第202274卷年1第01月0期 软件工程 SOFTWARE ENGINEERING Vol.O27c t.N 2o0.1204

文章编号:2096-1472(2024)10-0068-06 DOI:10.19644/j.cnki.issn2096-1472.2024.010.014 一种基于多模态特征增强网络的抑郁症检测方法

赵小明1,2, 范慧婷1, 张石清2

# (1.浙江理工大学信息科学与工程学院,浙江 杭州 310018; 2.台州学院智能信息处理研究所,浙江 台州 318000)

 tzxyzxm@163.com; courage\_f@163.com; tzczsq@163.com

摘 要:针对传统的多模态融合方法在抑郁症检测中忽略了模态之间的交互性、未能充分提取出更全面的特征表示的问题,本研究提出一种基于多模态特征增强网络的抑郁症检测方法,该方法有效地集成了视频、音频和远程光电容积脉搏(photoplethysmographic,rPPG)信号3种模态,通过模态间Transformer、模态内Transformer和多头

自注意力机制,共同学习输入模态序列每个时间步的模态内和模态间的动态关系,达到了特征增强的目的。最终, 拼接3个模态增强后的特征获得全面特征表示。在AVEC2013公共数据集上的实验结果显示,该方法的平均绝对误差为7.07,优于单模态抑郁症检测,表明该方法有效促进了模态之间的交互,并实现了特征增强,在自动抑郁症检测任务中展现出显著的有效性。

关键词:多模态;深度学习;抑郁症检测;卷积神经网络;特征增强;多模态融合中图分类号:TP391.41 文献标志码:A

ADepressionDetectionMethodBasedonMultimodalFeatureEnhancementNetwork

ZHAOXiaoming1,2,FANHuiting1,ZHANGShiqing2

(1.SchoolofInformationScienceandEngineering,ZhejiangSci-TechUniversity,Hangzhou310018,China; 2.InstituteofIntelligentInformationProcessing,TaizhouUniversity,Taizhou318000,China)  tzxyzxm@163com;courage\_f@163.com;tzczsq@163.com

Abstract:Traditi onalmu ltimodalfusion m ethodstend tooverloo kt heintera ctivit ybetween mod alities andfail to extract comprehensive feature representations in depression detection. To address these problems, this paper proposes a depressiondetectionmethodbasedonamultimodalfeatureenhancementnetwork,whicheffectivelyintegratesthreemodalities:video,audio,andremotephotoplethysmography(rPPG)signals.Byemployinginte-r modalTransformers,intra-modalTransformers,andamult-i headsel-f attentionmechanism,themethodlearnsthedynamicrelationshipsbothwithinandbetweenmodalitiesforeachtimestepoftheinputmodalitysequence,achievingfeatureenhancement.Ultimately,theenhancedfeaturesfromthethreemodalitiesareconcatenatedtoobtainacomprehensivefeaturerepresentation.ExperimentalresultsontheAVEC2013publicdatasetindicatethattheproposedmethodachievesanaverageabsoluteerrorof7.07,outperformingtraditionalunimodaldepressiondetectionmethods.Thisdemonstratesthattheproposedmethodeffectivelyfacilitatesinteractionbetweenmodalitiesandenhancesfeatures,showingsignificanteffectivenessinautomateddepressiondetectiontasks.

Keywords:multimodal;deeplearning;depressiondetection;convolutionalneuralnetwork;featureenhancement;

multimodalfusion

# 0引言(Introduction)

早期诊断抑郁症在促进治疗效果方面起着至关重要的作用。但是目前抑郁症的诊断依赖于主观行为,例如患者的自我

报告评估和临床判断症状严重程度,而这些因素容易受到环境因素的影响。

如何有效地进行自动多模态抑郁症检测,以辅助医生实现早期抑郁症的诊断,已成为当前一个既重要又具有挑战性的研究问题。因此,运用机器学习等技术进行抑郁症自动检测[1]的研究受到广大研究者的关注。然而,传统的融合方法通常直接采用简单的级联方式融合多模态特征,这种方式忽略了模态之间的交互性,无法充分提取出更全面的特征表示,从而影响了抑郁症的检测效果。

因此,本文探索了一种基于多模态特征增强网络的抑郁症检测方法,该方法融合了音频、视频及rPPG信号,其中rPPG 模态作为一种附加模态,增强了多模态抑郁症检测的效果,通过堆叠多个模态间和模态内Transformer,并配合多头自注意力机制,共同获取输入序列每个时间步的模态内和模态间的信息交互,以达到多模态特征增强的目的,从而提升抑郁症检测性能。

# 1相关研究(Relatedresearch)

目前,主流的抑郁症检测方法主要可以分为3类:基于视频的检测、基于音频的检测和基于多模态的检测。

抑郁症患者常常表现出面部表情的减少或呆滞,他们的面部表情可能缺乏生动度和情感表达。研究者通过机器学习分析面部特征在辅助诊断抑郁症方面取得了比较大的进展[2]。例如,孙浩浩等[3]基于人脸图像的全局和局部特征,构建了一种融合通道层注意力机制的多支路卷积网络模型。音频作为传达情感的媒介,抑郁症患者和非抑郁症患者之间的言语模式存在明显的差异[4]。MA等[5]提出了DepAudioNet深度模型, 结合卷积神经网络(CNN)和长短期记忆(LSTM),用于编码声道中的抑郁症相关特征,从而提供更全面的音频表示,取得了较好的检测效果。这些深度学习架构在提取有意义的音频或视频特征以及提高抑郁症检测的效果方面发挥着重要作用。然而,仅依赖音频或者视频特征可能会丢失测试对象的其他动态信息,从而限制了抑郁症检测的性能。

除了视频和音频模态,rPPG信号也可用于抑郁症检测。 rPPG信号使用非接触式光学技术测量和分析心率和血流量等生理信息。一些研究通过提取rPPG信号并计算统计特征和心率变异性(HRV)特征,探讨了抑郁症与HRV之间的关系[6-7]。这些特征随后被输入基于随机森林和多层感知机 (MultilayerPerceptron,MLP)的机器学习回归器中。这些发

现证明了基于rPPG的抑郁症检测方法的潜力。然而,很少有研究关注和探索用于抑郁症检测的rPPG信号。

除了上述单模态方法,通过多模态信息融合方法整合多种模态在提高抑郁症检测性能方面也显示出不错的效果。HE 等[8]通过特征层融合将提取的音频和视频特征串联成一个高维特征向量,并使用支持向量回归(SVR)进行抑郁症预测。然而,这种方法容易产生高维特征表示,从而导致维度灾难。 YANG等[9]将获得的音频视频结果和文本结果进行决策融合,以获得最终的抑郁症检测结果。但是,决策层融合单独考虑不同模态,无法捕捉它们之间的内在关系。更多的研究者通过模型层融合考虑模态之间的关系。NIU等[10]采用多模态注意力特征融合方法整合音频模态和视频模态。谷明轩等[11]结合了脑电信号和音频特征提出了基于全连接神经网络的多模态特征融合模型。但是,这些模型层融合方法在模态之间的交互性方面仍存在不足。近年来,Transformer[12]技术引起了广泛关注,Transformer模型中的编码器和解码器组件利用多头

自注[13]意提机出制了捕一捉种输将入语序言列之数外据的的信长息距融入离基上于下文Tra信ns息fo。rmILerIA的S 等模型,用于社交媒体中抑郁症和压力检测,这一方法展现出良好的应用前景。

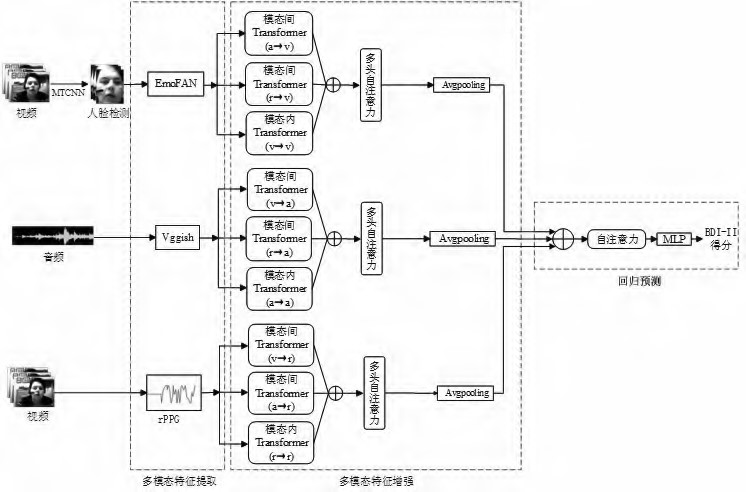
受到Transformer技术的优势和rPPG信号在抑郁症检测中潜力的启发,本研究提出一种基于多模态特征增强网络的抑郁症检测方法。首先,针对视频、音频和rPPG模态进行多模态特征提取;其次,通过基于Transformer的特征增强模块和多头自注意力机制,实现不同模态之间的交互;最后,利用多层

# 络的抑郁症检测方

## 法(Depressiondetectionmethodbasedon multimodalfeatureenhancementnetwork)

基于多模态特征增强网络的抑郁症检测方法的整体结构如图1所示,该结构主要包括多模态特征提取、多模态特征增强和和音回频归模预态测,本3文个采部用分深。度(1)多模[14态-15特]提征取提高取级:对视于频视和频音频模特态征。对于rPPG模态,采用短C时N端N到端rPPG估计框架[16]提取 rPPG信号值。(2)多模态特征增强:模态之间的Transformer

通过与其他模态之间进行信息交互,用于增强目标模态的特征。模态内Transformer聚焦于目标模态,对目标模态内部特征进行交互,关注到目标模态中最相关和有价值的信息。多头自注意力机制提取更丰富、更有用的特征,平均池化聚合目标模态特征。(3)回归预测:通过级联和自注意力机制处理增强后的特征并输入多层感知机网络进行最终的抑郁症预测。



### 图1 基 于多 模态特 征增强网 络 的抑郁症检 测方法的 整体结构

Fig.1The ov erallstructu reofde pressiondetec tionmethod basedonmultimodalfeatureenhancementnetwork

### 2.1多模态特征提取

帧任,务使对,用其于多次视任使频务用模级E态m联o,卷首FA积先N神从[14经每]预网个训络视练频样本中提[1取7]进10行0个人脸连续检的测

(M深T度CN卷N积)神经网络模型提

取每个对视于频音帧频的模面态部本特文征使。用预训练的 [15]深度卷积神

, VGGish

经网络模型进行特征提取。VGGish模型在一百万个音频片段上进行了预训练,并为每个谱图段生成128维特征。

|  |
| --- |
| 70 软件工程 2024年10月 |

对于rPPG模态,本文使用短时端到端rPPG估计框架[16],该框架能够从视频流中检测到由血容量脉搏引起的微小颜色变化,进而实现rPPG的有效估计。具体来说,在人脸检测之后,首先使用类似Unet[18]的深度学习模型选择和跟踪感兴趣区域并进行皮肤和非皮肤像素的语义分割训练;其次计算皮肤分割像素的空间RGB通道均值,并将其投影到垂直于肤色的平面上,通过调整投影信号的alpha值获得rPPG信号值。

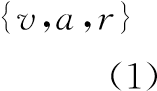
### 2.2多模态特征增强网络

本节将详细介绍多模态特征增强网络的相关模块,该网络由多个模态间Transformer(图2)、模态内Transformer(图3)和

多头自注意力机制组成,旨在共同捕捉输入序列每个时间步的模态内和模态间的动态关系,从而学习跨模态的渐进综合特征。

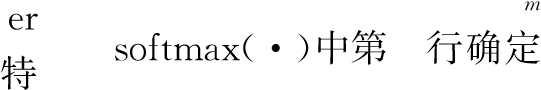
本文使用v、a、r分别表示视频模态、音频模态和rPPG模态。m和n分别代表两种不同的模态。通过使用Transformer 的注意力机制,逐步学习跨模态的全面特征。更具体地说,模态间Transformer通过与其他模态之间进行信息交互,学习到更全面的特征表示,以增强目标模态的特征。模态内 Transformer聚焦于目标模态,对其目标模态内部特征进行交互,从而捕捉输入特征中不同位置之间的依赖关系,关注到目标模态中最相关和有价值的信息,并且这些Transformer使用编码器结构,从另一个模态序列Un中提取当前模态序列Um 的信息。查询、键和值是Transformer编码器的3个输入。查询的来源是Um,键和值的来源是Un。因此,Transformer编码

器可以表示为

Un→m=Transformer(Um,Un,Un)∈RTm,d,m,n∈

Um→m=Transformer(Um,Um,Um)∈RTm,d,m∈{v,a,r}

(2)

考虑到目标模态与不同模态交互后的模态特征存在差异性,为了后续进行更好的融合,本文将从模态内Transform 和模态间Transformer获得的所有特征进行连接,作为增强征的输出,通过多头自注意力机制提取更丰富、更有用的特征, 该过程定义为

U=ConcaUt'm([=USnA1→m[i],,Umuln2(U→m),∈URm→Tmm,]3d)∈RTm,3d ((34)) 其中:m∈{v,a,r},n1、n2表示除m之外的其他两种模态, SA[i],mul表示多头自注意力机制,i表示第i个头部的注意力计算结果。考虑到跨模态的交互作用,期望增强的特征能够充分利用不同模态之间的互补性,生成更全面、更具表现力的特征表示。使用平均池化聚合目标模态特征,从而获得适用于下游任务的聚合特征。

#### 2.2.1模态间Transformer

模态间Transformer利用跨模态注意力机制处理不同模态之间的交互,使得一个模态的信息能够影响另一个模态的表示,从而对目标模态进行特征增强,模态间Transformer网络结构如图2所示。模态间Transformer是建立在跨模态注意力机制上的,下文将详细介绍跨模态注意力机制的原理。

假设有两种模m和n,分别对应着相应的序列Xm∈ R(Tm×dm)和Xn∈R(Tn×dn)。其中,T和d分别表示序列的长度和特征维度。受到原始Transformer[12]模型的启发,本文使

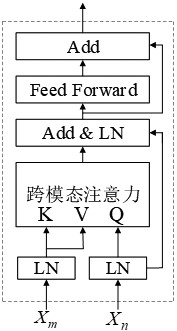
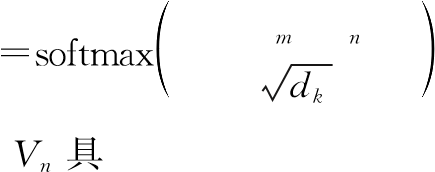
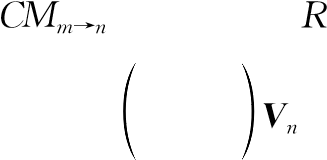


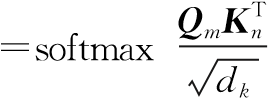
图2 模态间Tr ansformer网 络结构 Fig.2Inter-modelTransformernetwork

用m→n表示跨模态交互的潜在适应性,使用的模态包括视频、音频和rPPG信号。将查询、键、值矩阵分别d定m义dk为Qm=

XdmnWdQkm、Kn=XnWdKnn、dVkn=XnWVn,其中,WQm∈R × 、WKn∈

R × 和WVn∈R × 是权重矩阵。m→n的潜在适应性被定义为跨模态注意力,表示为

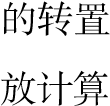
 Ym= (Xm,Xn)∈ Tm×dv

 (5)

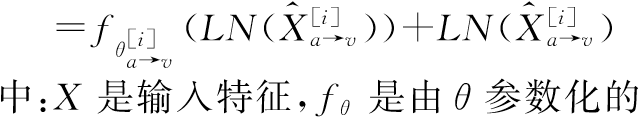
XmWQWKTXnT

XnWVn

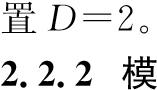
其中:Ym、Qm和 有相同的长度(即Tm和Tn),T表示向量

,操进作而。得在到公一式个(1得)分中,矩so阵ftmsoafxtm函ax数(通·过)∈除R以Tm×Tdnk,其进中行第缩

(i,j)个元素表示模态m的第i个时间步对模态n的第j个时间步的注意力。因此,Y 的第i个时间步是通过权重为 i 的对Vn进行加权汇总的结果。在这个意义上,公式(5)表示跨模态注意力。基于跨模态注意力块,本文设计了模态间Transformer,使一个模态可以从另一个模态获取交互信息。以将视觉(v)模态增强音频(a)模态为例,表示为v→a。每个模态间 Transformer由D层跨模态注意力块组成,其中i=1,…,D表示层数。模态间Transformer公式具体如下:

Xa[0→]v==CXMa[0a]^[iX→]v,ma[iu→l]v(LN(Xa[i→-v1]),LN(Xa[0→]v))+LN(Xa[0→]v)Xa[i→]v

(6)

其逐元素位置前馈子层,CMa[i→]v,mul表示多头注意力。LN表示层归一化。在文中设态内Transformer

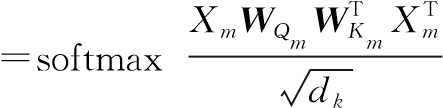
设计的模态内Transformer采用多头自注意力机制,专注于对每种模态内部的动力学进行建模,获得单一模态内的时间依赖性,使模块能够理解每种模态的信息随时间的演变。基于原始Transformer中的自注意力机制,本文设计了模态内

Transformer,用于跨单个模态的特征增强,模态内

XTrmaWnsQfom、r Kmemr=网X络mW结km构、 Q如m=图X3mW所V示m。为同查样询,、本键、文值定矩义阵Q,多m头= 自注意力表示为

Ym=SAm→m(Xm,Xm)∈RTm×dv

mn (7)

( )XmWVm

X

[

0

]

a

X

^

[

i

]

a

→

a

=

S

A

[

i

]

,

m

u

l

(

L

N

(

X

[

i

-

1

]

)

,

L

N

(

X

[

0

]

)

)

+

L

N

(

X

[

0

]

)

X

[

i

]

=

f

θ

[

i

]

a

→

a

(

L

N

(

X

^

[

i

]

a

→

a

)

)

+

L

N

(

X

^

[

i

]

a

→

a

)

8

S

A

[

i

]

,

m

u

l

a

→

a

表

示

多

头

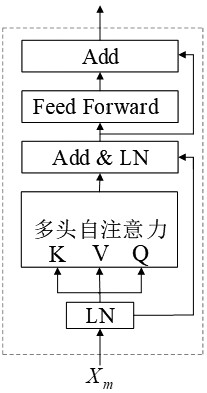
自

注

意

力

。



以音频模态为例,模态内Transformer表示为

Xa[0→]a=

a→a a→a a→a a→a a→a

() 其中,

图3 模态内Tr ansformer网 络结构

Fig.3Intra-modelTransformernetwork

通过结合模态间和模态内的Transformer输出,并利用多头自注意力机制,能够创建输入序列更全面的表示。这种方法使模型能够更全面地理解输入数据,并进一步实现整体的特征增强。在每种模态都通过多头自注意力机制获得统一增强的特征后,使用平均池化方法对这些特征进行展k平处理。

X'm=Avgpooling(U'm)∈R (9)

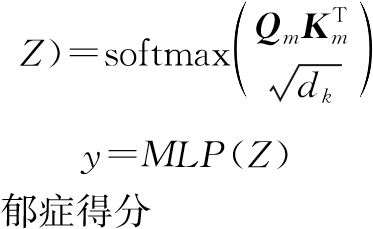
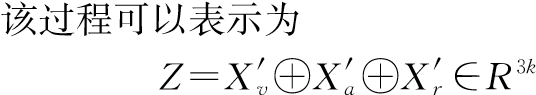
其中:U'm是多头自注意力的输出,k是输出特征的维度。

### 2.3抑郁症预测

本文将3个模态的特征级联在一起后进行最终的抑郁症预测,并且使用自注意 力机制过滤掉一些无用的特征,使用多层感知机(MultilayerPerceptron,MLP)进行最后的抑郁症分

数预测。

(10)

 Attention( Vm∈R3k (11)

(12)

其中,y是最后的抑 。

## 3实验与分析(Experimentandanalysis)

3.1数据集

本文在公开可用的抑郁症数据集———音频/视觉情感挑战 AVEC2013[19]上进行了实验。该数据集中的每个视频都附带有从贝克抑郁量表问卷(BDI-Ⅱ)回答中获得的标签,该量表将得分划分为0~63的范围。根据BDI-Ⅱ得分,抑郁的严重程度可以分为4个级别:最轻微(0~13)、轻度(14~19)、中度(20~ 28)和严重(29~63)。AVEC2013数据集由3个不同的部分组成:训练、验证和测试,每个部分包含50个视频,总共有150个

视频用于实验分析。

### 3.2评估指标

平均绝对误差(MAE)和均方根误差(RMSE)是评估自动抑郁检测任务中方法性能的常用指标。

N

MAE=N1i∑=1 yi-^yi (13)

=

1

N

∑

N

i

=

1

(

y

i

-

y

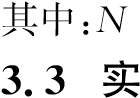
^

i

)

2

RMSE(14)

是样本数量,yi是样本标签,^yi是预测值。

### 验细节

对于视频模态,本文从每个视频中选择了100个连续帧, 并使用EmoFAN[14]预训练模型提取128维的面部特征,维度为(100,128)。对于音频模态,本文使用VGGish[15]预训练模型提取128维的音频特征,维度为(num\_segments,128),其中 num\_segments是分割后的频谱图段数。对于rPPG模态,获得了维度为(num\_seconds,10)的特征,其中num\_seconds是原始视频样本的持续时间。对于音频和rPPG模态,本文采用自适应平均池化[20]将提取的特征转换为(100,128)的固定特征维度供后续任务使用。使用的自适应平均池化[20]可以将具有任意空间维度的特征图转换为固定大小的表示。

所有深度学习 方法都在PyTorch框架上进行,并使用 NVIDIA RTX 3090GPU进行计算。使用Ad am优化器,初始学习率设置为0.001,权重衰减设置为0.00005。采用批量大小为4,并将最大训练轮数设置为1000。

### 3.4实验对比结果和分析

#### 3.4.1AVEC2013数据集实验结果

本研究在AVEC2013数据集上对比了单模态和多模态的结果。针对单模态情况,本文在模型中去除了模态间 Transformer模块。如表1所示,在AVEC2013数据集上,视频模态的表现优于音频模态和rPPG模态。视频模态的MAE 为8.67,而音频模态和rPPG模态分别为9.03和10.01。这一

优势可能源于视频中的面部表情为抑郁症检测提供了更多的线索。对于多模态融合方法,综合考虑3个模态的结果优于仅考虑两个模态的结果,这表明3个模态在一定程度上相互补充,为抑郁症检测提供了更全面的线索。同时,验证了rPPG 信号在多模态抑郁症检测中的有效性,为抑郁预测提供了额外的辅助信息。综上所述,将3个模态进行融合能够达到最佳的性能,验 证了本文提出模型的有 效性。

#### 表1 本 文方法 在AVEC 20 13 数据集上 取得的 结 果

Tab.1 Results obtained by the proposed method on theAVEC2013dataset

|  |  |  |  |
| --- | --- | --- | --- |
| 模态 | 特征类型 | MAE | RMSE |
| 单模态 | 音频视频 | 9.03  8.67 | 12.46  11.07 |
|  | rPPG | 10.01 | 12.95 |
| 多模态 | 音频+视频音频+rPPG | 8.06  8.83 | 10.65  12.21 |
|  | 视频+rPPG | 8.39 | 10.87 |
|  | 音频 视频 | 707 | 935 |

+ +rPPG . .

|  |
| --- |
| 软件工程 2024年10月 |

##### 3.4.2消融实验

为了评估模型中每个多模态特征增强模块的有效性,在 AVEC2013数据集上进行消融实验。本文进行了不同模块的组合实验,实验结果如表2所示。其中,模态间、模态内和多头自注意力分别表示模型多模态特征增强部分仅使用模态间 Transformer、模态内Transformer或多头自注意力机制,“+”

表示使用两个模块的组合。

表2中的结果显示:仅使用模态间Transformer的性能优于仅使用模态内Transformer,但低于这两个模块的组合使用。这表明,通过模态间Transformer中的跨模态注意力机制,能实现模态之间更有效地交互,从而对目标模态实现特征增强, 同时,模态内Transformer也能关注到目标模态在时间上的变化信息。因此,将这两个模块结合使用能够获得更好的效果。此外,单独使用多头自注意力机制模型效果并不理想,然而当与模态间Transformer和模态内Transformer模块结合使用时,达到了本模型的最优效果。这表明,多头自注意力机制在一定程度上能够学习到模态交互后更全面的特征,从而实现整体的特征增强。

#### 表2 不同模块对实验的影响

Tab.2Theimpactofdifferentmodulesonexperiments

|  |  |  |
| --- | --- | --- |
| 模块类型 | MAE | RMSE |
| 模态间模态内  模态间+模态内多头自注意力多模态特征增强网络 | 7.88  8.13  7.53 8.49  707 | 10.26  10.92  10.02  11.68  935 |

. .

3.4.3不同模型对比结果

为了更全面地评估本文提出模型的有效性,在AVEC2013 数据集上将其与目前较先进的方法进行了对比,对比结果如表 3所示。根据所使用的模态数量,这些方法可以分为3个主要类别:基于音频的抑郁症检测方法、基于视频的抑郁症检测方法以及基于音频和视频的双模态抑郁症检测方法。以下是对一些具有代表性检测方法的简要介绍,更多的信息可参考表3 中列出的相关文献。

基于音频的抑郁症检测方法:VALSTAR等[19]提取了 LLD声学特征,并采用支持向量回归进行抑郁症检测。HE 等[21]将深度音频特征与深度CNN和手工纹理特征相结合后, 通过全连接层进行抑郁得分预测。NIU等[22]提取短时MFCC 段的分段级特征并采用支持向量回归预测个体的抑郁水平。 ZHAO等[23]提出了一种混合特征提取网络,将DCNN与自注意力网络集成,用于从语音信号中检测抑郁严重程度。

基于视频的抑郁症检测方法:ZHU等[24]提取了LPQTOP特征,并通过稀疏编码进行学习,以进一步提高抑郁症检测的准确性。JAZAERY等[25]使用3D卷积神经网络(3DCNN)捕捉面部区域在两个不同尺度上的时空特征,并在决策层上进行融合。HE等[21]提出了一种名为DepNet的集成框架,用于捕捉视频中面部表情的时间动态特征,以进行抑郁症分析。

基于音频和视频的双模态抑郁症检测方法:MENG等[26] 使用LLD声学特征对音频特征进行编码,并使用运动历史直方图捕捉面部区域内每个像素的运动,最终在决策层上融合音频和视频特征后,进行最终的抑郁症预测。NIU等[10]提出了一种时空注意网络和多模态注意特征融合策略,用于通过音频和视频预测个体的抑郁水平。

#### 表3 AVEC 20 13 数据集 测 试对比结果

Tab.3 Comparison of test results on AVEC2013 test set

|  |  |  |  |
| --- | --- | --- | --- |
| 模态类型 | 方法 | MAE | RMSE |
| 音频 | 文献[19]方法文献[21]方法文献[22]方法文献[23]方法 | 10.35  8.20  7.48  7.38 | 14.12  10.00  9.79  9.65 |
| 视频 | 文献[10]方法 文献[24]方法文献[25]方法文献[21]方法文献[27]方法 | 7.14  7.58  7.37  7.55  7.52 | 9.50  9.82  9.28  9.20  9.22 |
| 音频+视频 | 文献[26]方法文献[28]方法文献[29]方法 | 8.72  7.68  9.09 | 10.96  9.44  11.19 |
| 音频 视频 | 本文方法 | 707 | 935 |

+ +rPPG . .

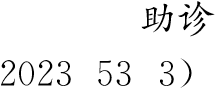
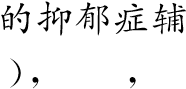
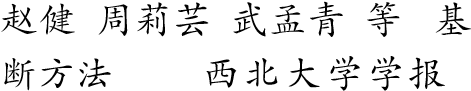
## 4结论(Conclusion)

针对自动抑郁症检测任务中传统的方法存在的问题,例如不能充分利用不同模态信息、未充分考虑多模态融合过程中模态间的交互等,本文提出了一种基于多模态特征增强网络的抑郁症检测方法。该方法通过与不同模态之间的交互,实现目标模态的特征增强,并融合了多种模态,将rPPG模态与视频模态和音频模态结合应用于多模态抑郁症检测任务。本文提出的方法利用模态间Transformer、模态内Transformer和多头自注意力机制逐步学习视频、音频和rPPG等不同模态的综合特征。在AVEC2013公共数据集上进行的大量实验证明,本文提出的方法在多模态抑郁症检测任务上展现出良好的性能。

本文提出的模型能更好地挖掘不同模态中的抑郁线索,为多模态融合提供了新思路。在未来工作中,我们将探索跨模态对齐的先进方法,以期进一步提高多模态融合效果。此外,研究发现,rPPG信号的性能并不优于视频模态和音频模态,这可能是由于提取rPPG信号值的方法不够精确。因此,使用更先进的方法提取更具表达力的生理信号,并将其应用于多模态融

合具有重要的研究意义。

## 参考文献(References)

[1] , , ,. 于人工智能

[J]. (自然科学版(:

325-335.

### [2]李欣,范青.机器学习在抑郁症患

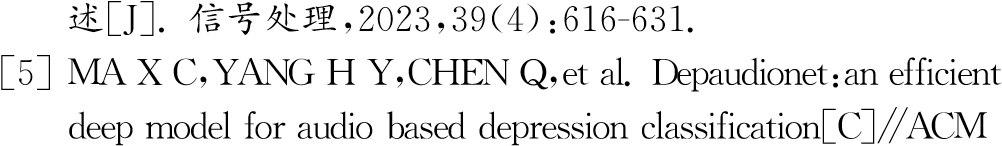
用进展[J].上海交通大学学报 2022421:

合通道层注意力机制的多

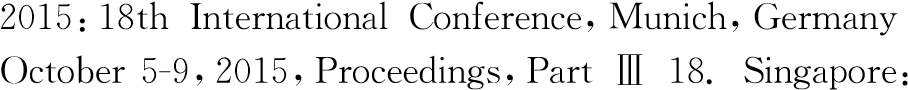
.中国图象图形学报,2022,

27113292-3302.

### [4]刘振焘,向春妮,刘陈陵,等.基于语音的抑郁检测研究综

.

ANPUREVSSUNGH

1. ssbPDEmiiprnoeAmoogenconGtnet sigeDeaoi-d nnnibndea eCgscrsheiliadnonlgflisecc tanar hlngeedeean6p. bit pinNhlo,gietcIewasnctytih esoYrntnoenorlamokotfigof:ayoAnra,CnlSomMHvWaeIj,lo2oNr 0rak1Busd6tAheo:op3nrpT5oe-so,m4esn2iticv.Aaet lruda.din isDosoi/erevdVneetilrsoru puea---l [19]2SV0p1rAi3nL:gSt ehTreAIcnRot nertMinna,utSiooCuns Hala[UuCPLdu]iLbo∥l/EiAsvRihCsiun BMagl,.,S2 e P0Mmr1Ioo5Tctie:oH2e/n3diV4Kan-in2sg,du4sea1t ldo.fe aElp tr.mheesoAst 3iivoorendnc,

afingertipphotoplethysmogr,a2p0h1y8s,6en(2so)r:6[J4-]7.4F.rontiersin rAecCoMgniItnitoenrncahtailolneanlgeWorkshoponAudio

1. ttrsaCPiiierwvnoAcoeganoScyg eCrAne: eoIiDdctE miioOnoEgpgnn uEsCittiwio ni[oÁftJgnht, ]ah) Cud.ne:s3A yd iI23nnEÑ0I0ag n1E5E,mt -5rESLei3eclIAL3 lmtin1vrAgHtoi6aeestSLnrneu.In sta aMaplIcthHniat ooi tLnot.endo,rnaaLplsMacl Ó etCuoitlPoonthinnEy fmasZeforfm edeMonacgltc [ireBCvade o.pe] n hp∥cDryIoeeAE psfmfr sErfpioeeEouscmn---.  [[2201]] ttevCHHreixhaddEpUanerlisolAeLnf-esodbNns ,rauiGgGossmeeUtn er.dQri Or aNe[d lJcHeeCoi]pwng,r.Gtn eH IeYis n,ltUsolifTiirooIgoAkrnenWNmn: wAtaAaGintCtRis ahoCyIlMnys  gtPsQ, rsie2isc,0d,mei-1et[WJn3w)uacAi]l:s:se33i.N.es-8nI1 1Gg,Dan205tte 0t-edpX.e23re nN18n eZt3eapi,tt55o,ie8ln:o.tae0 n an aaa:rnll3nad5.iju-n o tvF5goui4asrmcfun.ioaaaa -rlll facial,20vi2d3eo,1s4(4

[8]HEL,JIANGDM

udiocues

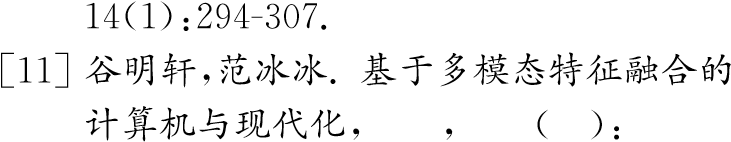
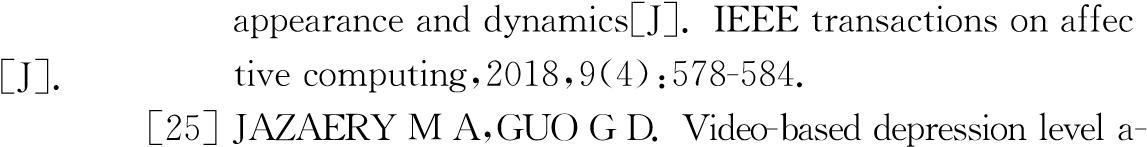
(ACII).Piscat- ofintelligentsystems,2022,37(7

,2015:260-266. [22]NIUM,TAOJ,LIUB,etal.Automaticdepressionlevelde-

[9]sbpYhruiAatdillNn oagGrwc,2h L0imt2,eoJ1cItAurNe sGD M ,SA H LI H.Inte gra tingdeep —ahnyd- [23] toZefcH tIinAotneOrvs Zipae, elLcpI-h nQ .o r,ACmuUstprMoioalMi:InI NgNT[SCE N]R∥S,IePtNE aTlEE.C RHHSyP,2bE0r i1Ed9C n:H4e5t.5w9P[o-Jrr4ok]5c.6e f 3eeIadn.ittnuegrr-es delsfo[Jr):]2m.3Iu9l-Et2iE-5E3mot.rdaanlsadcetpiroensssioonnaaffneacltyisviescom- extractionfordepressionassessmentfromspeech

,12(1

[10] pNoIrUal[J Mr]e. pY IreE,sTEeAnEtO attr Jiaon nHsa f,coLtriIo Uanus Bt oo,enmta a atfilfce. c dMteivuplertei csmosiomodnpaul lte sipnagti otem- [24] nZaHtioUna Yl j, oSuHrnAalN oGf i n Yte lYlig,Se nHtA syOst Ze m Hs,,2e0t2 a0l,.6 A(3u t)o:8m-1a2te. d de- tion vel,d2e0t2e3c-, pressiondiagnosisbasedondeepnetworkstoencodefacial

抑郁症识别 -

# ,2017

[[1142]]  tttppTVirarroatOeoAinnsIcoSssSeinfissWaOoosl rniaUA nlsm ilLgoNne c I yrsi As-oa yobAlus,ca tKis ,seanSyeOlemsdHS e 2sm Sd0AmeA2[ZdoIJ3i EdFa],eI2E2[.lJ 00sJR ,223]A , 4f30,.dNoA Iv ,(r112E,Sa P10niEK)dcA(:E2Oee2Rs1n2 t )Ut7r5 M:ii-a1-fNn22n9yAI,2 i3s7SnnRa5 9.egc- D u.t1N irs.9oat,9lnC er 0steaisln .oi asfbl no r.ar ca ntomAidotmatt npeido nueon---f  [[26]] tonnumMrifastaioli Etnnytoihss gN noieasnGcp b3t a iabrCrHosdy,th neia asdYeAllln oeC,olcnneHnoMag dasUeidf tInfy.An, e gsn tcNNqea tdriueGmvneawieercDep cYf, orsoaeWmcrpg i Akapratle uiNsotaisGtinneog dmnH,p2v[,o0oCe rct2a]a1 l al∥l , f1e.Ae2 xaDCpt (eur1Mper)rse .ess: 2isPos6[/riJn2oov-]cinf2s.e e6 rue aIe8datcEliu.onr EEggeE--ss

[13]ILIASL,MOUZAKITISS

tems ational:WAoCrkMs,H20o1p3o:n21-A3u0di.o

BULATAetal.Estimationof 27DUZYLIWXHUANGD,etal.Encodingvisualbehav-

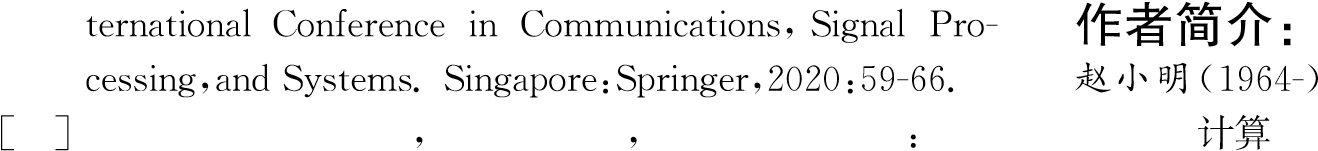
ccoonntdiintuouss[Jva]l.e nNcaetuarned amraocuhsianlel ienvteellslifgreonmcef,a2c0e2s1in,3n:a4t2u-r5a0lis.tic itioorsn[wCit]h∥IatEtEenEti.v ePrtoecmeepdoirnaglsc oonfv tohleut 2io0n19fo 1r4dtehp IreEsEsiEon Inptreerdniac-ion

1. IfcCeHaErNtEEeaNnERwcaSa.eryHPcorhnEiotYceAeccetSoduiu,rnsCegtsiHscfsoAorfUltaDhreHge2U-s0Rc1Ia7 leSIaE,uEEdLiEoLIIcnlSatesDrsinfaiPctiaotWinoa,nle[tCCoa]nl∥-.[28]t(iKFoAnGa Yl2A0C1o 9nHf)e,.reÇPnIicsLecLaotIanwFAa ,yuSt:IAoEmLEaAtEiHc,2[F0C Aa1c]9e∥A :1A&-.7CE.GMnes/s.etVuPmirrseboulcaeRelee CcEdoiCmgnAngoitstifioooonnrf

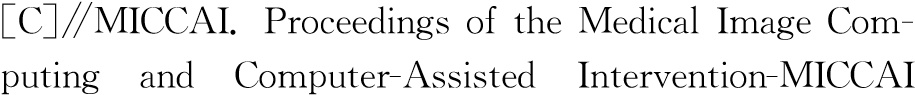
,SpeechandSignalProcessing.Pis- continuousemotionprediction

:IEEE,2017:131-135. the4thInternationalWorkshoponAudio

1. pLhI B,JIA NG W,PENG J,et al. Dee p learn ing-based re mote- [29] CKhÄ alCleHnEgeL.E N Mew ,GYLorOkD:EA KC MM,2 Z0 H14A:R19K-2O6V. D ,et al. Fu sion vidoetoop[Jle]t.hPyshymsoioglroagpihcaylmmeeaassuurreemmeenntt,2f0r2o2m,43s(h1o1rt)-:t1i1m5-e12f3ac.ial ofaudio-visualfeaturesusinghierarchicalclassifiersys-
2. JvIolAuNtioGn aB l, nReEurNal Q ne,DtwAoIr kFs,e ftor a l r.ea lM-tuiltmi -et adsykn aca msicca fdae cde c r oenc-- tdeepmrse sfsoiro tnhe[J re]c.o gRneictio ognni otfio anff eacptipvliec asttiaotness aanndd t hmee stthatoeds of, ognitionmethod[C]∥LIANGQ,LIUX,NAZ,etal.In- 2014,1(1):671-678.

,男,硕士,教授。研究领域:模式识别,情感

18RONNEB ERGERO FIS CHERPB ROXT .U-netCon- 范慧婷 。女硕士生 研究领域人工智能

volutionalnetworksforbiomedicalimagesegmentation 张石清(1998-),男,博士。教授 研究:领域模式。识别

(1980-),, , 。 : ,情感

计算。