智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

融合关键区域信息的双流网络视频表情识别

孔英会, 崔文婷, 张珂, 车辚辚

引用本文:
孔英会, 崔文婷, 张珂, 等. 融合关键区域信息的双流网络视频表情识别[J]. 智能系统学报, 2025, 20(3): 658-669.
KONG Yinghui, CUI Wenting, ZHANG Ke, et al. Two-stream network video expression recognition by fusing key region information[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(3): 658-669.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202401031

DOI: 10.11992/tis.202401031
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250428.1322.006

融合关键区域信息的双流网络视频表情识别

孔英会[1,2], 崔文婷[1], 张珂[1,2], 车辚辚[1,2]
（1. 华北电力大学 电子与通信工程系, 河北 保定 071003; 2. 华北电力大学 河北省电力物联网技术重点实验室, 河北 保定 071003）

摘　要: 人脸表情识别是计算机视觉领域中的一个重要研究课题, 而视频中的表情识别在很多场景下具有实用价值。视频序列包含丰富的帧内空间信息与帧间时间信息, 同时面部关键区域的提取也对表情识别结果有重要影响, 本文提出一种融合关键区域信息的双流网络表情识别方法。构建空间−时间双流网络, 其中空间网络分支结合面部运动单元和 CSFA(channel-spatial frame attention), 重点关注影响表情识别结果的面部关键区域, 以实现空间特征的有效提取; 时间分支通过 Farneback 提取光流获得帧间的表情运动信息, 并借助空间关键区域掩模选取降低光流计算复杂度。对空间−时间双流网络识别结果进行决策融合, 得到最终视频表情识别结果。该方法在 eNTERFACE'05、CK+数据集上进行实验测试, 结果表明本文所提方法可有效提升识别精度, 且提高了运行效率。

关键词: 视频表情识别; 双流网络; 注意力机制; 光流; 卷积神经网络; 掩模; 特征融合; 面部表情识别
中图分类号: TP39　　**文献标志码:** A　　**文章编号:** 1673−4785(2025)03−0658−12

中文引用格式: 孔英会, 崔文婷, 张珂, 等. 融合关键区域信息的双流网络视频表情识别 [J]. 智能系统学报, 2025, 20(3): 658–669.
英文引用格式: KONG Yinghui, CUI Wenting, ZHANG Ke, et al. Two-stream network video expression recognition by fusing key region information[J]. CAAI transactions on intelligent systems, 2025, 20(3): 658–669.

Two-stream network video expression recognition by fusing key region information

KONG Yinghui[1,2], CUI Wenting[1], ZHANG Ke[1,2], CHE Linlin[1,2]
(1. Department of Electronic and Communication Engineering, North China Electric Power University, Baoding 071003, China; 2. Hebei Key Laboratory of Power Internet of Things Technology, North China Electric Power University, Baoding 071003, China)

Abstract: Facial expression recognition is an important research topic in the field of computer vision, and facial expression recognition in video has practical value in many scenes. Video sequences contain rich intra-frame spatial information and inter-frame temporal information, and key facial regions also have an important impact on the expression recognition results. This paper proposes a two-stream network expression recognition method by fusing key region information. First, a spatial-temporal two-stream network is constructed. The spatial network branch combines the facial motion unit and the CSFA attention mechanism to focus on the key facial regions that affect the expression recognition results, so as to realize the effective extraction of spatial features. The temporal branch extracts the optical flow through Farneback to obtain the expression motion information between frames and uses the spatial key region mask selection to reduce the computational complexity of optical flow. Finally, the final video expression recognition results are obtained by decision fusion of the spatial-temporal two-stream network recognition results. The method is tested on the eNTERFACE'05 and CK+ datasets. The results show that the proposed method can effectively improve the recognition accuracy and operating efficiency.

Keywords: video expression recognition; two-stream network; attention mechanism; optical flow; convolutional neural networks; mask; feature fusion; facial expression recognition

面部表情[1]是人类重要的沟通方式之一, 随着视频监控设备的普及和计算机技术的迅速发展, 基于视频数据的表情识别具有广泛的应用前

收稿日期: 2024−01−24.　**网络出版日期:** 2025−04−28.
基金项目: 国家自然科学基金项目 (62076093).
通信作者: 孔英会. E-mail: kongyh2005@163.com.

景, 在智能监控、人机交互、医疗健康等领域发挥着越来越重要的作用。面部表情识别 FER(facial expression recognition) 方法主要分为基于手工提取特征和基于深度学习的表情识别方法。传统的基于手工提取特征的方法效率低、鲁棒性差[2-9], 基于深度学习[10] 的方法成为研究热点。针对视频表情识别任务, 视频序列既包括帧内空间全局与局部表情相关关键信息, 又包括帧间表情变化过程的帧间运动时间信息, 因此融合帧内、帧间信息的双流网络视频表情识别的方法[10-12] 被广泛采用。

双流网络的构建有不同的方法, Zhang 等[10] 融合了 PHRNN(part-based hierarchical bidirectional recurrent neural network) 时间网络和 MSCNN (multi-signal convolutional neural network) 空间网络, 以提取 FER 的部分和整体、几何外观和静态/动态特征信息, 最终通过两个网络融合, 在 CK+、Oulu-CASIA 及 MMI 数据集上识别准确度分别为 98.50%、86.25% 和 81.18%, 较当时其他方法有更高识别率, 但训练和计算资源消耗较大, 泛化能力也有待进一步验证。Feng 等[11] 提出了一种新的双流卷积神经网络 (convolutional neural network, CNN) 架构: 空间流以最后表情帧作为输入提取面部图像信息; 时间流对输入的图像序列提取 LBP-TOP(local binary pattern-three orthogonal planes) 特征, 最终将两个流 CNN 输出结果平均作为网络最终输出结果。通过 CK+数据集证明其有效性, 但该双流网络对面部关键区域感知这一部分并未展开研究。Chen 等[12] 提出一种通过分析单幅表情峰值图像和视频序列中面部关键点轨迹的空域特征和时域特征双流网络, 最后采用微调的特征融合策略取得了最优的时域特征和空域特征融合效果。该方法在 CK+、MMI 和 Oulu-CASIA 数据集上的识别准确率分别为 98.46%、82.96% 和 87.12%, 接近或超越了当前同类表情识别方法中的最好性能。

除了双流网络视频表情识别方法外, 还有基于 RNN、LSTM(long short-term memory) 提取空间时间以及语音等更多模态信息的表情识别方法。Hasani 等[13] 提出了一种用于视频面部表情识别的 3D 卷积神经网络方法, 其网络体系结构由 3D Inception-ResNet 层和 LSTM 单元组成, 共同提取面部图像中的空间关系以及视频中不同帧之间的时间关系。该网络将面部关键点用作输入, 强调对面部表情有重大贡献的面部关键区域。在公开 CK+、MMI、FERA 和 DISFA 数据集上进行了评估, 表明所提出的方法优于许多最先进的方法, 但该方法也存在着运算量大、识别速度慢的缺点。张隽睿[14] 建立基于 ConvLSTM 的视频表情识别模型并加入了注意力模块, 使模型提取特征图的时间和空间特征的能力提升, 但在 eNTERFACE'05 数据集上的识别准确度并不好。刘菁菁等[15] 设计了基于长短时记忆网络的多模态情感识别模型, 采用双路 LSTM 分别模拟人类听觉和视觉处理通路处理语音和面部表情的情感信息, 并模拟人脑边缘系统情感区进行决策层加权特征融合, 在 eNTERFACE'05 数据集上的准确度为 74.7%。Farhoudi 等[16] 开发了基于 Mobel 模型的深度学习特征视听融合模型, 视觉模态使用 3D-CNN 模型学习视觉表达的时空特征, 而听觉模态将语音信号的 Mel 谱图输入卷积递归神经网络 (convolutional recurrent neural network, CRNN) 进行时空音频特征提取, 在 eNTERFACE'05 数据集上准确度为 62%。多模态表情识别确实可以实现较好的识别准确度, 但数据集需要额外获取语音等模态信息, 实际应用受限制。

注意力机制对网络选择性聚焦关键区域特征信息具有重要作用。Fernandez 等[17] 提出一种带有注意力机制的人脸表情识别网络结构, 该网络模型的注意力机制模块采用基于 U-Net[18] 的网络结构对输入的人脸图像进行人脸分割并输出一张掩模, 再把特征提取模块输出的特征图张量和掩模进行融合, 以去除特征图张量中与表情无关的特征, 从而取得更好的表情分类效果, 相较未添加注意力机制模块的模型, 识别准确度提升了 1.25%。Meng 等[19] 设计了帧注意网络 (frame attention network, FAN) 来进行视频表情识别, 在端到端框架中识别一些有区别的帧, 去除与表情识别无关的冗余信息; 并嵌入帧注意力计算模块以学习多个注意力权重, 用于自适应地聚合特征向量以形成单个具有判别力的视频特征表示。与其他基于 CNN 的方法相比, 在 CK+和 AFEW8.0 数据集上的识别准确度分别为 99.69% 和 51.18%, 显示出帧注意网络的卓越性, 但该方法存在着侧重提取全局特征, 并没有考虑局部关键区域的问题。李同霞[20] 提出了一种基于帧注意力机制的表征流嵌入网络 (residual network embedding representation flow with frame attention), 在基于表征流嵌入的深度残差网络基础上加入了帧注意力模块, 经实验验证在 AFEW 和 eNTERFACE'05 数据集上也达到了较好的效果, 准确率分别为 47.51% 和 55.42%。

通过分析上述文献发现, 实验室环境下人为

表情数据 (如 CK+数据集) 识别准确度比较高，但对于实验室环境下自发表情 (如 eNTERFACE'05 数据集)，由于情感强度变化并不明显，其识别仍具有挑战性。

综上所述，融合时间空间信息的双流网络视频表情识别方法取得了不错的识别效果，但如何准确感知关键区域、去除冗余信息的影响、降低计算量、提升有挑战数据集的识别精度仍是需要解决的关键问题。本文提出一种融合关键区域信息的双流网络视频表情识别方法，首先构建空间−时间双流网络，空间网络分支结合面部运动单元和 CSFA(channel-spatial frame attention) 提取利于表情识别的关键区域，以实现空间特征的有效提取；其次，在通过 Farneback 光流提取帧间的表情运动信息基础上，利用空间关键区域掩模控制光流提取区域，从而减少计算量；最后，利用空间、时间网络决策融合得到最终识别结果。在 eNTERFACE'05 和 CK+数据集上进行实验，验证本文方法有效性。

1　提出的方法与模型

本文提出的融合关键区域信息的双流网络识别方法包括空间网络和时间网络两个分支，如图 1 所示。其中，空间网络分支使用 CSFA，以通道空间注意力 (channel-spatial attention, CSA) 模块提取空间全局和局部特征。ResNet-18 以视频表情帧作为输入，从中提取表情关键特征信息，并且通过帧注意 (frame attention, FA) 模块实现显著视频帧的有效选择。在时间网络分支中，以光流图作为输入提取时间特征，采用卷积神经网络结构，进一步提取视频帧间的表情变化特征。最后，对空间、时间识别结果进行决策融合。

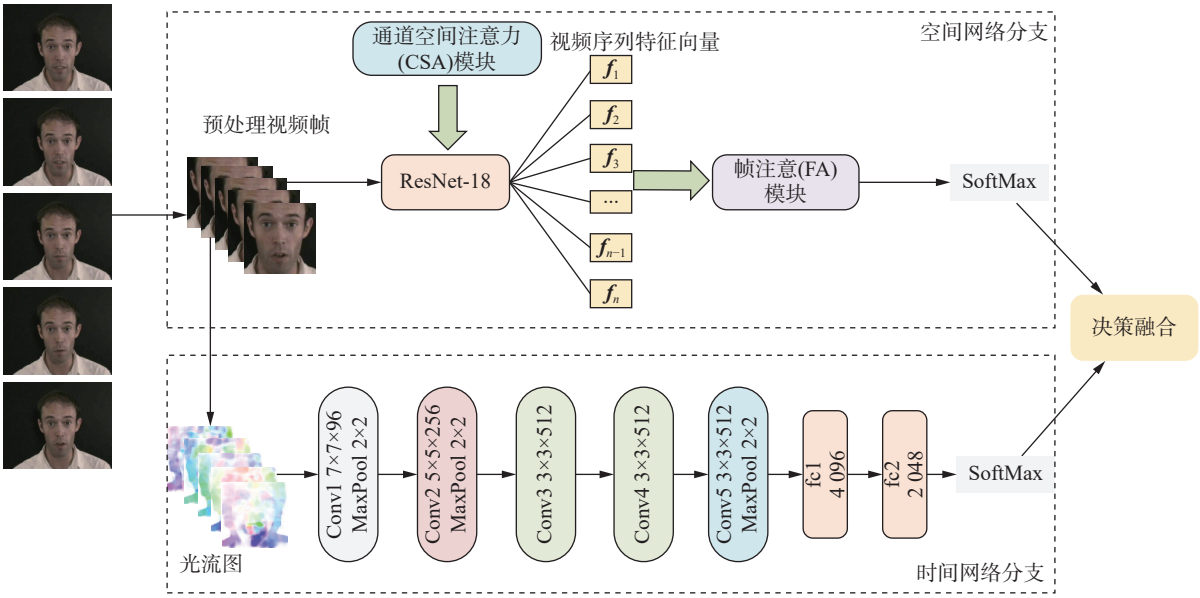


图 1　融合关键区域信息的双流网络表情识别方法框架

Fig. 1　Framework of two-stream network expression recognition method with key region information

图 1 的工作过程如下：首先通过定位和标记确定视频中的人脸位置，再进行人脸对齐、人脸归一化及数据增强操作的预处理。空间网络分支对预处理后的视频数据进行帧采样，将视频分割成 z 个片段，然后从每个片段中随机选择一帧，以 M 个视频帧作为 2D 卷积神经网络的输入，其中 $z \neq M$，z 代表视频分成片段数、M 代表网络的输入帧数。并在 ResNet-18 添加 CSA 关注重要面部信息，即关键区域与面部肌肉运动区域。再通过 FA 模块学习两级注意权重，用于自适应地聚合特征向量，以形成单个有判别力的视频表示。光流特征分支基于空间特征信息进行目标掩模提取图像中感兴趣的区域，再利用 Farneback 光流法计算出帧之间的光流，将得到的光流送入时间卷积网络提取特征并得到识别结果。最后通过最大概率的决策融合得到网络最终识别结果。融合函数为

$$W(x)=\sum_{i=1}^{2}\alpha_i(C_i(x)+P_i(x)) \tag{1}$$

式中：α_i=0.5；$P_i(x)(0<P_i(x)<1)$ 是表情预测概率，它是空间分支网络和时间分支网络中 SoftMax 层的输出；$C_i(x)$是表情类别的预测排序，公式为

$$C_i(x_1),C_i(x_2),\cdots,C_i(x_n)=\mathrm{sort}(P_i(x_1),P_i(x_2),\cdots,P_i(x_n)) \tag{2}$$

式中 n 表示表情类别的总数。基于 $P_i(x)$的值对 n 个表情类别进行排序。换句话说，较大的 $P_i(x)$

对应于较大的$C_i(x)(C_i(x)\in\{1,2,\cdots,n\})$。式 (1) 是指当网络模型预估面部表情时, 会优先考虑预测排序。如果网络不能根据$C_i(x)$的表情预测类别值决定, 则将会对$P_i(x)$进行比较, 其结果作为双流网络的输出, 以此确定模型融合后的最终判别结果。

为方便理解和验证, 下面给出融合关键区域信息的双流网络表情识别方法的伪代码, 其仅概述了整体流程, 并未完全展开每一个细节函数的具体实现。

融合关键区域信息的双流网络表情识别方法伪代码

```
def preprocess_video(self, video_frames):
# 定位和标记人脸, 进行对齐、归一化等预处理操作
preprocessed_frames = preprocess(video_frames)
def extract_spatial_features(self, preprocessed_frames):
# 使用空间网络分支提取关键区域的空间特征
spatial_features = self.spatial_model(preprocessed_frames)
def compute_temporal_features(self, preprocessed_frames):
# 计算光流并结合关键区域掩模减少计算量
optical_flow = compute_optical_flow(preprocessed_frames)
masked_optical_flow = apply_mask(optical_flow, key_region_masks)
temporal_features = self.temporal_model(masked_optical_flow)
def fuse_features(self, spatial_features, temporal_features):
# 双流特征融合
fused_features = self.fusion_method(spatial_features, temporal_features)
def recognize_expression(self, fused_features):
# 识别表情类别
expression_prediction = classify_expression(fused_features)
# 定义双流网络类
class DualStreamNetwork:
def __init__(self, spatial_model, temporal_model, fusion_method):
self.spatial_model = spatial_model # 空间分支模型
self.temporal_model = temporal_model # 时间分支模型
self.fusion_method = fusion_method # 融合方法
def train_and_evaluate(self, dataset, epochs, optimizer, criterion):
# 训练和验证过程
for _ in range(epochs):
for data in dataset:
# 执行前向传播、损失计算、反向传播和优化步骤
pass
# 计算和记录性能指标
pass
```

1.1　基于 CSFA 的表情识别空间子网络构建

为提升性能并降低训练成本, 选择 ResNet-18[21] 网络作为骨干网络进行提取特征。空间注意力模块包含 2 部分: CSA 模块和 FA 模块。CSA 模块应用于提取特征向量部分, 增加对人脸中与表情相关区域的关注。FA 模块在基于全局和局部关键区域特征向量的基础上进行帧聚合, 以实现帧间信息融合。

CSA 注意力由 ECA(efficient channel attention) 模块和 SA(spatial attention) 模块两部分构成。ECA[22] 模块是一种有效的通道注意模块, 可以实现适当的跨信道交互, 进而显著降低模型复杂度, 同时保持识别精度性能。ECA 模块采用了不降维的局部跨信道交互策略, 该策略可以通过一维卷积有效地实现。此外, ECA 模块还包含自适应选择一维 ($1\times k$) 卷积核大小的方法, 可以确定局部跨信道交互的覆盖率。通过特征通道数 C 自适应地映射 k 的取值。实践证明, 该方法保证了 ECA 机制的计算能力与运行效率。

$$w=\sigma(\mathrm{C}_{1\mathrm{D}_k}(y)) \tag{3}$$

式中: σ是 Sigmoid 函数, $\mathrm{C}_{1\mathrm{D}}$表示 1D 卷积。该模块将输入特征图进行全局平均池化操作, 再进行卷积核大小为 k 的 1 维卷积操作, 并经过 Sigmoid 激活函数得到各个通道的权重 w。最后, 将归一化权重和原输入特征图逐通道相乘, 生成加权后的特征图。

SA[23] 模块是一种空间注意模块, 可以帮助网络在学习特征表示时更加集中于图像中空间上的重要区域。该模块将 ECA 注意力输出的特征图作为输入, 基于通道的全局平均池化, 得到 2 个 $H\times W\times 1$ 的特征图, 然后将这 2 个特征图基于通道做通道拼接。然后经过一个 7×7 卷积操作, 降维为 1 个通道, 即 $H\times W\times 1$, 再经过 Sigmoid 激活函数得到空间注意力特征。最后将该特征和其输入特征相乘, 得到最终生成的特征。

两部分注意模块需要有效结合，首先利用 ECA 模块对输入特征进行加权，然后通过两个卷积层构建了一个 SA 模块，将加权后的特征传入该模块。SA 模块通过减少通道数的方式来降低计算量，并通过 ReLU 函数的非线性变换增强模块的表征能力和学习能力。同时，Sigmoid 函数计算空间注意力权重，得到 0 ~ 1 的权重。最后将 SA 权重与 ECA 加权后的特征相乘，得到最终的注意力加权特征。

ResNet-18 网络是一种 18 层深度卷积神经网络，其中包含 4 个残差块，每个残差块由两个基本块 (BasicBlock) 组成。通过残差连接和跳跃连接解决了梯度消失问题，使得网络更易训练，最终通过全连接层将特征映射到类别数量上。将 CSA 模块放在基本块部分末尾，如图 2 所示，可以在每个基本块中引入注意力机制，增强对面部特征的选择性。这有助于网络更加关注与表情识别任务相关的重要面部区域和特征，提高模型对表情变化的敏感度。

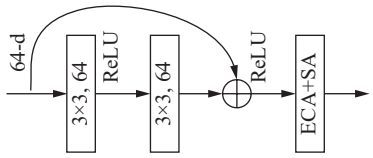


图 2　结合 ECA+SA 注意力机制基本块结构

Fig. 2　Combined ECA+SA attention mechanism BasicBlock structure

FA 模块[19] 如图 3 所示，首先视频序列特征向量 $\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n$ 通过全连接层与 Sigmoid 函数，得到每帧图像对应的自注意力权重值 $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_n$，再将所有输入帧特征聚合为全局表示 $\boldsymbol{f}'_v$，其计算式为

$$\boldsymbol{f}'_v=\frac{\sum_{i=1}^{n}\boldsymbol{\alpha}_i\boldsymbol{f}_i}{\sum_{i=1}^{n}\boldsymbol{\alpha}_i} \tag{4}$$

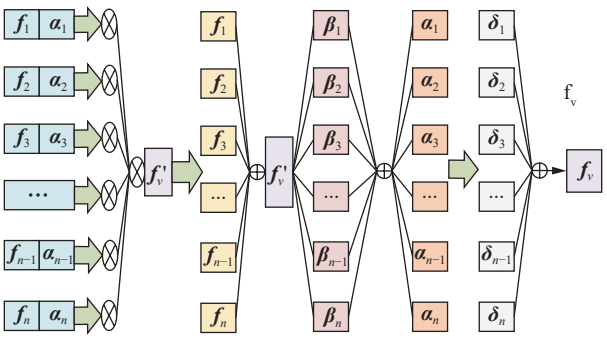


图 3　帧注意模块结构

Fig. 3　Frame attention module structure

通过对帧特征和这个全局表示的关系进行建模，得到第 i 帧的帧间信息注意力权重为

$$\boldsymbol{\beta}_i=\sigma([\boldsymbol{f}_i:\boldsymbol{f}'_v]^{\mathrm{T}}\boldsymbol{q}^1) \tag{5}$$

最后，通过自我注意和关系注意权重，将所有帧特征聚合成一个新的紧凑特征，即将所有帧的特征经过加权聚合后得到一个单一的、压缩的表示：

$$\boldsymbol{f}_v=\frac{\sum_{i=0}^{n}\boldsymbol{\alpha}_i\boldsymbol{\beta}_i[\boldsymbol{f}_i:\boldsymbol{f}'_v]}{\sum_{i=0}^{n}\boldsymbol{\alpha}_i\boldsymbol{\beta}_i} \tag{6}$$

FA 模块主要用于视频帧的帧间信息融合，通过融合各关键帧的特征向量，得到较为准确的视频表示形式，并与自注意力权重融合，权重越大表示此视频帧对整段视频的意义越显著。

1.2　基于光流特征的表情识别时间网络构建

光流[24] 是在时变图像中用于表征模式运动速度的概念。物体在运动过程中图像上对应点的亮度模式也会呈现移动，它可以捕捉并呈现图像随时间的变化，蕴含了目标运动的信息。

时间网络从预处理的视频数据帧随机抽取系列视频帧，该网络分支利用 Farneback 法[25] 计算两帧之间的光流通量，该方法是一种稠密光流法，可计算图像中每一个像素点的光流信息，从而计算出整个图像的稠密光流场。它通过金字塔多级分辨率的处理和多项式拟合的方法，能够捕捉微小的面部运动，从而捕获表情中的细微变化。可得到 x 方向和 y 方向光流图，将光流图按照时间顺序进行堆叠，形成堆叠光流图，经过多轮卷积处理及全连接处理后，在 SoftMax 逻辑回归层进行多元 L2 范式标准化处理。

密集的光流可以看作是连续帧 t 和 $t+1$ 之间的一组位移向量场 $\boldsymbol{d}_t$，通过 $\boldsymbol{d}_t(u,v)$ 表示在第 t 帧中的点 (u,v) 处的位移向量，它将该点移动到下一帧 $t+1$ 中相应的点。向量场的水平分量和垂直分量 $\boldsymbol{d}_t^x$ 和 $\boldsymbol{d}_t^y$ 可以看作是图像通道。为表示一系列帧间的运动，将 K 个连续帧的光流通道 $\boldsymbol{d}_t^{x,y}$ 堆叠起来，形成总共 $2K$ 个输入通道。设 w 和 h 是视频的宽度和高度，对于任意帧 τ 的输入卷积 $\boldsymbol{I}_\tau\in\mathbf{R}^{w\times h\times 2L}$ 的构造：

$$\boldsymbol{I}_\tau(u,v,2k-1)=\boldsymbol{d}^x_{\tau+k-1}(u,v) \tag{7}$$

$$\boldsymbol{I}_\tau(u,v,2k)=\boldsymbol{d}^y_{\tau+k-1}(u,v) \tag{8}$$

$$u=[1;w],v=[1;h],k=[1;L] \tag{9}$$

式中：输入堆叠了 K 视频帧序列的水平向量 $\boldsymbol{x}$ 和垂直向量 $\boldsymbol{y}$，τ 为位移偏量。

1.3　基于空间关键区域掩模控制的光流信息获取

为解决稠密光流法计算图像中每个像素点的光流信息而导致计算量大的问题，本文通过提取面部图像中感兴趣的区域来建立目标掩膜，从而减少光流的计算量。根据面部 AU(action unit) 运

动单元及注意力分析可知, 面部表情的运动大多集中在眼睛、鼻子、嘴巴及其周围相邻的区域, 利用这些表情识别区域既可以降低计算量、去除冗余面部信息, 也可以去除非关键区域的干扰。根据 68 个人脸关键点对关键区域标定, 以此获取这些关键点的坐标信息, 然后对原始的表情视频帧建立目标掩模, 如图 4 所示。

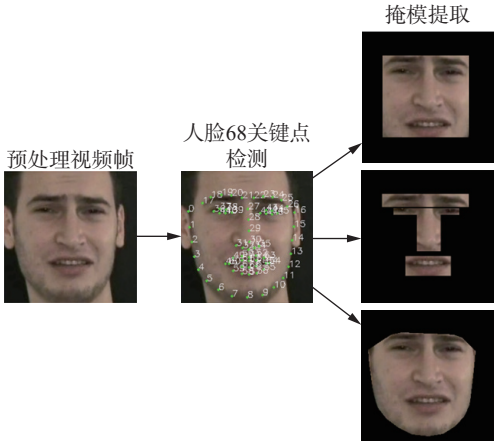


图 4　人脸关键区域掩模流程
Fig. 4　Mask process of face key region

掩模建立可以有不同方法, 但都是基于 AU 运动单元和注意力机制, 如图 4 右侧所示为 3 种不同掩模处理后的人脸实验样本效果。

1) 矩形区域 AU 单元掩模法: 根据 68 个面部关键点的位置坐标信息, 可以计算出包含眉毛、眼睛、鼻子和嘴巴区域的最小矩形框, 从而将这些重要区域提取在一个矩形框中, 以便进行进一步的分析和处理。

2) 分区单独提取 AU 单元掩模法: 根据 68 个面部关键点的位置坐标信息, 可以针对每个区域单独计算出其边界框, 将眉毛、眼睛、鼻子和嘴巴各自分别框出, 这有助于对每个区域进行更精细的分析和特征提取。

3) 面部外围轮廓关键点掩模法: 考虑到眉毛、眼睛、鼻子和嘴巴关键点附近的肌肉区域也对表情识别有重要影响, 选取包含所有关键点的面部外围轮廓以内的近似椭圆形区域, 将面部外围以外的部分去除, 只保留与表情识别相关的区域。

1.4　实验数据集

CK+数据集[26] 是一个经典的人脸表情识别数据库, 包含 123 名演员的人脸表情图像序列 (部分数据如图 5(a) 所示)。该数据集涵盖愤怒 (Anger, An)、厌恶 (Disgust, Di)、恐惧 (Fear, Fe)、快乐 (Happy, Ha)、悲伤 (Sad, Sa)、惊讶 (Surprise, Su) 和中性 (Contempt, Co)7 种情感类别, 每种类别以时间顺序展示了表情从起始到峰值的变化过程。每个表情序列都配有情感标签和强度评分, 为研究者提供准确的情感类别和表情强度信息。CK+数据集的多样性和实用性使其成为评估表情识别算法性能、进行比较和研究的重要资源, 为人脸情感识别领域的进展提供了宝贵基础。表 1 给出了 CK+数据集样本数。

(a) CK+

(b) eNTERFACE'05

图 5　表情数据集部分样本
Fig. 5　Partial samples of expression dataset

表 1　CK+数据集样本数
Table 1　Sample number of CK+ dataset

类别	愤怒	厌恶	恐惧	快乐	悲伤	惊讶	中性
样本数	45	59	25	69	28	83	18

eNTERFACE'05 数据集[27] 是在实验室环境下录制的视听情感数据集, 包含来自 14 个国家的 44 名被试者 (其中 81% 为男性, 19% 为女性, 31% 戴眼镜, 17% 留有胡子)。实验过程中, 被试者需聆听 6 段用英文表述的短篇小说 (每段小说对应不同的情感), 并使用给定的 5 个句子来回答, 表达他们对情感的理解。两名专家对被试者的反应进行判断, 被试者们做出的反应如图 5(b) 所示。该数据集是实验室环境下自发生成的表情, 但数据集中第 6 号被试者出现录制表情表达不完整, 且第 23 号被试者缺少高兴视频, 同时未划分训练集和测试集。本文随机选择 32 个人的数据作为训练集, 其余 10 个人的数据作为测试集。表 2 给出了 eNTERFACE'05 数据集样本数。

表 2　eNTERFACE'05 数据集样本数
Table 2　Sample number of eNTERFACE'05 dataset

表情类别	生气	厌恶	害怕	开心	悲伤	惊讶
样本数	210	210	210	210	210	210

2　实验与结果分析

2.1　实验环境及网络模型微调

实验基于 PyTorch 深度学习框架的 GPU 版本完成, 版本号为 torch1.7.1, torchvision0.8.2。空间

子网络在 MSCeleb-1M 面部识别数据集和 FER Plus 表达数据集上进行预训练处理。

对模型的训练和测试基于 eNTERFACE'05 和 CK+数据集，优化器选择随机梯度下降 (stocgastic gradient descent, SGD) 优化器，动量值设置为 0.9，权重衰减为 10^{-4}，学习速率 (l_r) 初始化为 0.001，并且在 30 次迭代时将其修改为 0.000 2，并且在 100 次迭代之后停止训练。由于 eNTERFACE'05 和 CK+数据集均未划分出训练集和验证集，本文将 eNTERFACE'05 和 CK+数据集随机取出 2/3 用作训练集，余下用作验证集。

2.2　数据集预处理

由于 eNTERFACE'05 及 CK+是视频数据集，而神经网络只能输入张量，因此需从视频中提取人脸图片变成能直接输入神经网络的张量。CK+数据集不是动态视频，本文针对 eNTERFACE'05 数据集利用 ffmpeg 函数提取视频帧完成分帧处理。通过 Dlib 工具箱实现对两数据集的人脸检测和对齐，并以 25% 的比例扩展人脸边界框，然后将裁剪后的人脸大小调整为 224×224。并以一定的概率随机水平翻转人脸图像，提高模型的泛化能力。

2.3　实验结果分析

2.3.1　空间网络实验结果及可视化

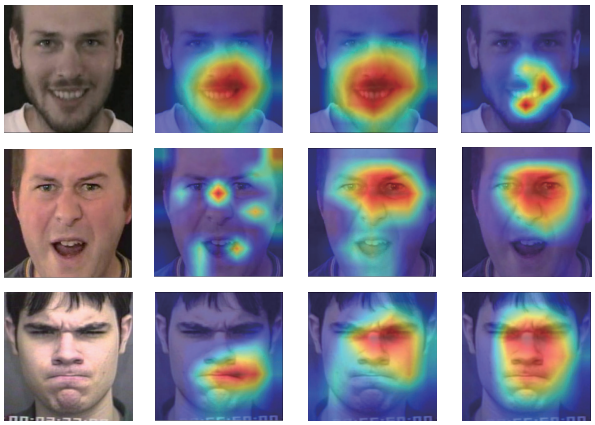
基于不同注意力机制的实验，参数量与识别准确度的实验结果对比见表 3。结果显示，以 FAN 为基线模型，在参数量方面，引入不同注意力机制后，参数量均略有增加，分别为 0.35、0.12 及 1.15 MB；识别准确度方面均有提升，而结合 SA 和 ECA 后的 eNTERFACE'05 和 CK+数据集准确度最高，分别达到 50.769% 和 99.982%。在 FAN 模型实际训练中，CK+准确度实际为 99.808%，与原论文相差约 0.88%。综合来看，不同注意力机制的引入显著提升了表情识别性能，采用多重注意力机制使网络集中关注有效特征信息，减弱无用信息的提取，有效提升识别准确度。

表 3　不同注意力机制的实验结果对比

Table 3　Comparison of experimental results for different attention mechanisms

不同注意力机制	参数量/MB	准确度/%	
		eNTERFACE'05	CK+
FAN	44.74	48.205	98.808
FAN+CBAM	(+0.35)	48.718	99.455
FAN+ECA	(+0.12)	49.744	99.148
FAN+CSA	(+1.15)	50.769	99.982

此外，本文还采用梯度加权类激活映射 (gradient-weighted class activation mapping, Grad-CAM)[28] 技术，对添加注意力机制后的空间网络特征图进行了可视化，图 6 给出了在高兴和生气表情的人脸图像中，添加注意力机制后网络关注的重要区域。在高兴的表情中，注意力机制的空间网络更倾向于关注嘴巴的中心区域，激活咧嘴、嘴角上扬等特征。而在生气的表情中，注意力机制的空间网络认为眼睛区域更为重要，激活张嘴、皱鼻等特征。图 6 热力图结果也表明，引入注意力机制后增强了网络针对不同表情的关键区域关注，网络不仅关注表情主要集中的如眼睛及嘴巴等关键区域，而且也有效关注面部关键点附近肌肉在不同表情下的变化，这为后续掩模的选取提供参考依据。



(a) FAN　(b) FAN+CBAM　(c) FAN+ECA　(d) FAN+SA+ECA

图 6　数据集部分样本的类激活热力图

Fig. 6　Class activation heatmaps for some samples of the dataset

2.3.2　时间网络不同掩模下的实验对比结果

表 4 给出基于关键区域不同掩模处理后的实验结果对比。与原始图像相比，3 种掩模选取策略在减少光流计算的同时能够去除冗余信息的干扰，显著提升了表情识别的准确度。其中，矩形区域 AU 单元掩模法将面部眼睛、鼻子、嘴巴等主要特征区域整合在一起，减少了背景和其他干扰区域的影响；分区单独提取 AU 单元掩模法可以更细致地捕捉每个特征区域的信息。尽管上述两种方法均涵盖了关键表情区域，却无法捕捉关键点附近细微的面部肌肉运动。与之不同，面部外围轮廓关键点掩模法将面部的关键表情区域信息及周围肌肉运动均涵盖其中，充分考虑了表情微小变化，相对于上述两种掩模法达到最佳的表情识别结果。

通过光流可视化箭头图（设置像素采样点间隔为 40），绘制光流采样点及其偏移情况。从图 7 中能够直观地观察到在人脸表情发生变化时，光流场中每个采样点的运动方向和大小。在表情变化显著区域，箭头的明显偏移突显了人脸表情的

动态变化。同时, 实验基于 3 种不同的掩模处理方式, 验证了其在光流计算中的有效性, 避免不相关位置的光流计算, 减少计算量并提高模型的运行效率。

表 4　不同掩模处理后的实验结果（准确度）对比
Table 4　Comparison of experimental results after different mask treatments　%

掩模处理方式	数据集	
	eNTERFACE'05	CK+
原图像	30.724	50.402
矩形区域AU单元掩模法	36.077	57.240
分区单独提取AU单元掩模法	36.324	58.416
面部外围轮廓关键点掩模法	37.376	59.474

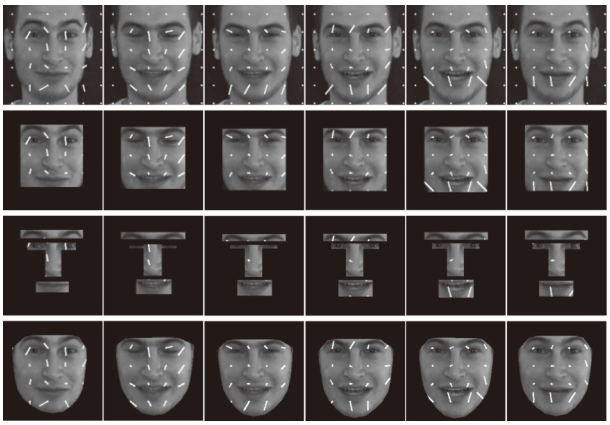


图 7　数据集部分光流可视化箭头图
Fig. 7　Arrow diagram of partial optical flow visualization of the dataset

2.3.3　双流网络融合前后实验对比结果

针对空间和时间双流网络的融合, 表 5 给出了不同融合策略下的表情识别性能。由图 8 可见, 最大概率的决策融合方法有效实现了空间和时间双流网络的互补融合。

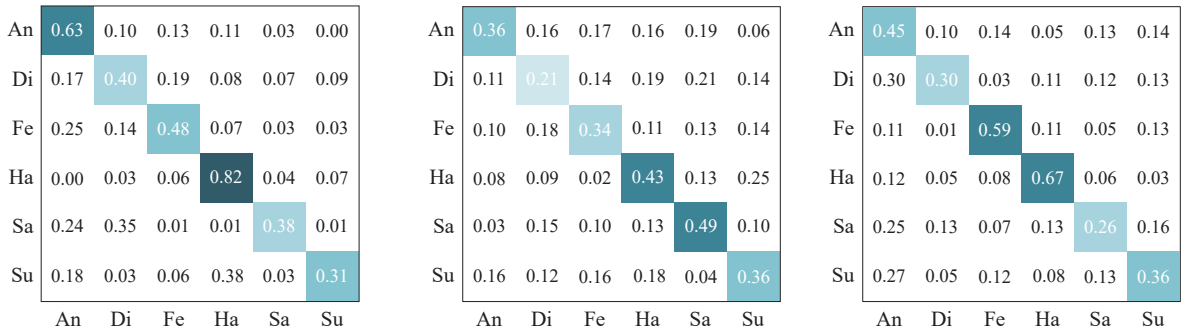
此外, 表 5 给出了单独空间网络和单独时间网络的人脸表情识别准确度。显然, 通过充分结合表情帧内的空间信息和帧间的时间信息, 融合时空域特征的人脸表情识别方法取得了表情识别精度的显著提升。同时, 也可以看出加权平均决策融合得到的识别准确度低于单个网络的识别准确度, 这是因为简单的加权平均融合并没有考虑两个网络的互补性。

表 5　不同融合策略下的表情识别结果（准确度）对比
Table 5　Comparison of expression recognition results under different fusion strategies　%

方法	数据集	
	eNTERFACE'05	CK+
空间网络	50.42	99.03
时间网络	36.45	58.61
决策融合(加权平均)	43.89	90.95
决策融合(最大概率)	70.69	99.87

图 8　数据集部分视频帧的融合效果
Fig. 8　Effect of the fusion of some video frames in the dataset

为了比较网络对每一类别的分类效果, 通过混淆矩阵对空间网络、时间网络、加权平均融合和最大概率融合网络进行评估, 如图 9 所示。混淆矩阵以矩阵形式按照真实标签类别与预测类别两个判断标准进行汇总: 预测类别分布在方阵中的每一列中, 真实类别分布在方阵中的每一行中, 每一列的数据表示预测为该类别的数量, 每一行的数据表示该类别真实标签类别数量。图中的混淆矩阵代表各类表情的识别概率, 对角线代表表情正确的识别概率。可以看出, 本文最大概率融合网络提高了各类表情的识别准确度。同时, 图 10 给出了空间网络、时间网络、加权平均融合和最大概率融合网络在 eNTERFACE'05 和 CK+数据集训练过程中损失函数的变化过程, 相对于其他 3 个网络, 最大概率融合网络收敛速度最快, 且收敛平稳。



(a) 空间网络eNTERFACE'05　(b) 时间网络eNTERFACE'05　(c) 加权平均融合eNTERFACE'05

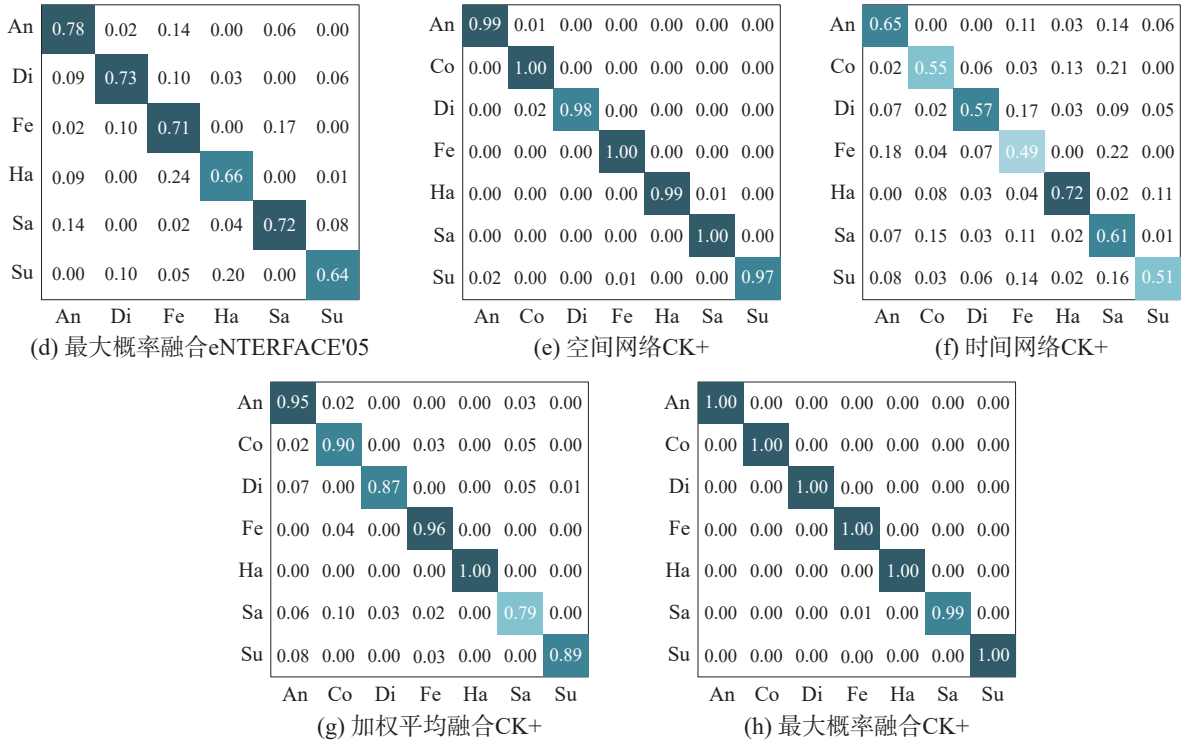


图 9　eNTERFACE'05 和 CK+数据集各网络的混淆矩阵

Fig. 9　Confusion matrices for each network for eNTERFACE'05 and CK+ datasets

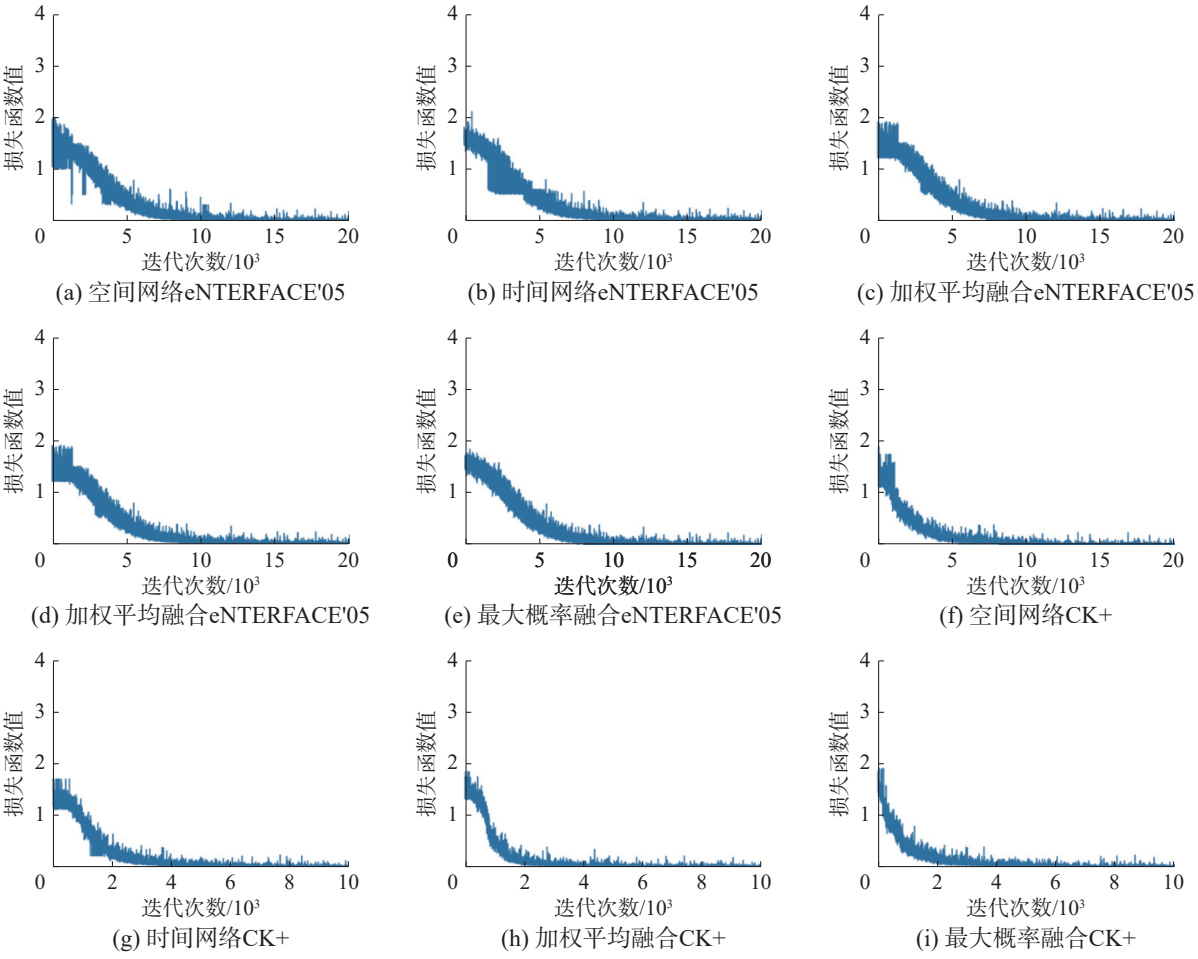


图 10　eNTERFACE'05 和 CK+数据集各网络训练中的损失函数

Fig. 10　Loss functions in the training of each network for eNTERFACE'05 and CK+ dataset

2.3.4 与其他方法的比较

将本文方法与采用 eNTERFACE'05 数据集的其他文献方法进行比较, 包括 5 种典型方法: ConvLSTM[14]、双路 LSTM[15]、3D-CNN 及 CRNN 融合模型[16]、HGR 最大相关性[29]、FAN[19]。目前大多数已提出的多模态情感识别方法多采用语音与人脸表情两种模态进行情绪识别, 如刘菁菁等[15] 采用双路 LSTM 分别模拟人类听觉和视觉处理通路处理语音和面部表情的情感信息, 并通过模拟人脑边缘系统情感区进行决策层加权特征融合, 实现表情识别。同时, 基于多模态情感识别方法相较于其他方法取得了更加理想的识别结果, 这是因为多模态情感识别具有更全面的信息、上下文理解、抗噪性提高和准确结果好的优点, 但它同时也存在复杂性增加、数据整合难度、多模态依赖性和模型复杂度等缺点。与之相比, 从表 6 中可以看出, 本文提出的视频表情识别方法相较于 LSTM 及 FAN 网络也取得了较好的识别结果。

表 6　eNTERFACE'05 数据集不同方法的识别准确度
Table 6　Recognition accuracy of different methods for eNTERFACE'05 dataset　%

方法	准确度
Farhoudi等(多模态)[16]	62.00
Ma等(多模态)[29]	62.16
张隽睿(LSTM)[14]	44.97
Meng等[19]	45.75
本文方法	70.69

同时, 将不同的双流表情识别方法在 CK+数据集上的识别结果与本文提出的融合关键区域信息的双流网络表情识别方法进行比较, 包括融合 PHRNN 时间网络和 MSCNN 空间网络[10]、基于 LBP-TOP 特征 (时间流) 和表情帧特征 (空间流) 并通过 CNN 输出平均作为最终结果[11] 等方法, 比较结果如表 7 所示。从表 7 中可以看出本文方法在 CK+数据集上识别率也有所提升。

表 7　CK+数据集不同方法的识别准确度
Table 7　Recognition accuracy of different methods for CK+ dataset　%

方法	准确度
Zhang等[10]	98.50
Feng等[11]	93.70
Chen等[12]	98.46
Zhao等[30]	98.77
Meng等[19]	99.69
本文方法	99.87

图 11 给出了表情识别正确与错误的一些样本。可以发现, 网络能够将“高兴”或“悲伤”等表情正确分类, 双流网络实现了较好的融合效果, 捕捉到了在“高兴”表情中嘴角上扬和眼睛放松的特征, 以及在“悲伤”表情中眉毛下垂和嘴角下弯的特征。

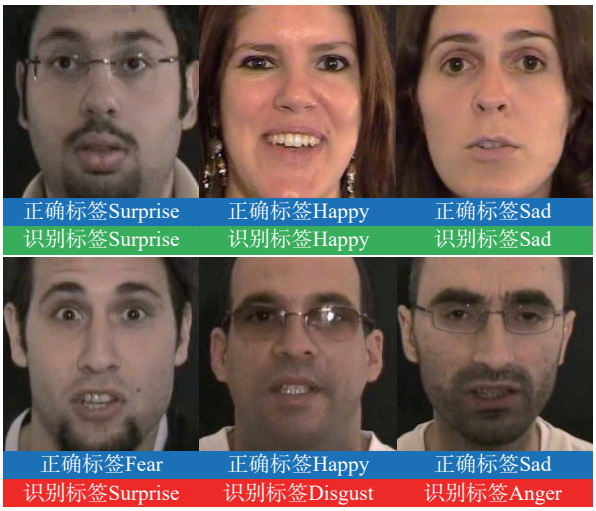


图 11　数据集部分正确分类和错误分类样本
Fig. 11　Plots of partially correctly classified and misclassified samples of the dataset

然而, 对于表情之间差异微小的表情容易产生混淆而造成错误分类, 从图 11 中也可以看到“害怕”被识别为“惊讶”, 分析原因可能是由于两者的表情在某些关键特征上存在相似之处, 如眼睛的张开程度或眉毛的位置; 类似地, 在“悲伤”和“生气”之间也出现误分类, 可能是因为在眉毛的位置、眼神的表现等方面存在相似之处。事实上, 消极类别的表情特征本身类似, 大多具有嘴角向下、眉头皱起等特征, 容易出现混淆, 导致网络识别出现错误。

3　结束语

针对视频表情识别问题, 本文提出了一种融合关键区域信息的双流网络表情识别模型。该模型分别构建空间和时间网络模型以提取视频帧内的空间信息与帧间时间信息; 空间网络利用 CSFA 有效关注表情识别的空间关键区域, 增强了识别的准确性; 提取光流获取时间信息的同时通过空间关键区域掩模控制减少光流计算量。实验采用了 CK+和 eNTERFACE'05 数据集进行有效性验证。实验结果表明, 所提出的方法在 eNTERFACE'05 和 CK+数据集上的准确度分别达到 70.69% 和 99.87%, 由此看出提出的方法可以较好提升表情识别准确度。未来的工作重点在表情遮挡及姿态变化时的表情识别, 并且也要考虑因网络结构复杂导致的参数量过大的问题, 研究轻量化网络的表情识别。

参考文献:

[1] 彭小江, 乔宇. 面部表情分析进展和挑战[J]. 中国图象图形学报, 2020, 25(11): 2337−2348.
PENG Xiaojiang, QIAO Yu. Advances and challenges in facial expression analysis[J]. Journal of image and graphics, 2020, 25(11): 2337−2348.

[2] SHAN Caifeng, GONG Shaogang, MCOWAN P W. Facial expression recognition based on local binary patterns: a comprehensive study[J]. Image and vision computing, 2009, 27(6): 803−816.

[3] ZHAO Guoying, PIETIKÄINEN M. Dynamic texture recognition using local binary patterns with an application to facial expressions[J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(6): 915−928.

[4] ZHI Ruicong, FLIERL M, RUAN Qiuqi, et al. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition[J]. IEEE transactions on systems, man, and cybernetics Part B, Cybernetics, 2011, 41(1): 38−52.

[5] ZHONG Lin, LIU Qingshan, YANG Peng, et al. Learning active facial patches for expression analysis[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012: 2562−2569.

[6] 应自炉, 张有为, 李景文. 融合人脸局部区域的表情识别[J]. 信号处理, 2009, 25(6): 963−966.
YING Zilu, ZHANG Youwei, LI Jingwen. Facial expression recognition by fusing local facial regions[J]. Journal of signal processing, 2009, 25(6): 963−966.

[7] 何俊, 蔡建峰, 房灵芝. 基于 LBP 特征的融合脸部关键表情区域的表情识别方法[C]//第 27 届中国控制与决策会议. 青岛: 信息科技, 2015: 1209−1213.
HE Jun, CAI Jianfeng, FANG Lingzhi. Facial expression recognition method based on LBP feature fusion of key facial expression regions [C]// 27th China Control and Decision Conference. Qingdao: Information Technology, 2015: 1209−1213.

[8] JAIN S, HU Changbo, AGGARWAL J K. Facial expression recognition with temporal modeling of shapes[C]// 2011 IEEE International Conference on Computer Vision Workshops. Barcelona: IEEE, 2011: 1642−1649.

[9] SIKKA K, SHARMA G, BARTLETT M. LOMo: latent ordinal model for facial analysis in videos[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 5580−5589.

[10] ZHANG Kaihao, HUANG Yongzhen, DU Yong, et al. Facial expression recognition based on deep evolutional spatial-temporal networks[J]. IEEE transactions on image processing, 2017, 26(9): 4193−4203.

[11] FENG Duo, REN Fuji. Dynamic facial expression recognition based on two-stream-CNN with LBP-TOP[C]// 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems. Nanjing: IEEE, 2018: 355−359.

[12] CHEN Tuo, XING Shuai, YANG Wenwu, et al. Spatio-temporal features based human facial expression recognition[J]. Journal of image and graphics, 2022, 27(7): 2185−2198.

[13] HASANI B, MAHOOR M H. Facial expression recognition using enhanced deep 3D convolutional neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017: 2278−2288.

[14] 张隽睿. 基于深度学习的静态和动态面部表情识别研究[D]. 成都: 电子科技大学, 2022.
ZHANG Junrui. Research on static and dynamic facial expression recognition based on deep learning[D]. Chengdu: University of Electronic Science and Technology of China, 2022.

[15] 刘菁菁, 吴晓峰. 基于长短时记忆网络的多模态情感识别和空间标注[J]. 复旦学报 (自然科学版), 2020, 59(5): 565−574.
LIU Jingjing, WU Xiaofeng. Real-time multimodal emotion recognition and emotion space labeling using LSTM networks[J]. Journal of Fudan University (natural science), 2020, 59(5): 565−574.

[16] FARHOUDI Z, SETAYESHI S. Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition[J]. Speech communication, 2021, 127: 92−103.

[17] FERNANDEZ P D M, PENA F A G, REN T I, et al. FERAtt: facial expression recognition with attention net[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2019: 837−846.

[18] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[M]// Medical Image Computing and Computer-Assisted Intervention. Cham: Springer International Publishing, 2015: 234−241.

[19] MENG Debin, PENG Xiaojiang, WANG Kai, et al. Frame attention networks for facial expression recognition in videos[C]//2019 IEEE International Conference on Image Processing. Taipei: IEEE, 2019: 3866−3870.

[20] 李同霞. 基于表征流嵌入网络的动态表情识别[D]. 南京: 南京邮电大学, 2022.
LI Tongxia. Dynamic expression recognition based on representation stream embedding network[D]. Nanjing:

Nanjing University of Posts and Telecommunications, 2022.

[21] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[22] WANG Qilong, WU Banggu, ZHU Pengfei, et al. ECA-net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11531−11539.

[23] WANG Dong, GAO Feng, DONG Junyu, et al. Change detection in synthetic aperture radar images based on convolutional block attention module[C]//2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp). Shanghai: IEEE, 2019: 1−4.

[24] SIMONYAN K, ZISSERMAN A, SIMONYAN K, et al. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1. [S.l.]: ACM, 2014: 568−576.

[25] FARNEBÄCK G. Two-frame motion estimation based on polynomial expansion[C]//Image Analysis. Berlin: Springer Berlin Heidelberg, 2003: 363−370.

[26] LUCEY P, COHN J F, KANADE T, et al. The extended Cohn-kanade dataset (CK): a complete dataset for action unit and emotion-specified expression[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. San Francisco: IEEE, 2010: 94−101.

[27] OLIVIER M, IRENE K, BENOIT M, et al. The eNTERFACE'05 audio-visual emotion database[C]//Proceedings of the 22nd International Conference on Data Engineering Workshops. [S.l.]: IEEE, 2006: 8−15.

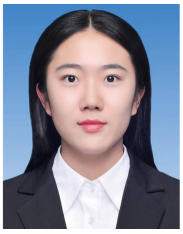
[28] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks *via* gradient-based localization[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 618−626.

[29] MA Fei, HUANG Shaolun, ZHANG Lin. An efficient approach for audio-visual emotion recognition with missing labels and missing modalities[C]//2021 IEEE International Conference on Multimedia and Expo. Shenzhen: IEEE, 2021: 1−6.

[30] ZHAO Jianfeng, MAO Xia, ZHANG Jian. Learning deep facial expression features from image and optical flow sequences using 3D CNN[J]. The visual computer, 2018, 34(10): 1461−1475.

作者简介:

孔英会, 教授, 主要研究方向为机器学习与计算机视觉。发表学术论文 50 余篇。E-mail: kongyh2005@163.com。



崔文婷, 硕士研究生, 主要研究方向为计算机视觉。E-mail: 1595470319@qq.com。

张珂, 教授, 主要研究方向为计算机视觉和电力人工智能。发表学术论文 100 余篇。E-mail: zhangkeit@ncepu.edu.cn。