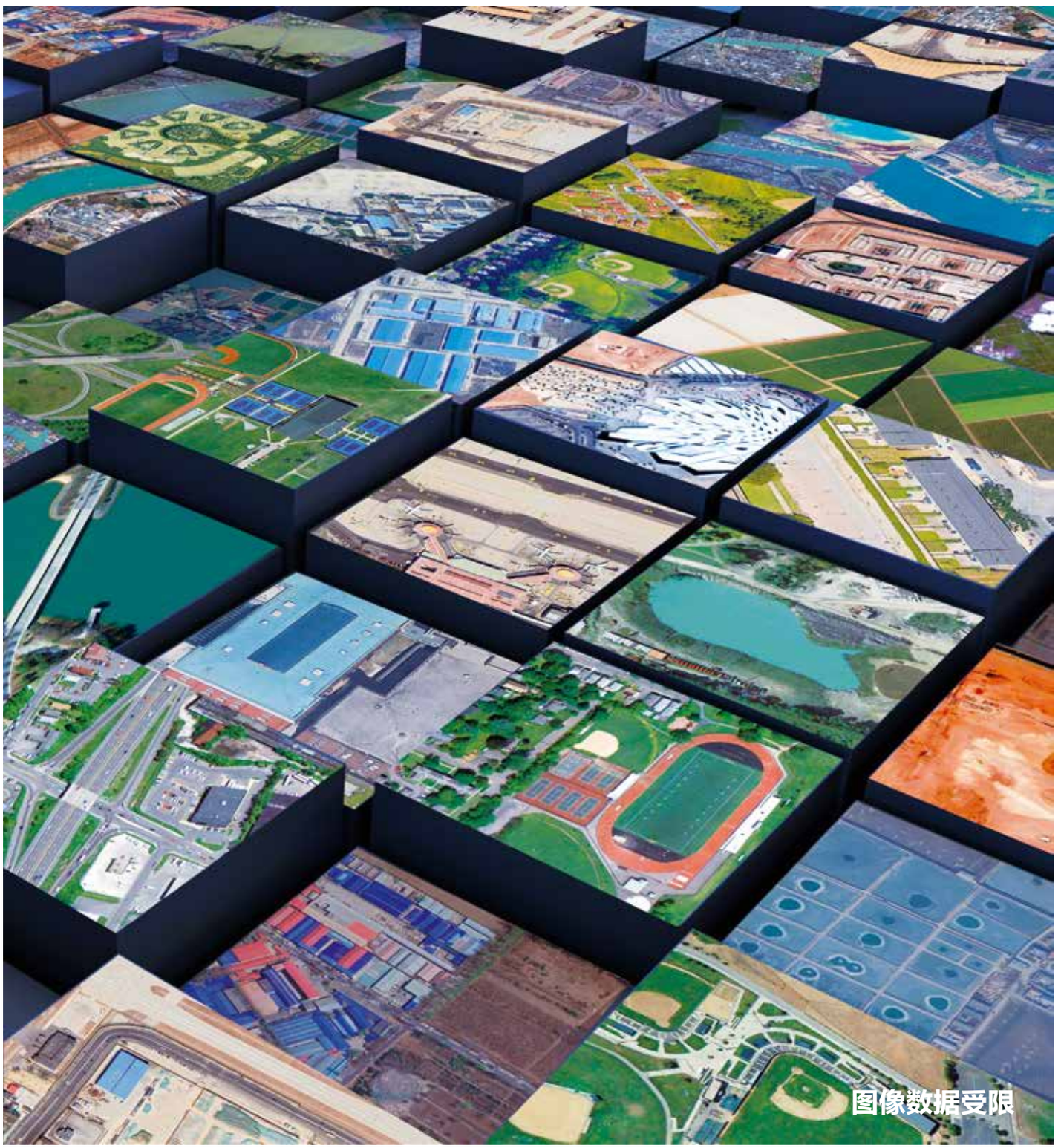
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
10
VOL.27

ISSN1006-8961
CN11-3758/TB



中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第10期（总第318期）
2022年10月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

版权声明

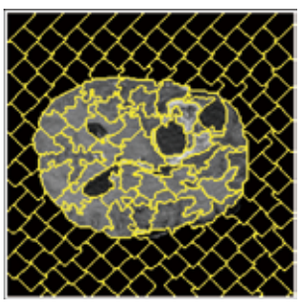
凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

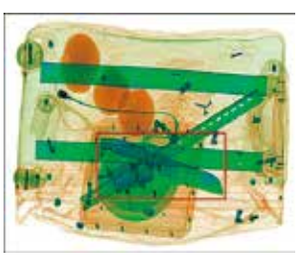
All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

2022年10月 第27卷 第10期（总第318期）
中国图象图形学报 Zhongguo Tuxiang Tuxing Xuebao

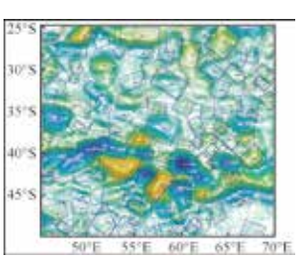
目次



MRI脑肿瘤图像的超像素/体素分割及发展现状(第2897页)



融合多尺度特征与全局上下文信息的X光违禁物品检测(第3043页)

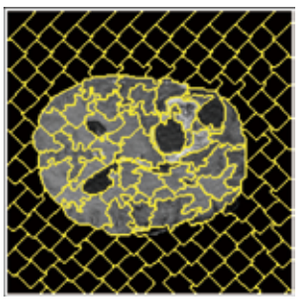


融合多尺度旋转锚机制的海洋中尺度涡自动检测(第3092页)

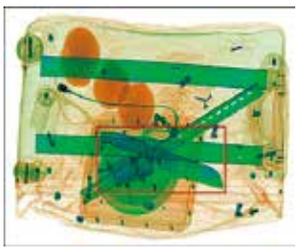
封面素材来自中国科学院空天信息创新研究院二部孙显研究员团队

Volume 27, Number 10
Published October 16, 2022

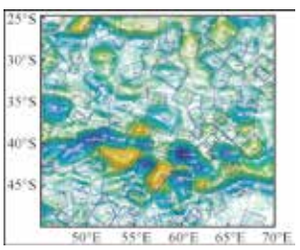
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



The review of superpixel/voxel segmentation of MRI brain tumor images(P2897)



Integrated multi-scale features and global context in x-ray detection for prohibited items(P3043)



Multi-scale rotating anchor mechanism based automatic detection of ocean mesoscale eddy(P3092)

E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2022)10-3022-09

论文引用格式:Chen G, Zhang S Q and Zhao X M. 2022. Video sequence-based human facial expression recognition using transformer networks. Journal of Image and Graphics,27(10):3022-3030(陈港,张石清,赵小明. 2022. 采用 Transformer 网络的视频序列表情识别. 中国图象图形学报,27(10):3022-3030)[DOI:10.11834/jig.210248]

采用 Transformer 网络的视频序列表情识别

陈港[1,2],张石清[1],赵小明[1,2*]

1. 台州学院智能信息处理研究所,台州　318000;2. 浙江理工大学机械与自动控制学院,杭州　310018

摘　要:**目的**　相比于静态人脸表情图像识别,视频序列中的各帧人脸表情强度差异较大,并且含有中性表情的帧数较多,然而现有模型无法为视频序列中每帧图像分配合适的权重。为了充分利用视频序列中的时空维度信息和不同帧图像对视频表情识别的作用力差异特点,本文提出一种基于 Transformer 的视频序列表情识别方法。**方法**　首先,将一个视频序列分成含有固定帧数的短视频片段,并采用深度残差网络对视频片段中的每帧图像学习出高层次的人脸表情特征,从而生成一个固定维度的视频片段空间特征。然后,通过设计合适的长短时记忆网络(long short-term memory network,LSTM)和 Transformer 模型分别从该视频片段空间特征序列中进一步学习出高层次的时间维度特征和注意力特征,并进行级联输入到全连接层,从而输出该视频片段的表情分类分数值。最后,将一个视频所有片段的表情分类分数值进行最大池化,实现该视频的最终表情分类任务。**结果**　在公开的 BAUM-1s(Bahcesehir University multimodal)和 RML(Ryerson Multimedia Lab)视频情感数据集上的试验结果表明,该方法分别取得了 60.72% 和 75.44% 的正确识别率,优于其他对比方法的性能。**结论**　该方法采用端到端的学习方式,能够有效提升视频序列表情识别性能。

关键词:视频序列;人脸表情识别;时空维度;深度残差网络;长短时记忆网络(LSTM);端到端;Transformer

Video sequence-based human facial expression recognition using Transformer networks

Chen Gang[1,2], Zhang Shiqing[1], Zhao Xiaoming[1,2*]

1. Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, China;

2. School of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China

Abstract: **Objective** Human facial expression is as one of the key information carriers that cannot be ignored in interpersonal communication. The development of facial expression recognition promotes the resilience of human-computer interaction. At present, to the issue of the performance of facial expression recognition has been improved in human-computer interaction systems like medical intelligence, interactive robots and deep focus monitoring. Facial expression recognition can be divided into two categories like static based image sand video sequence based dynamic images. In the current "short video" era, video has adopted more facial expression information than static images. Compared with static images, video sequences are composed of multi frames of static images, and facial expression intensity of each frame image is featured. Therefore, video sequence based facial expression recognition should be focused on the spatial information of each frame

收稿日期:2021-04-07;**修回日期**:2021-06-30;**预印本日期**:2021-07-06

*通信作者:赵小明　tzxyzxm@163.com

基金项目:国家自然科学基金项目(61976149);浙江省自然科学基金项目(LZ20F020002)

Supported by: National Natural Science Foundation of China (61976149); Natural Science Foundation of Zhejiang Province, China (LZ20F020002)

and temporal information in the video sequences, and the importance of each frame image for the whole video expression recognition. The early hand-crafted features are insufficient for generalization ability of the trained model, such as Gabor representations, local binary patterns (LBP). Current deep learning technology has developed a series of deep neural networks to extract facial expression features. The representative deep neural network mainly includes convolutional neural network (CNN), long short-term memory (LSTM). The importance of each frame in video sequence is necessary to be concerned for video expression recognition. In order to make full use of the spatio-temporal scaled information in video sequences and driving factors of multi-frame images on video expression recognition, an end-to-end CNN + LSTM + Transformer video sequence expression recognition method is proposed. **Method** First, a video sequence is divided into short video clips with a fixed number of frames, and the deep residual network is used to learn high-level facial expression features from each frame of the video clip. Next, the high-level temporal dimension features and attention features are learned further from the spatial feature sequence of the video clip via designing a suitable LSTM and Transformer model, and cascaded into the full connection layer to output the expression classification score of the video clip. Finally, the expression classification scores of all video clips are pooled to achieve the final expression classification task. Our method demonstrates the spatial and the temporal features both. We use transformer to extract the attention feature of fragment frame to improve the expression recognition rate of the model in terms of the difference of facial expression intensity in each frame of video sequence. In addition, the cross-entropy loss function is used to train the emotion recognition model in an end-to-end way, which aids the model to learn more effective facial expression features. Through CNN + LSTM + Transformer model training, the size of batch is set to 4, the learning rate is set to 5×10^{-5}, and the maximum number of cycles is set to 80, respectively. **Result** The importance of frame attention features learned from the Transformer model is greater than that of temporal scaled features, the optimized accuracy of each is 60.72% and 75.44% on BAUM-1 s (Bahcesehir University multimodal) and RML (Ryerson Multimedia Lab) datasets via combining CNN + LSTM + Transformer model. It shows that there is a certain degree of complementarity among the three features learned by CNN, LSTM and Transformer. The combination of the three features can improve the performance of video expression recognition effectively. Furthermore, our averaged accuracy has its potentials on BAUM-1 s and RML datasets. **Conclusion** Our research develops an end-to-end video sequence-based expression method based on CNN + LSTM + Transformer. It integrates CNN, LSTM and Transformer models to learn the video features of high-level, spatial features, temporal features and video frame attention. The BAUM-1 s and RML-related experimental results illustrate that the proposed method can improve the performance of video sequence-based expression recognition model effectively.

Key words: video sequence; facial expression recognition; spatial-temporal dimension; deep residual network; long short-term memory network (LSTM); end-to-end; Transformer

0 引 言

人脸表情在人际交往中是不可忽视的重要信息载体,表情使人与人之间的沟通更加自然。人脸表情识别技术的进步促进着人机交互技术的发展。如何提升人脸表情识别性能是智慧医疗、聊天机器人和学生专注度监测等人机交互系统中一个重要的热点研究课题。目前已有的人脸表情识别技术(李珊和邓伟洪,2020)可以分为基于静态图像的表情识别和基于动态视频序列图像的表情识别两类。基于静态图像表情识别的方法大多考虑如何提取有效的空间特征(王善敏 等,2020)以及通过注意力机制关注各局部区域关键信息(Li 等,2019)和互相关性(Li 等,2020)来提高模型的识别性能。刘帅师等人(2011)采用 Gabor 滤波器从图像中提取不同尺度的多方向特征,再将多个多方向 Gabor 特征进行融合,并分块计算每块的直方图用于人脸表情识别。相比于静态图像,视频序列是由若干帧静态图像组成的,并且每帧图像的人脸表情强度具有一定的差异。因此,基于视频序列的人脸表情识别方法不仅要考虑每帧人脸表情图像的空间维度信息,还要关注视频序列包含的时间维度信息以及每帧图像对于整个视频表情识别的重要性的不同。在当前短视频流行背景下,视频包含的人脸表情信息相比静态图像更加丰富,视频序列的人脸表情识别比静态图像更具有

挑战性。De Silva 和 Ng(2000)对每个视频序列特征采用一对速度和位移特征向量表示,并采用光流算法(Anandan,1989)跟踪视频序列中的特征运动,实现了基于视频序列的表情识别。为了从每帧图像中提取有效的人脸表情特征,付晓峰等人(2015)对局部二值模式(local binary pattern,LBP)进行方向编码,形成局部方向角模式(local oriented pattern,LOP),再扩展到3维空间,提出时空局部方向角模式,以提取视频的全局表示特征进行视频人脸表情识别。然而,上述方法采用的手工特征是低层次的,可靠性不够,导致训练出的模型泛化能力不强。

随着深度学习技术的发展,提出了一系列可用于提取视频人脸表情特征的深度神经网络,代表性的有卷积神经网络(convolutional neural network,CNN)(李彦冬 等,2016)和长短时记忆网络(long short-term memory network,LSTM)(Sun 等,2016)。王晓华等人(2020)采用堆叠 LSTM 的方式学习视频序列数据的分层表示,使用注意力机制进行相应的权重分配,构建出有效的情感层次特征表示,提高了模型的识别准确率。Hu 等人(2019)提出一个基于级联的局部增强运动历史图像(Ahad 等,2012)和 CNN-LSTM 网络集成结构的视频表情识别方法。其中,CNN 用于提取每帧图像的空间特征,LSTM 用于学习视频图像帧之间的时间维度信息表示。该集成框架用于提取视频人脸表情的全局特征,较大程度地改善了识别模型的性能。Fan 等人(2016)提出一个混合网络框架,分别采用循环神经网络(recurrent neural network,RNN)和 3D 卷积神经网络对外观和运动信息进行编码。其中 RNN 的输入为 CNN 从每帧图像学习到的特征,然后将 RNN 和 3D 卷积神经网络输出的特征进行级联,用于实现视频情感分类。

现有基于深度学习方法的视频表情识别模型大多具有较强的特征提取能力,同时也考虑了视频序列的时间维度信息,但很少关注视频序列中每帧图像对视频表情识别的重要性。对此,以自注意力机制(self-attention)为核心的 Transformer(Bahdanau 等,2016;Brown 等,2020;Dosovitskiy 等, 2021)方法提出了解决方案。Transformer 模型将注意力机制作为编解码器的核心执行翻译操作,并成功应用于自然语言处理领域,取得了很好的机器翻译性能。Dosovitskiy 等人(2021)提出一个视觉 Transformer 模型,将处理文本序列的 Transformer 模型成功应用在图像识别领域,取得了不错的图像识别性能。考虑到 Transformer 强大的注意力特征的学习能力,本文提出一种基于 Transformer 的视频表情识别方法。首先采用在 ImageNet 数据集(Deng 等,2009)预训练好的深度残差网络(residual network,ResNet)学习出视频片段序列中每帧图像的人脸表情特征,并将其作为后续 LSTM 模块和 Transformer 模块的输入。然后将 LSTM 学习到的时间维度特征和 Transformer 输出的注意力特征进行级联,构建出一个视频片段注意力特征,输入到 softmax 层,得到每个片段的分类分数值。最后将一个视频中所有片段的表情分类分数值(classification scores)进行最大池化,得到视频样本的表情类别。该方法融合 Transformer 机制对视频序列中的每帧表情图像进行加权特征学习,学习出不同帧图像对视频表情识别的作用力差异性,提取到判别力强的视频表情注意力特征。在两个公开视频表情数据集 BAUM-1s(Bahcesehir University multimodal)(Zhalehpour 等,2017)和 RML(Ryerson Multimedia Lab)(Wang 等,2012)的实验结果表明,本文方法能取得较好的识别准确率。

1 本文方法

本文提出的端到端的 CNN + LSTM + Transformer 的视频人脸表情识别模型结构如图1所示,具体包括以下步骤:1)片段空间特征提取。将视频分成若干包含16帧的片段,通过深度残差网络学习出一个片段特征 $\boldsymbol{G} \in \mathbf{R}^{N\times d}$ 并进行归一化处理。2)片段时间维度特征提取。将视频片段特征输入到 LSTM 模块中,学习出视频片段的时间维度特征信息 $\boldsymbol{e} \in \mathbf{R}^{1\times d}$。3)片段帧注意力特征提取。将视频片段特征输入到 Transformer 模块中,学习得到一个片段帧注意力特征 $\boldsymbol{Z}_{\text{class}} \in \mathbf{R}^{1\times d}$。4)视频表情分类。将时间维度特征和帧注意力特征进行级联后,输入到全连接层中,输出一个 1×6 维的片段分类分数值,然后采用最大池化操作,输出视频人脸表情的类别。

综上所述,本文方法不仅考虑了视频中每帧图像的空间特征,还考虑了时间维度特征。根据视频序列每帧图像中人脸表情强度的差异性,提出采用 Transformer 提取片段帧注意力特征,提高了模型的表情识别率。此外,采用交叉熵损失函数以端到端的方式训练情感识别模型,有助于模型学习到更有

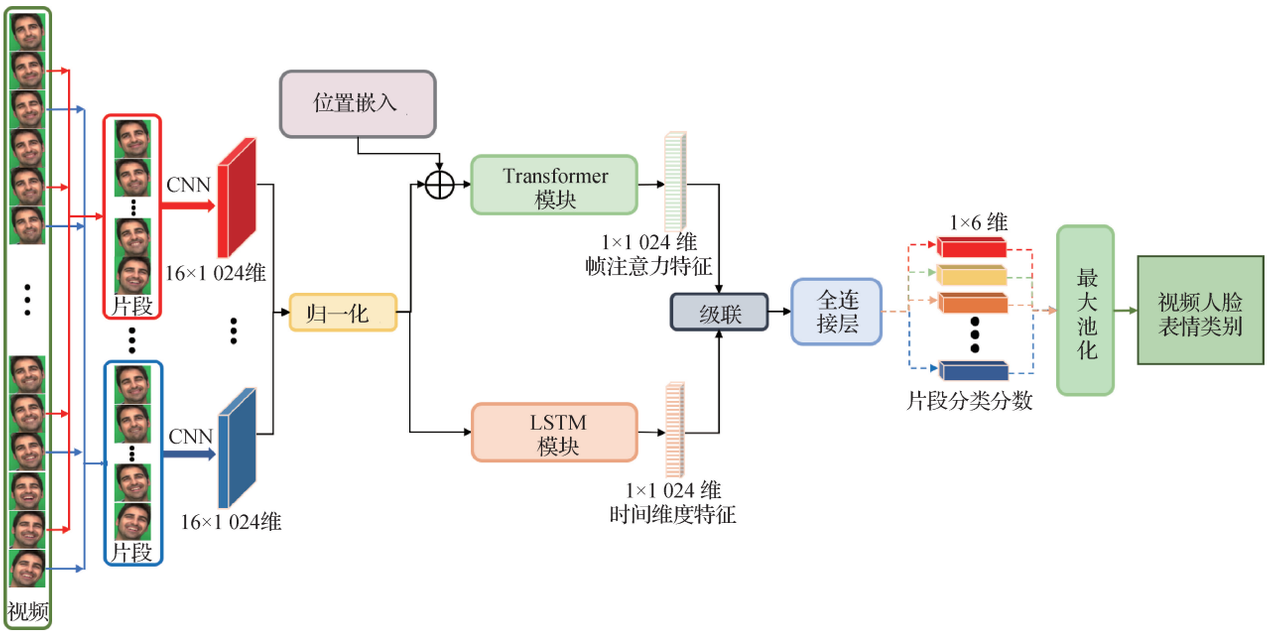


图 1　CNN + LSTM + Transformer 模型框架示意图

Fig. 1　Schematic diagram of CNN + LSTM + Transformer model framework

效的人脸表情特征。该损失函数定义为

$$\min \gamma(P(x_i),y_i) = -\sum_x (P(x_i)\log y_i) \quad (1)$$

式中，x_i 表示输入的第 i 个片段序列；y_i 表示第 i 个片段的表情标签；$P(x_i)$ 表示模型预测值。

1.1　片段空间特征提取

为了有效提取视频片段中每帧的特征，采用在 ImageNet 数据集上预训练好的 ResNet-18 提取每帧图像的人脸表情特征。给定一个视频的帧为 $\boldsymbol{s}_i \in \mathbf{R}^{N\times W\times 3}$，其中，$H$ 和 W 分别为每帧图像的高和宽，生成的特征为 $\boldsymbol{G}_i \in \mathbf{R}^{1\times d}$，其中，$d = 1\,024$ 为片段特征的维度，残差网络表示为 $E(\cdot)$，特征学习表示过程为

$$\boldsymbol{G}_i = E(\boldsymbol{W}_{\text{Resnet}},\boldsymbol{s}_i), \quad i \in [1,N] \quad (2)$$

式中，$\boldsymbol{W}_{\text{Resnet}}$ 为残差网络的可学习权重参数；$N = 16$ 为片段包含的帧数；后文的 N 和 d 表示同一个含义。这样，每帧图像生成的特征是 1 024 维，一个片段含有 16 帧图像，因此，一个片段特征 $\boldsymbol{G}$ 的维度为 16 × 1 024 维。

1.2　片段时间维度特征提取

相比于静态的图像，视频片段中的时间维度特征信息有助于视频表情分类。因此，将 LSTM 网络用于视频片段中帧与帧之间的时间动态信息建模，以学习出视频片段的时间维度特征信息。给定输入一个视频片段图像序列 $\boldsymbol{N}(\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_N)$，时间步长 $t \in [1,N]$，LSTM 中的细胞(cell)单元处理序列数据过程为

$$\boldsymbol{I}_t = f_{\text{sigmoid}}(\boldsymbol{W}_I[\boldsymbol{G}_t,\boldsymbol{H}_{t-1},\boldsymbol{C}_{t-1}] + b_I) \quad (3)$$

$$\boldsymbol{F}_t = f_{\text{sigmoid}}(\boldsymbol{W}_F[\boldsymbol{G}_t,\boldsymbol{H}_{t-1},\boldsymbol{C}_{t-1}] + b_F) \quad (4)$$

$$\boldsymbol{O}_t = f_{\text{sigmoid}}(\boldsymbol{W}_O[\boldsymbol{G}_t,\boldsymbol{H}_{t-1},\boldsymbol{C}_{t-1}] + b_O) \quad (5)$$

$$\boldsymbol{C}_t = \boldsymbol{F}_t\boldsymbol{C}_{t-1} + \boldsymbol{I}_t\tanh(\boldsymbol{W}_C[\boldsymbol{G}_t,\boldsymbol{H}_{t-1}] + b_C) \quad (6)$$

$$\boldsymbol{H}_t = \boldsymbol{O}_t\tanh(\boldsymbol{C}_t) \quad (7)$$

式中，$\boldsymbol{I}_t$，$\boldsymbol{F}_t$ 和 $\boldsymbol{O}_t$ 分别表示 LSTM 单元中的输入门、遗忘门和输出门。$\boldsymbol{G}_t$ 表示当前 LSTM 单元当前时刻的输入；$\boldsymbol{H}_{t-1}$ 和 $\boldsymbol{H}_t$ 分别表示 LSTM 单元的上个时刻和当前时刻的隐藏状态；$\boldsymbol{C}_{t-1}$ 和 $\boldsymbol{C}_t$ 分别表示上个时刻和当前时刻的输出状态。$\boldsymbol{W}$ 和 b 表示可学习的门权重参数和偏置项(下标表示对应的门)；sigmoid 和 tanh 表示 sigmoid 激活函数和 tanh 函数。

给定输入序列 $\boldsymbol{G}$，将其输入到 LSTM 网络，相应的学习过程为

$$\boldsymbol{e} = L(\boldsymbol{W}_{\text{LSTM}},\boldsymbol{G}) \quad (8)$$

式中，时间维度特征 $\boldsymbol{e} \in \mathbf{R}^{1\times d}$ 由函数 $L(\cdot)$ 计算得到；$\boldsymbol{W}_{\text{LSTM}}$ 为 LSTM 的可学习网络权重参数。考虑到 LSTM 的输入是 $N \times d$ 维的片段特征，所以选择 1 层结构的 LSTM 就可以提取出相关的时间维度特征信息。

1.3　片段帧注意力特征提取

1.3.1　位置嵌入编码

相比 LSTM 网络结构自身存在输入的位置信

息,Transformer 无法主动为片段中的每帧图像添加相应的位置信息。因此,本文采用位置嵌入(position embedding)编码方法为输入的每帧图像特征添加对应的位置信息。给定片段特征 $\boldsymbol{G}\in\mathbf{R}^{N\times d}$,随机生成一个位置矩阵 $\boldsymbol{M}\in\mathbf{R}^{N\times d}$,具体的位置编码过程为

$$\boldsymbol{X}_{i,j}=\boldsymbol{G}_{i,j}\oplus\boldsymbol{M}_{i,j} \tag{9}$$

式中,$i\in[1,N]$,$j\in[1,d]$;位置矩阵 $\boldsymbol{M}_{i,j}$ 随训练过程更新;$\boldsymbol{X}\in\mathbf{R}^{N\times d}$为带有位置信息的片段特征;⊕代表按元素相加。

1.3.2 Transformer 模块设计

Transformer 模块的网络结构如图 2 所示。本文方法采用类似 Dosovitskiy 等人(2021)方法中的 Transformer 模块。该模块主要由 1 个线性映射层,1 个可迭代的 MHA(multi-head attention)层,1 个由 2 个全连接层和 2 个 GELU(Gaussian error linear unit)激活函数组成的多层感知器(multilayer perceptron, MLP),最后输出一个片段帧注意力特征 $\boldsymbol{Z}_{\text{class}}\in\mathbf{R}^{1\times d}$,具体计算为

$$\boldsymbol{X}=C_{\text{concat}}^{(v)}(\boldsymbol{X}_{\text{class}},\boldsymbol{X}_1,\cdots,\boldsymbol{X}_N) \tag{10}$$

$$\boldsymbol{X}^l=MHA(LP(\boldsymbol{X}^{l-1}))+\boldsymbol{X}^{l-1},\ l\in[1,L] \tag{11}$$

$$\boldsymbol{Z}_{\text{class}}=LN(MLP(\boldsymbol{X}^L(0))+\boldsymbol{X}^L(0)) \tag{12}$$

式中,$C_{\text{concat}}^{(v)}$表示沿竖直方向进行级联操作,$MHA(\cdot)$表示多层注意力机制计算函数;$LP(\cdot)$表示线性映射层;$LN(\cdot)$表示层归一化;$\boldsymbol{X}^{l-1}$和 $\boldsymbol{X}^l$表示上一层的注意力矩阵和当前层的注意力矩阵;当式(11)迭代 L 次即可输出注意力矩阵 $\boldsymbol{X}^L$(L 表示 Transformer 的层数),并将第 1 行的特征向量 $\boldsymbol{X}^L(0)$作为下一个 MLP 的输入,最后输出片段帧注意力特征 $\boldsymbol{Z}_{\text{class}}$。

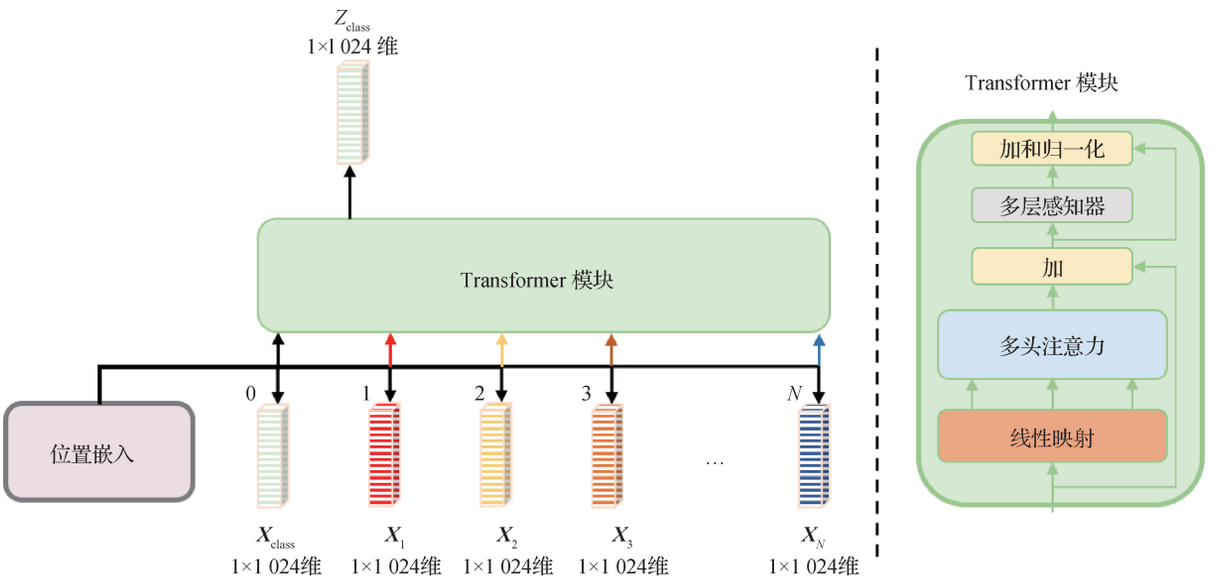


图 2 Transformer 模块结构示意图

Fig. 2 Schematic diagram of Transformer module structure

1.3.3 多头注意力机制

多头注意力机制(multi-head attention, MHA)最初主要用于处理文本序列数据,Dosovitskiy 等人(2021)将其首次用于图像识别。受此启发,如图 3 所示,本文将带有帧注意力特征 $\boldsymbol{X}_{\text{class}}$ 的片段特征 $\boldsymbol{X}\in\mathbf{R}^{(N+1)\times d}$沿水平方向分成 z 个矩阵 $\boldsymbol{X}'_i\in\mathbf{R}^{(N+1)\times\left(\frac{d}{z}\right)}$,即

$$\boldsymbol{X}=C_{\text{concat}}^{(h)}(\boldsymbol{X}'_1,\boldsymbol{X}'_2,\cdots,\boldsymbol{X}'_z) \tag{13}$$

式中,函数 $C_{\text{concat}}^{(h)}(\cdot)$表示沿水平方向进行级联操作,通过训练更新后的帧注意力特征表示为 $\boldsymbol{Z}_{\text{class}}\in\mathbf{R}^{1\times d}$,计算为

$$\boldsymbol{Z}=C_{\text{concat}}^{(v)}(\boldsymbol{Z}_{\text{class}},\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_N) \tag{14}$$

$$\begin{cases}\boldsymbol{Q}_i=\boldsymbol{X}'_i\cdot\boldsymbol{W}_i^Q\\ \boldsymbol{K}_i=\boldsymbol{X}'_i\cdot\boldsymbol{W}_i^K\\ \boldsymbol{V}_i=\boldsymbol{X}'_i\cdot\boldsymbol{W}_i^V\end{cases} \tag{15}$$

$$\boldsymbol{h}_i=f_{\text{softmax}}\left(\frac{\boldsymbol{Q}_i\cdot\boldsymbol{K}_i^{\text{T}}}{\sqrt{d_k}}\right)\cdot\boldsymbol{V}_i \tag{16}$$

$$\boldsymbol{h}_{\text{all}}=C_{\text{concat}}^{(h)}(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_z) \tag{17}$$

$$\boldsymbol{Z}=\boldsymbol{h}_{\text{all}}\cdot\boldsymbol{W}_{\text{out}} \tag{18}$$

式中,$\boldsymbol{Z}_i\in\mathbf{R}^{1\times d}$表示注意力分数;$\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i\in\mathbf{R}^{N\times\left(\frac{d}{z}\right)}$

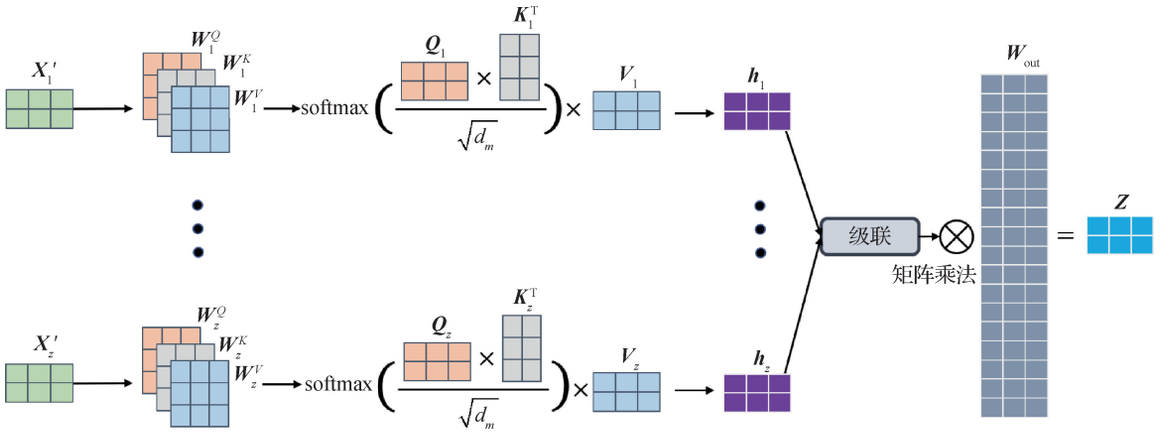


图 3　多头注意力机制计算原理图

Fig. 3　Multi-head attention mechanism calculation principle diagram

分别表示注意力矩阵对应的查询量(query)、键(key)和值(value);$\boldsymbol{h}_i$ 表示注意力矩阵;z 表示注意力头数;$\boldsymbol{W}_{\text{out}}$ 表示可学习的权重参数;$f_{\text{softmax}}(\cdot)$ 表示 softmax 函数。

2　实验及分析

2.1　实验参数设置

实验平台为显存 24 GB 的 NVIDIA GPU。识别模型训练时,batch 设为 4,学习速率开始设为 5×10^{-5},最大循环次数设为 80。实验测试采用与测试对象无关的交叉验证方法。对超过 10 个人的 BAUM-1s 数据集平均分成 5 组,进行 5 次交叉验证,对包含 8 个人的 RML 数据集采用 8 次交叉验证,最后取所有交叉验证结果的平均准确率作为实验的最终结果。

2.2　数据集

实验在视频表情数据集 BAUM-1s(Zhalehpour 等,2017)和 RML(Wang 等,2012)上进行。BAUM-1s 是一个自然情感数据集,由 31 个人的 8 种基本表情组成,共 1 222 个视频片段。实验只采用其中 6 种基本表情,分别为生气(anger)、厌恶(disgust)、害怕(fear)、高兴(joy)、悲伤(sadness)和惊奇(surprise),共 520 个视频片段。视频中每帧图像的原始分辨率为 720×576×3。图 4 为 BAUM-1s 数据集中的人脸图像样例。RML 是一个模拟情感数据集,由不同国家的 8 个人组成,共 720 个视频片段,包含生气、厌恶、害怕、高兴、悲伤和惊奇 6 种基本表情,视频中每帧图像的分辨率为 720×480×3。图 5 是 RML 数据集中的人脸图像样例。

图 4　BAUM-1s 数据集中的人脸图像样例

Fig. 4　Samples of face images from the BAUM-1s dataset

图 5　RML 数据集中的人脸图像样例

Fig. 5　Samples of face images from the RML dataset

2.3　数据预处理

由于视频数据集的每个视频时长不一致,所以需要对数据集中的每个视频进行片段化处理。对每个视频样本采用 MTCNN(multi-task convolutional neural network)网络模型(Zhang 等,2016)进行人脸检测,首先将输入的图像进行不同尺寸的缩放,形成一个图像金字塔。图像中小的人脸通过放大进行检测,大的人脸通过缩小进行检测,从而对统一大小的人脸图像进行检测。然后将所有检测到的人脸图像的维度采样为 112×112×3,再将视频的所有帧按间隔数 φ 提取到片段中,形成 φ 个包含 16 帧的片段。间隔数 φ 计算为

$$\varphi = [n/16] \tag{19}$$

式中,[ ]为取整函数;n 表示一个视频的所有帧数。

2.4 消融实验及分析

本文模型主要由 CNN、LSTM 和 Transformer 3 个模块组成。为了验证各模块的有效性,在 BAUM-1s 和 RML 数据集上按 CNN 与 LSTM 组合、CNN 与 Transformer 组合及 CNN 与 LSTM 和 Transformer 组合进行 3 组实验,消融实验结果如表 1 所示。可以看出,CNN + LSTM 模型在 BAUM-1s 和 RML 数据集上分别取得了 57.73% 和 70.20% 的表情识别准确率,CNN + Transformer 模型在 BAUM-1s 和 RML 数据集上分别取得 59.04% 和 74.96% 的准确率,明显优于 CNN + LSTM。由此可知,结合 Transformer 模型学习到的帧注意力特征的重要性大于时间维度特征。结合 CNN + LSTM + Transformer 模型表现最好,在 BAUM-1s 和 RML 数据集上分别取得 60.72% 和 75.44% 的准确率。说明 CNN 学习到的片段空间特征、LSTM 学习到的时间维度特征与 Transformer 学习到的注意力特征三者存在一定互补性,三者结合能有效改善视频表情识别的性能。

表 1 消融实验对比结果

Table 1 Comparison results of ablation experiments

/%

识别模型	BAUM-1s	RML
CNN + LSTM	57.73	70.20
CNN + Transformer	59.04	74.96
CNN + LSTM + Transformer	**60.72**	**75.44**

注:加粗字体表示各列最优结果。

2.5 与其他现有方法的对比

为了验证本文方法的性能,与其他方法进行对比,结果如表 2 所示,其中对比方法结果为文献数据。Zhang 等人(2018)采用深度 3D-CNN 提取人脸表情特征用于人脸表表情识别,在 BAUM-1s 和 RML 数据集上分别取得 50.11% 和 68.09% 的识别准确率。Kansizoglou 等人(2019)采用 CNN 与 LSTM 组合进行人脸表情特征提取,在 BAUM-1s 和 RML 上分别取得 55.36% 和 70.55% 的准确率。Cornejo 和 Pedrini(2019)首先采用局部约束的立体匹配算法对人脸图像进行处理,即 census 变换,用于减少光照差异引起的误匹配,然后采用基于 CNN 的 census 变换方法提取人脸表情特征,最后使用逻辑回归分类器进行人脸表情识别,在 BAUM-1s 和 RML 数据集上分别取得 59.52% 和 75% 的准确率。Ma 等人(2019)首先利用 3D-CNN 从面部表情序列中提取特征,最后使用支持向量机(support vector machine,SVM)进行表情分类,在 BAUM-1s 和 RML 数据集上分别获得 54.69% 和 73.88% 的准确率。本文方法在 BAUM-1s 和 RML 数据集上分别获得 60.72% 和 75.44% 的平均准确率,优于现有其他方法。实验结果表明,本文提出的模型能有效进行视频人脸表情识别。

表 2 不同方法在 BAUM-1s 和 RML 数据集上的准确率对比

Table 2 Comparison of accuracy among different methods on baum-1s and RML datasets

方法	BAUM-1s 数据集		RML 数据集	
	视频特征	准确率/%	视频特征	准确率/%
Zhang 等人(2018)	3D-CNN	50.11	Vnet	68.09
Kansizoglou 等人(2019)	CNN + LSTM	55.36	CNN + LSTM	70.55
Ma 等人(2019)	3D-CNN	54.69	3D-CNN	73.88
Cornejo 和 Pedrini(2019)	CT based VGG	59.52	CT based VGG	75.00
本文	CNN + LSTM + Transformer	**60.72**	CNN + LSTM + Transformer	**75.44**

注:加粗字体表示各列最优结果。

为了更直观地观察本文方法对各类表情的识别情况,图 6 和图 7 分别给出了本文方法在 BAUM-1s 和 RML 数据集上获得最佳识别结果的混淆矩阵。由图 6 可见,高兴和悲伤的识别效果比较好,正确识别率分别为 85.47% 和 78.36%,而生气和害怕的识别准确率较低,分别为 8.93% 和 13.51%,这两种表情容易误判为悲伤,原因可能是这 3 种表情的区分度不高,造成网络模型误判。由图 7 可以看出,害怕

的识别性能最低，正确识别率为 37.5%，生气的正确识别率为 65%，其他表情的识别效果较好，正确识别率超过 70%。原因可能是 RML 数据集中害怕表情的样本数目比其他表情样本数目少很多，致使网络模型无法较好地识别此类表情。

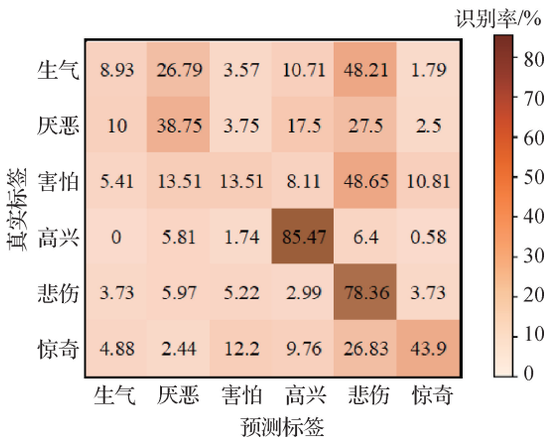


图 6　BAUM-1s 数据集中 CNN + LSTM + Transformer 模型的识别结果混淆矩阵

Fig. 6　Confusion matrix of the recognition results of the CNN + LSTM + Transformer model in the BAUM-1s dataset

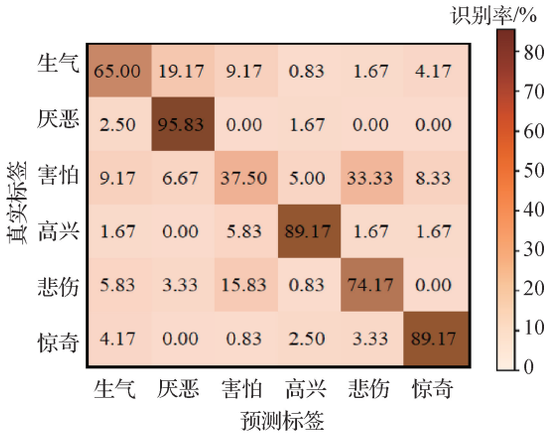


图 7　RML 数据集中 CNN + LSTM + Transformer 模型的识别结果混淆矩阵

Fig. 7　Confusion matrix of the recognition results of the CNN + LSTM + Transformer model in the RML dataset

3　结　论

本文提出一种基于 Transformer 的视频序列表情识别模型，用于实现视频人脸表情识别。该方法集成了 CNN、LSTM 和 Transformer 模型，分别用于学习高层次的视频片段空间特征、视频片段时间特征以及视频片段帧注意力特征。在 BAUM-1s 和 RML 两个视频情感数据集上的实验结果表明，本文方法能有效改善视频人脸表情识别模型的性能。与对比方法相比，本文方法的整体性能占优，但识别某些数据样本数量较少的表情类别依然表现欠佳，并且在面对难以区分的两种表情时，如害怕和悲伤两种表情，本文方法表现一般。在以后的工作中，将尝试通过采用有效的数据增强技术扩大数据集，建立更具鲁棒性的深度注意力机制模型。同时，根据各类人脸表情存在的差异性，构建一个区分表情注意力模块，以进一步提高区分度不高的表情识别性能。

参考文献(References)

Ahad M A R, Tan J K, Kim H and Ishikawa S. 2012. Motion history image: its variants and applications. Machine Vision and Applications, 23(2): 255-281 [DOI: 10.1007/s00138-010-0298-4]

Anandan P. 1989. A computational framework and an algorithm for the measurement of visual motion. International Journal of Computer Vision, 2(3): 283-310 [DOI: 10.1007/BF00158167]

Bahdanau D, Cho K and Bengio Y. 2016. Neural machine translation by jointly learning to align and translate [EB/OL]. [2021-05-21]. https://arxiv.org/pdf/1409.0473.pdf

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal A, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D. 2020. Language models are few-shot learners [EB/OL]. [2020-07-22]. https://arxiv.org/pdf/2005.4165.pdf

Cornejo J Y R and Pedrini H. 2019. Audio-visual emotion recognition using a hybrid deep convolutional neural network based on census transform//Proceedings of 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). Bari, Italy: IEEE: 3396-3402 [DOI: 10.1109/SMC.2019.8914193]

De Silva L C and Ng P C. 2000. Bimodal emotion recognition//Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition. Grenoble, France: IEEE: 332-335 [DOI: 10.1109/AFGR.2000.840655]

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words:

transformers for image recognition at scale. [EB/OL]. [2021-06-03]. https://arxiv.org/pdf/2010.11929.pdf

Fan Y, Lu X J, Li D and Liu Y L. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks//Proceedings of the 18th ACM International Conference on Multimodal Interaction. Tokyo, Japan: Association for Computing Machinery: 445-450 [DOI: 10.1145/2993148.2997632]

Fu X F, Fu X J, Li J J and Yu Z S. 2015. Facial expression recognition using multi-scale spatiotemporal local orientational pattern histogram projection in video sequences. Journal of Computer-Aided Design and Computer Graphics, 27(6): 1060-1066 (付晓峰, 付晓鹃, 李建军, 余正生. 2015. 视频序列中基于多尺度时空局部方向角模式直方图映射的表情识别. 计算机辅助设计与图形学学报, 27(6): 1060-1066)

Hu M, Wang H W, Wang X H, Yang J and Wang R G. 2019. Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks. Journal of Visual Communication and Image Representation, 59: 176-185 [DOI: 10.1016/j.jvcir.2018.12.039]

Kansizoglou I, Bampis L and Gasteratos A. 2022. An active learning paradigm for online audio-visual emotion recognition. IEEE Transactions on Affective Computing, 13(2): 756-768 [DOI: 10.1109/TAFFC.2019.2961089]

Li J, Jin K, Zhou D L, Kubota N and Ju Z J. 2020. Attention mechanism-based CNN for facial expression recognition. Neurocomputing, 411: 340-350 [DOI: 10.1016/j.neucom.2020.06.014]

Li S and Deng W H. 2020. Deep facial expression recognition: a survey. Journal of Image and Graphics, 25(11): 20-34 (李珊, 邓伟洪. 2020. 深度人脸表情识别研究进展. 中国图象图形学报, 25(11): 20-34) [DOI: 10.11834/jig.200233]

Li Y, Zeng J B, Shan S G and Chen X L. 2019. Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, 28(5): 2439-2450 [DOI: 10.1109/TIP.2018.2886767]

Li Y D, Hao Z B and Lei H. 2016. Survey of convolutional neural network. Journal of Computer Applications, 36(9): 2508-2515, 2565 (李彦冬, 郝宗波, 雷航. 2016. 卷积神经网络研究综述. 计算机应用, 36(9): 2508-2515, 2565) [DOI: 10.11772/j.issn.1001-9081.2016.09.2508]

Liu S S, Tian Y T and Wan C. 2011. Facial expression recognition method based on gabor multi-orientation features fusion and block histogram. Acta Automatica Sinica, 37(12): 1455-1463 (刘帅师, 田彦涛, 万川. 2011. 基于 Gabor 多方向特征融合与分块直方图的人脸表情识别方法. 自动化学报, 37(12): 1455-1463) [DOI: 10.3724/SP.J.1004.2011.01455]

Ma Y X, Hao Y X, Chen M, Chen J C, Lu P and Košir A. 2019. Audio-visual emotion fusion (AVEF): a deep efficient weighted approach. Information Fusion, 46: 184-192 [DOI: 10.1016/j.inffus.2018.06.003]

Sun B, Wei Q L, Li L D, Xu Q H, He J and Yu L J. 2016. LSTM for dynamic emotion and group emotion recognition in the wild//Proceedings of the 18th ACM International Conference on Multimodal Interaction. Tokyo, Japan: Association for Computing Machinery: 451-457 [DOI: 10.1145/2993148.2997640]

Wang S M, Shuai H and Liu Q S. 2020. Facial expression recognition based on deep facial landmark features. Journal of Image and Graphics, 25(4): 813-823 (王善敏, 帅惠, 刘青山. 2020. 关键点深度特征驱动人脸表情识别. 中国图象图形学报, 25(4): 813-823) [DOI: 10.11834/jig.190331]

Wang X H, Pan L J, Peng M Z, Hu M, Jin C H and Ren F J. 2020. Video emotion recognition based on hierarchical attention model. Journal of Computer-Aided Design and Computer Graphics, 32(1): 27-35 (王晓华, 潘丽娟, 彭穆子, 胡敏, 金春花, 任福继. 2020. 基于层级注意力模型的视频序列表情识别. 计算机辅助设计与图形学学报, 32(1): 27-35) [DOI: 10.3724/SP.J.1089.2020.17719]

Wang Y J, Guan L and Venetsanopoulos A N. 2012. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. IEEE Transactions on Multimedia, 14(3): 597-607 [DOI: 10.1109/TMM.2012.2189550]

Zhalehpour S, Onder O, Akhtar Z and Erdem C E. 2017. BAUM-1: a spontaneous audio-visual face database of affective and mental states. IEEE Transactions on Affective Computing, 8(3): 300-313 [DOI: 10.1109/TAFFC.2016.2553038]

Zhang K P, Zhang Z P, Li Z F and Qiao Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10): 1499-1503 [DOI: 10.1109/LSP.2016.2603342]

Zhang S Q, Zhang S L, Huang T J, Gao W and Tian Q. 2018. Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology, 28(10): 3030-3043 [DOI: 10.1109/TCSVT.2017.2719043]

作者简介

陈港,男,硕士研究生,主要研究方向为图像处理和情感计算。E-mail: 904699855@qq.com

赵小明,通信作者,男,教授,主要研究方向为模式识别和情感计算。E-mail: tzxyzxm@163.com

张石清,男,教授,主要研究方向为模式识别和情感计算。E-mail: tzczsq@163.com