法将语音幅度谱与视频分割为固定长度片段,采用特征池化聚合模态片段级特征,生成包含模态互补信息的多模态表征,最终通过支持向量回归模型实现抑郁程度预测。

Meng 等[23]提出一种融合视听模态的抑郁症程度自动预测方法,利用运动历史直方图分别从视频和音频数据中提取面部微表情与语音语调的动态特征,捕捉抑郁状态下非言语行为的细微变化。针对单模态分析,采用偏最小二乘回归算法建立动态特征与抑郁量表的映射关系;在决策层融合视觉与听觉线索的预测结果,形成最终抑郁评估结果。需要说明的是,该研究仅基于 AVEC2013 数据展开,未涉及对 AVEC2014 数据的研究。

表 4 STSCNet 与其他先进模型在 AVEC 2014 数据集的实验对照

方法	模态	AVEC2014	
		RMSE	MAE
Uddin 等[4]	V	8.78	6.86
Melo 等[5]	V	7.65	6.06
Niu 等[6]	V	8.07	6.14
Shang 等[7]	V	7.84	6.08
于明等[8]	V	8.67	6.48
Pan 等[9]	V	7.30	5.99
Niu 等[22]	V+A	7.03	5.21
Meng 等[23]	V+A	–	–
Pan 等[9]	V+A	**6.83**	5.32
STSCNet	V+T	6.89	**5.14**

对于单一视频模态,STSCNet 在 AVEC2013 数据集上 RMSE 和 MAE 值均为最低,相比 Uddin 等[4]、Melo 等[5]、Niu 等[6]、Shang 等[7]、于明等[8]、Pan 等[9],RMSE 值分别低了 2.01、0.63、1.21、1.28、1.93、0.34,MAE 值分别低了 1.59、0.79、0.83、0.93、1.34、0.52。同样的在 AVEC2014 中,RMSE 和 MAE 均为最优。这是因为本章提出的 STSCNet 融合了人

脸表情的文本特征,使模型对于人脸细微表情捕捉更加敏感,弥补视频模态在语义理解上的不足。

对于多模态,STSCNet 在 AVEC2013 数据集上,RMSE 和 MAE 值均为最低,相比 Niu 等[22]、Meng 等[23]和 Pan 等[9] RMSE 值分别低了 1.24、4.04 和 1.07,MAE 值分别低了 0.69、3.27 和 0.78。而在 AVEC2014 中,STSCNet 延续 MAE 最优表现,仅 RMSE 以 0.06 的细微差距略高于 Pan 等[9]0.06。造成这一差异的原因,是因为 STSCNet 对于长时数据的处理能力更强,在复杂场景下稳定性更优,但在视频时长较短的 AVEC2014 中,对于短时数据的适配性略有不足,导致 RMSE 稍高。

图 6 和图 7 显示了不同数据集上的预测值和标签值分布情况。AVEC2013 包含 50 例测试样本,AVEC2014 包含 100 例测试样本。通过视频片段识别值均值计算最终预测结果,采用 PCC 评估预测值与真实标签的相关性。可视化结果显示,预测值与真实值在整体趋势上高度契合,尤其在 AVEC2014 中,随着样本量增至 100 例,预测值对真实值的拟合更为紧密,波动趋势几乎一致。PCC 指标在 AVEC2013 数据集上为 0.749,在 AVEC2014 数据集上为 0.771,表明 STSCNet 在不同规模数据集中均能有效捕捉抑郁特征,并在样本量更充足的 AVEC2014 中,STSCNet 对抑郁分数预测能力更强,验证了模型在数据集上的预测稳定性。

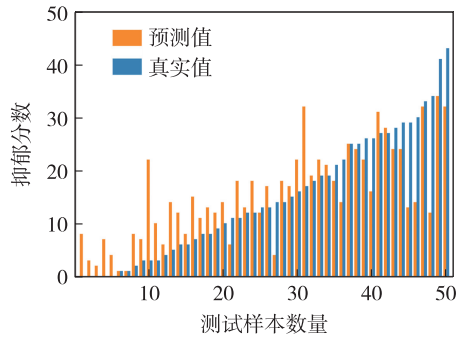


图 6　STSCNet 在 AVEC2013 数据集上的可视化对比

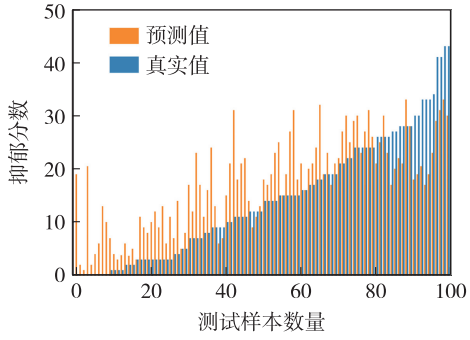


图 7　STSCNet 在 AVEC2014 数据集上的可视化对比

图 8 展示了模型在不同抑郁程度分类下的 PCC 结果对比。由图中数据可知,在 AVEC2013 上抑郁样本平均 PCC 值为 0.784,较非抑郁样本的 0.594 提升 32%,在 AVEC2014 上抑郁样本平均为 0.790,较非抑郁样本为 0.503 提升 57%。相较于非抑郁样本,模型对抑郁样本的识别性能更优。这是因为随着抑郁程度加重,视频时空特征和文本语义特征的关联性加强,有效提升了识别准确性。

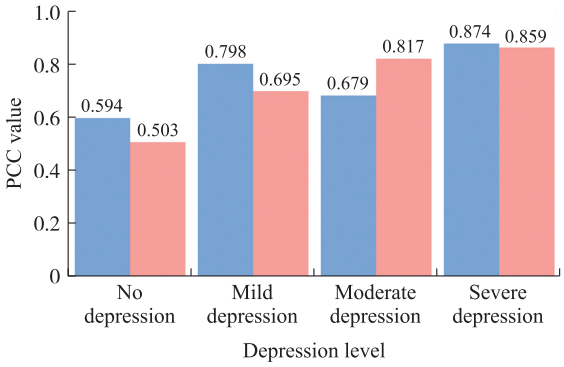


图 8　不同抑郁程度分类下 STSCNet 模型的 PCC 结果对比

3　总结

本文主要提出了一种融合大语言模型与视觉传感的抑郁症筛查方法。实验表明本文提出的方法在识别效果上优于现有方法,并有更好的工成部署可行性。但是在处理长视频时,本方法的处理效果和耗时还有优化空间,希望在后续工作中,对于长视频处理能力,进一步优化本文方法,并结合实际使用场景进行验证。

参考文献:

[1] World Health Organization. Depression and Other Common Mental Disorders:Global Health Estimates[R]. Geneva, Switzerland:World Health Organization,2017.

[2] 谢丽颖. 抑郁症诊断方法与相关干预研究综述[J]. 社会科学前沿,2024,13(1):1-7.

[3] Thapar A, Eyre O, Patel V, et al. Depression in Young People[J]. The Lancet,2022,400(10352):617-631.

[4] Uddin M A,Joolee J B,Lee Y K. Depression Level Prediction Using Deep Spatiotemporal Features and Multilayer Bi-LTSM[J]. IEEE Transactions on Affective Computing,2020,13(2):864-870.

[5] de Melo W C,Granger E,Lopez M B. MDN:A Deep Maximization-Differentiation Network for Spatio-Temporal Depression Detection[J]. IEEE Transactions on Affective Computing,2021,14(1):578-590.

[6] Niu M,He L,Li Y,et al. Depressioner:Facial Dynamic

Representation for Automatic Depression Level Prediction [J]. Expert Systems with Applications, 2022, 204: 117512.

[7] Shang Y, Pan Y, Jiang X, et al. LQGDNet: A Local Quaternion and Global Deep Network for Facial Depression Recognition[J]. IEEE Transactions on Affective Computing, 2021, 14(3): 2557-2563.

[8] 于明,徐心怡,师硕,等. 基于面部深度空时特征的抑郁症识别算法[J]. 电视技术,2020,44(11):12-18.

[9] Pan Y, Shang Y, Liu T, et al. Spatial-Temporal Attention Network for Depression Recognition from Facial Videos [J]. Expert Systems with Applications, 2024, 237: 121410.

[10] Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[11] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016: 770-778.

[12] Liu Z, Mao H Z, Wu C Y, et al. A ConvNet for the 2020s [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022: 11966-11976.

[13] Abdin M, Aneja J, Awadalla H, et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone[EB/OL]. (2024-08-30). https://arxiv.org/abs/2404.14219.

[14] Chen Y. Convolutional Neural Network for Sentence Classification[D]. Waterloo, ON, Canada: University of Waterloo, 2015.

[15] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality [C]//2013 27th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 2013: 3111-3119.

[16] Pennington J, Socher R, Manning C D. Glove: Global Vectors for Word Representation[C]//2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatqr, 2014: 1532-1543.

[17] Devlin J, Chang M W, Lee K, et al. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding[C]//2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2019: 4171-4186.

[18] Valstar M, Schuller B, Smith K, et al. Avec 2013: the Continuous Audio/Visual Emotion and Depression Recognition Challenge[C]//3rd ACM International Workshop on Audio/Visual Emotion Challenge, Barcelona, Spain, 2013: 3-10.

[19] Valstar M, Schuller B, Smith K, et al. Avec 2014: 3D Dimensional Affect and Depression Recognition Challenge [C]//4th International Workshop on Audio/Visual Emotion Challenge, Orlando, FL, USA, 2014: 3-10.

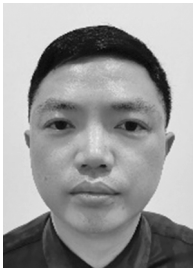
[20] Touvron H, Lavril T, Izacard G, et al. Llama: Open and Efficient Foundation Language Models[EB/OL]. (2023-02-27). https://arxiv.org/abs/2302.13971

[21] Mesnard T, Hardin C, Dadashi R, et al. Gemma: Open Models Based on Gemini Research and Technology [EB/OL]. (2024-04-16). https://arxiv.org/abs/2403.08295.

[22] Niu M, Tao J, Liu B, et al. Multimodal Spatiotemporal Representation for Automatic Depression Level Detection [J]. IEEE Transactions on Affective Computing, 2020, 14(1): 294-307.

[23] Meng H Y, Huang D, Wang H, et al. Depression Recognition Based on Dynamic Facial and Vocal Expression Features Using Partial Least Square Regression [C]//3rd ACM International Workshop on Audio/Visual Emotion Challenge, Barcelona, Spain, 2013: 21-30.

张嗣勇(1994—),男,浙江工业大学硕士研究生,研究方向为嵌入式操作系统和计算机视觉,221122120281@zjut.edu.cn;



钱晓峰(1978—),男,硕士,通讯作者,浙江工业大学之江学院助理研究员,中共党员,主要研究方向为人工智能应用研究,02662@zzjc.edu.cn。