DOI:10.16136/j.joel.2022.06.0725

融合局部语义与全局信息的人脸表情识别

潘海鹏,郝 慧,苏 雯*

(浙江理工大学 机械与自动控制学院,浙江 杭州 310018)

摘要:人脸表情识别在人机交互等人工智能领域发挥着重要作用,当前研究忽略了人脸的语义信息。本文提出了一种融合局部语义与全局信息的人脸表情识别网络,由两个分支组成:局部语义区域提取分支和局部-全局特征融合分支。首先利用人脸解析数据集训练语义分割网络得到人脸语义解析,通过迁移训练的方法得到人脸表情数据集的语义解析。在语义解析中获取对表情识别有意义的区域及其语义特征,并将局部语义特征与全局特征融合,构造语义局部特征。最后,融合语义局部特征与全局特征构成人脸表情的全局语义复合特征,并通过分类器分为7种基础表情之一。本文同时提出了解冻部分层训练策略,该训练策略使语义特征更适用于表情识别,减少语义信息冗余性。在两个公开数据集JAFFE和KDEF上的平均识别准确率分别达到了93.81%和88.78%,表现优于目前的深度学习方法和传统方法。实验结果证明了本文提出的融合局部语义和全局信息的网络能够很好地描述表情信息。

关键词:人脸表情识别;人脸解析;迁移学习;局部-全局特征融合;解冻部分层训练策略

中图分类号:TP183 **文献标识码**:A **文章编号**:1005-0086(2022)06-0652-08

Facial expression recognition based on fusion of local semantic and global information

PAN Haipeng, HAO Hui, SU Wen*

(School of Mechanical Engineering & Automation, Zhejiang Sci-Tech University, Hangzhou, Zhejiang 310018, China)

Abstract: Facial expression recognition plays an important role in artificial intelligence such as human-computer interaction. However, current researchers ignore the semantic information of human faces. In this paper, we propose a facial expression recognition network fusing local semantic and global information, which consists of two branches: the local semantic region extraction branch and the local-global feature fusion branch. Firstly, the face semantic parsing is achieved by training semantic segmentation network on face parsing dataset. The semantic parsing of facial expression dataset is obtained by transfer training. Then the meaningful regions and their semantic features are extracted and fused with the global features to obtain the semantic local features. Finally, the global semantic composite features of facial expressions are constructed by combining semantic local features with global features. They are classified into one of the 7 basic facial expressions by the classifier. We also propose a training strategy of unfreezing partial layers, which makes semantic features more suitable for facial expression recognition and reduces the redundancy of semantic information. The average recognition accuracy on two public datasets, JAFFE and KDEF, reaches 93.81% and 88.78%, respectively. The performance outperforms the current deep learning methods and traditional methods. The experimental results demonstrate that the network proposed can describe the expression information comprehensively by integrating local semantic and global information.

* **E-mail**: wensu@zstu.edu.cn

收稿日期:2021-10-22 **修订日期**:2021-11-25

基金项目:国家自然科学基金(62006209)、浙江省自然科学基金(LQ20F020001)、浙江理工大学科研启动基金(1802225-Y)和浙江理工大学基本科研业务费专项资金(2020Q014)资助项目

Key words: facial expression recognition; face parsing; transfer learning; local-global feature fusion; training strategy of unfreeze partial layers

1 引　言

在日常生活中,人脸表情扮演着十分重要的角色,它可以传递人们的心情与想法。随着计算机计算能力的大幅提升和人工智能的应用落地,人脸表情识别已经成为流行的研究方向,广泛应用在人机交互、司法侦察、监测监控等领域。人脸表情一般分为 7 种基础表情:生气、厌恶、恐惧、高兴、中性、悲伤和惊讶。

深度学习能从海量数据中自主学习所需的特征。国内外研究学者设计出了很多优秀的模型用于表情识别,并取得了较好的实验效果。现有的人脸表情识别网络主要从局部-全局、身份-表情和损失层 3 个方向展开。LIU 等[1]针对较大姿态变化导致识别率不高的问题,提出了多通道姿态感知神经网络模型(multi-channel pose-aware convolution neural networks,MPCNN)。不足之处是只在每个子卷积神经网络(convolutional neural network,CNN)的全连接层进行特征融合,导致特征过于抽象,对细节信息的感知能力较差。YOLCU 等[2]提出了 4 支路分割分类级联网络,将人脸的眉毛、眼睛和嘴巴分割出来。但本文方法在训练分割网络之前,需要生成训练掩膜。CAI 等[3]针对 Center 损失[4]只减小类内距离的问题,提出了经典的 Island 损失。

上述方法在识别人脸表情的任务中效果显著,且解决了角度转换、身份不同等带来的影响,但忽略了人脸的语义信息。本文将语义分割网络引入到人脸表情识别领域,以此提高识别准确率。全局信息描述了人脸表情的颜色、形状等整体特征,且表示直观。但其特征维数高,会忽略掉一些细节信息,因此我们将局部语义特征融入到全局特征中,使网络能更好地描述表情信息。另外,为了减少语义信息的冗余性,本文提出了解冻部分层训练策略,使提取到的语义特征更适用于表情识别任务。本文的主要贡献是:

1) 本文尝试引入语义信息到人脸表情识别领域。首先在人脸解析数据集上,训练语义分割网络解析人脸,然后利用迁移的策略得到人脸表情数据集的语义解析。最后提取其中对人脸表情有意义的区域及其语义特征。2) 为了充分利用人脸表情的空间信息,提出了局部-全局特征融合的计算结构。利用语义区域提取分支的掩膜全局信息得到局部语义特征,将其与全局特征融合得到语义局部特征。此方法使得全局特征得到语义特征的补充,屏蔽了无意义语义区域对人脸识别引入的冗余影响,使复合特征更注重于描述重点语义区域。3) 本文提出了解冻部分层训练策略。在语义提取分支上,只训练部分层:编码器的第 3、第 4、第 5 层最大池化层之前的卷积层,解码器的第 1、第 2 层反卷积层及其对应的批标准化(Batch Normalization,BN)层。该训练策略使语义分支的部分层参数不断调整,提高其对表情识别任务的适应性。4) 在人脸表情识别的公开数据集 JAFFE (Japanese female facial expression dataset)[5]和 KDEF(Karolinska directed emotional faces dataset)[6]上对本文提出的方法进行了验证,得到的人脸表情平均识别率分别为 93.81% 和 88.78%。

2 提出的方法

本文提出了一种融合局部语义和全局信息的人脸表情识别网络,用来分类 7 种基础表情。网络结构如图 1 所示,由局部语义区域提取分支(上分支)和局部-全局特征融合分支(下分支)构成。上分支的目的是解析人脸,获取人脸的语义区域,并提取其中对识别表情有意义的局部区域。下分支利用语义区域掩膜全局特征得到有意义的局部语义特征,并与全局特征融合,使网络能更好地描述表情信息。

2.1 局部语义区域提取分支

当前很多学者对人脸进行解析,以此获取像素级的语义特征。但在人脸表情识别领域中,很少有工作考虑到像素级信息对表情识别的重要性。究其原因,主要是因为人脸表情数据集没有精细的像素级标签,而人工标注的成本很高。考虑到人脸解析数据集和人脸表情数据集的相似性,本文利用迁移学习方法,将人脸解析数据集的解析模型迁移至人脸表情数据集,从而辅助无语义分割标签的人脸表情数据集进行人脸解析。局部语义区域提取分支选取全卷积网络-8S(fully convolutional networks,FCN-8S)[7],人脸像素标签作为监督,以此获取其语义信息。FCN-8S 网络由编码器和解码器构成,编码器为去除全连接接层和分类层的预训练视觉几何组网络-16(visual geometry group network,VGG-16)[8],作用是得到输入图像的热图。解码器为结合不同深度层的跳级结构,利用上采样的方式使输出特征图的尺寸和原始输入图像相同。

输入图像 $\boldsymbol{I}\in\boldsymbol{R}^{C\times H\times W}$ 到语义分割网络 S 逐像素

预测概率，将所有像素分为 11 类，得到语义特征图 $\boldsymbol{Mask}_{11}\in\boldsymbol{R}^{11\times H\times W}$ 及掩膜 $\boldsymbol{Mask}\in\boldsymbol{R}^{H\times W}$。其中 C 为图像通道数，H 和 W 分别为输入图像及特征图的高度和宽度。语义分割计算过程可用式(1)表示：

$$\boldsymbol{Mask}=S_{\theta_S}(\boldsymbol{I}),\tag{1}$$

式中，θ_S 为局部语义区域提取分支参数，$\boldsymbol{Mask}\in\boldsymbol{R}^{H\times W}$ 是大小为 $H\times W$ 矩阵。其中的元素 k_{mn} 表示输入图像的第 m 行第 n 列像素所属的类别。

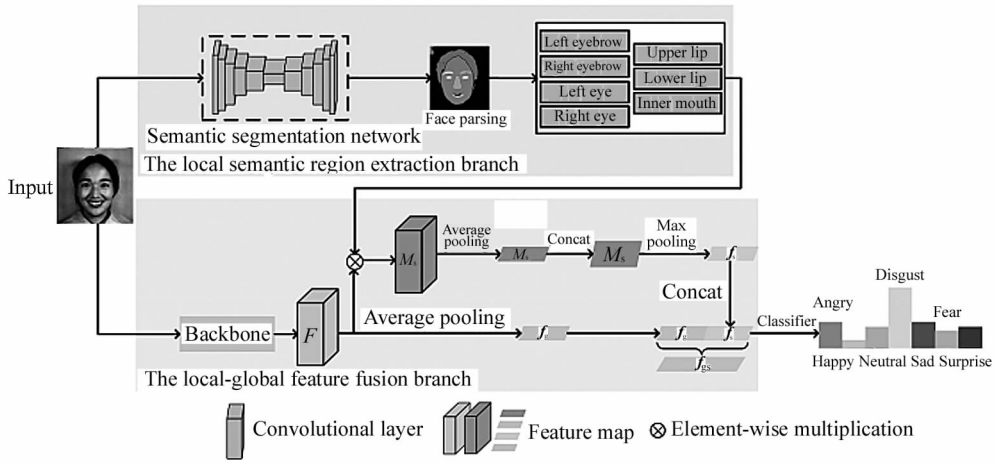

图 1　网络结构图

Fig. 1　Structure of network

语义分割网络解析的 11 类为背景、脸部皮肤、左眉毛、右眉毛、左眼睛、右眼睛、鼻子、上嘴唇、下嘴唇、嘴巴内部和头发。参考 YOLCU 等[2]、JUNG 等[9]论文的一般结论，眉毛、眼睛和嘴巴对表情识别至关重要。本文选取了其中对表情识别有意义的 7 类区域及其语义特征：左眉毛、右眉毛、左眼睛、右眼睛、上嘴唇、下嘴唇和嘴巴内部，作为后续特征融合的一部分。值得注意的是，这些语义区域的解析依赖于人脸解析数据，对人脸表情识别任务并无特别的适应性。在引入重点语义区域的同时，也可能引入某些冗余的语义信息。

2.2　局部-全局特征融合分支

全局特征可以获取人脸的外观和整体轮廓等信息，但往往会冗余，而且当卷积神经网络较深时，特征的分辨率往往很低，容易丢失细节信息。局部语义信息在像素级上描述了人脸表情部分区域的特征，反映了身份、性别等信息，且鲁棒性更强。因此本文将局部语义特征与全局特征相融合，使网络学习不同分辨率的特征，增强泛化能力。

本文选择了 HUANG 等[10]提出的密接卷积网络 121(densely connected convolutional networks，DenseNet121)来提取全局特征。此网络通过建立前面所有层与后面层的密集连接，实现了特征在通道维度上的复用，更有利于提取表情特征。输入表情图像 $\boldsymbol{I}\in\boldsymbol{R}^{C\times H\times W}$ 到去除分类层的预训练 DenseNet121 网络 D，得到具有人脸整体轮廓等信息的全局特征 $\boldsymbol{F}\in\boldsymbol{R}^{C_1\times H_1\times W_1}$。其中 C_1、H_1 和 W_1 分别为特征图的通道数、高度及宽度。然后，输入到平均池化层，保留主要特征的同时降低计算量，并经过降维得到特征 $\boldsymbol{f}_g\in\boldsymbol{R}^{C_1\times 1}$。采用双线性插值方法下采样语义特征图 $\boldsymbol{Mask}_{11}\in\boldsymbol{R}^{11\times H\times W}$，使其高度及宽度和全局特征 $\boldsymbol{F}$ 相同。提取其中对表情识别有意义的 7 类语义特征 $\boldsymbol{Mask}'\in\boldsymbol{R}^{7\times H_1\times W_1}$，将每类特征 $\boldsymbol{Mask}'_j\in\boldsymbol{R}^{1\times H_1\times W_1}$ 分别与全局特征 $\boldsymbol{F}$ 逐元素相乘，得到重点区域语义特征 $\boldsymbol{M}_i$。该操作赋予全局特征 $\boldsymbol{F}$ 中的每个元素一个权重，强调人脸重要局部区域对表情识别的影响，并抑制一些无关的细节信息。$\boldsymbol{M}_i$ 经过平均池化并降维得到融合特征 $\boldsymbol{M}_{ia}\in\boldsymbol{R}^{C_1\times 1}$。将 7 类融合特征 $\boldsymbol{M}_{ia}$ 拼接，以此获取全部选定区域的语义特征 $\boldsymbol{M}_s\in\boldsymbol{R}^{C_1\times 7}$，即语义局部特征。上述计算过程可用式(2)—(6)表示：

$$\boldsymbol{F}=\boldsymbol{D}_{\theta_D}(I),\tag{2}$$

$$\boldsymbol{f}_g=Avgpool(\boldsymbol{F}),\tag{3}$$

$$\boldsymbol{M}_i=\boldsymbol{F}\otimes\boldsymbol{Mask}'_i,\tag{4}$$

$$\boldsymbol{M}_{ia}=Avgpool(\boldsymbol{M}_i),\tag{5}$$

$$\boldsymbol{M}_s=Concat(\boldsymbol{M}_{2a},\boldsymbol{M}_{3a},\boldsymbol{M}_{4a},\boldsymbol{M}_{5a},\boldsymbol{M}_{7a},\boldsymbol{M}_{8a},\boldsymbol{M}_{9a}),\tag{6}$$

式中，θ_D 为去除分类层的预训练 DenseNet121 网络的参数，$Avgpool$ 表示平均池化操作。$\otimes$表示逐元素相乘，$\boldsymbol{Mask}'_i\in\boldsymbol{R}^{1\times H_1\times W_1}$ 表示第 i 类语义特征。

$\boldsymbol{Mask'}_i$ 与全局特征 $\boldsymbol{F}$ 融合之后得到对应类别的语义特征图 $\boldsymbol{M}_i \in \boldsymbol{R}^{C_1 \times H_1 \times W_1}$,$Concat$ 为拼接操作。

语义局部特征 $\boldsymbol{M}_s$ 经过最大池化层生成特征 $\boldsymbol{f}_s \in \boldsymbol{R}^{C_1 \times 1}$,保留更多的纹理信息。$\boldsymbol{f}_g$ 与 $\boldsymbol{f}_s$ 按通道拼接得到全局语义复合特征 $\boldsymbol{f}_{gs} \in \boldsymbol{R}^{2C_1 \times 1}$,实现局部特征对全局特征的补充。最后经过分类器进行分类,分类器由全连接层、BN 层、激活函数和 Softmax 层组成。上述特征融合的计算过程可用式(7)—(8)表示:

$$\boldsymbol{f}_s = Maxpool(\boldsymbol{M}_s), \tag{7}$$

$$\boldsymbol{f}_{gs} = Concat(\boldsymbol{f}_g, \boldsymbol{f}_s), \tag{8}$$

式中,$Maxpool$ 表示最大池化,$Concat$ 为拼接操作。

2.3　解冻部分层训练策略

当冻结上分支全部参数时,语义区域的解析完全依赖于人脸解析,导致其不能很好地适用于表情识别任务。若将所有层都解冻,即根据网络损失更新每层参数,会发现整个网络效果极差(表现为网络不收敛)。究其原因,在于该损失为单纯的分类损失,不包括像素分类的损失。虽然人脸解析和人脸表情识别任务具有相似性,都需要处理人脸的五官和背景等,但关注点不同。人脸解析更注重每个人人脸五官成分的分割,而人脸表情识别更注重人与人之间五官的相似与不同。因此本文只解冻上分支的部分层:编码器的第 3、第 4、第 5 层最大池化层之前的卷积层;解码器的第 1、第 2 层反卷积层及其对应的 BN 层,该训练策略使网络效果变得更好。一方面是底层信息具有共性,表示整体轮廓等信息,因此不需要重新训练;另一方面是高级信息会关注不同的任务,能高度概括不同任务的属性。

2.4　训练策略与损失函数

语义特征提取阶段:此阶段在人脸解析数据集上进行全监督人脸解析。输入图像到语义分割网络得到预测结果。将解析结果与像素级标签之间的损失记为分割损失 $\boldsymbol{L}_{fp}$,采用逐像素交叉熵损失来计算。分割损失反向传播到语义分割网络,使网络参数不断更新,逐渐准确地分类像素。$\boldsymbol{L}_{fp}$ 可用式(9)计算:

$$\boldsymbol{L}_{fp} = -\sum_{i=1}^{c} \boldsymbol{z}_i^{pixel} \log(\boldsymbol{y}_i^{pixel}), \tag{9}$$

式中,c 为像素种类数量,$\boldsymbol{z}_i^{pixel}$ 为指示变量,如果该类别和输入图像像素的所属类别相同则取 1,否则为 0,$\boldsymbol{y}_i^{pixel}$ 表示输入图像的像素属于该类别的预测值。

局部-全局特征融合阶段:人脸表情数据集没有像素级标签,无法得到其分割损失。因此,此阶段将上分支的参数全部冻结,分类损失只在下分支反向传播(其中骨干网络的参数被冻结)。将人脸表情的语义特征与全局信息融合,并采用经典的交叉熵损失计算分类损失 $\boldsymbol{L}_{gs}$,如式(10)所示:

$$\boldsymbol{L}_{gs} = -\sum_{i=1}^{k} \boldsymbol{z}_i \log(\boldsymbol{y}_i), \tag{10}$$

式中:k 为表情种类数量,$\boldsymbol{z}_i$ 为指示变量,如果该类别和输入图像所属类别相同则取 1,否则为 0,$\boldsymbol{y}_i$ 表示输入图像属于该类别的预测值。

局部-全局特征融合阶段(解冻部分层训练策略):此阶段分类损失 $\boldsymbol{L}_{gs}$ 反向传播至上分支的解冻层和下分支,不断更新上下分支的参数(其中骨干网络的参数被冻结)。

3　实验与分析

3.1　数据集

为了验证提出方法的有效性,本文在公开的人脸表情数据集 JAFFE 和 KDEF 上进行了实验,在公开的人脸解析数据集 Helen 数据集[11]上进行了人脸解析的训练。JAFFE 数据集包含 10 位日本女性正脸的 7 类表情图像,一共有 213 张,均为像素 256×256 的 8 位灰度图,实验采用三折交叉验证。KDEF 数据集包含 20—30 岁的 70 位业余演员(35 位男性和 35 位女性)的 7 类表情图像,一共有 4 900 张,均为像素 562×762 的彩色图。每种表情有 5 种角度,分别为±90°、±45°以及正面角度。实验选用正面角度图像,一共 980 张,采用五折交叉验证。Helen 数据集包含 2 330 张人脸图像,每张图片在像素级上被标记为 11 类:背景、脸部皮肤、左眉毛、右眉毛、左眼睛、右眼睛、鼻子、上嘴唇、下嘴唇、嘴巴内部和头发。在本文的实验中训练集有 2 000 张,验证集有 230 张,测试集有 100 张。

3.2　实现细节

本文实验基于深度学习框架 Pytorch 实现,采用 GTX1650 进行 GPU 加速,CPU 为 i5-9300HF@2.40 GHz。在训练过程中,将输入图像和人脸解析数据集标签的大小调整为 224×224,并进行标准化,其参数为[0.485,0.456,0.406],[0.229,0.224,0.225]。网络的批大小设为 8。对于 JAFFE 数据集,采用随机水平翻转作为其数据增强。对于 KDEF 数据集,在对输入图像调整大小之前,将其中心裁剪为大小 450×450 的图片。局部-全局特征融合分支的网络学习率设置为 0.001,优化器算法采用 Adam,权重设置为 5×10^{-4}。余弦退火衰减作为学习率衰减策略,使学习率按照周期变化,设置余弦函数周期

为111，最小学习率为0。人脸解析分支的学习率为0.01，并采用随机梯度下降算法（stochastic gradient descent，SGD）优化，动量值设为0.7。本文使用Kaiming方法初始化分类器的全连接层和BN层，防止激活输出爆炸或消失。

3.3 局部语义区域提取实验

本文采用FCN-8S网络提取人脸语义特征。在Helen数据集上进行了全监督训练，将人脸分为11类，生成的掩膜大小为224×224。当训练到第125轮时，网络损失不再降低，平均像素分类准确率和平均交并比都达到了最高，分别为93.76%和66.68%。将此网络迁移至人脸表情数据集进行语义解析，并提取对识别表情有意义的7个区域。如图2所示，分别为JAFFE数据集和KDEF数据集的部分解析可视化结果。由于人脸表情数据集没有语义分割标签，因此得不到平均像素分类准确率和平均交并比等定量指标，但从可视化结果能看出此分支对7个语义区域的划分比较准确。

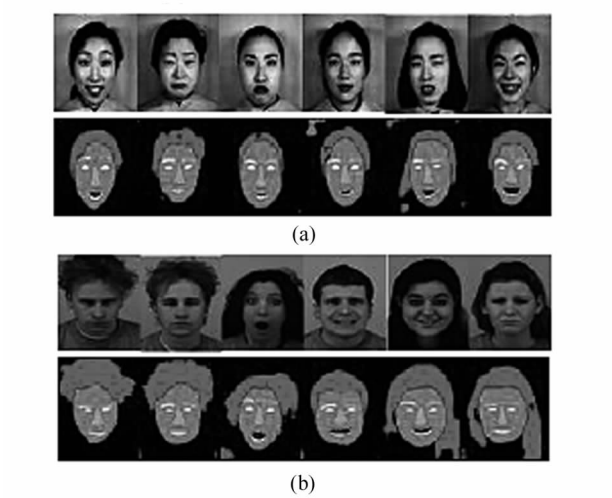

图2 JAFFE和KDEF数据集部分人脸解析结果：(a) JAFFE；(b) KDEF

Fig. 2 Partial face parsing results of JAFFE and KDEF datasets: (a) JAFFE; (b) KDEF

3.4 消融实验

3.4.1 全局特征提取实验

实验选用预训练的DenseNet121网络作为主干。输入图像到此网络，得到大小7×7的全局特征$\boldsymbol{F}$，通道数为1024。然后接平均池化层进行下采样并降维，输出尺寸为1024×1的向量$\boldsymbol{f}_{\mathrm{g}}$。此操作在一定程度上防止网络过拟合。最后经过分类器进行分类，将特征由1024降为256，再由256降到7。在KDEF数据集上，在最后一层全连接层之前加上Dropout层，随机使50%的神经元失活。全局特征实验也即本文的Baseline实验，在JAFFE数据集和KDEF数据集上的平均准确率分别为86.19%和81.12%。

3.4.2 局部-全局特征融合实验

此实验将局部语义区域提取分支的参数全部冻结，即不更新网络参数，网络损失只传播到下分支。具体操作如下：对3.3实验的11类语义特征下采样，使其尺寸变为7×7，提取其中的7类特征分别与全局特征$\boldsymbol{F}$逐元素相乘，然后自适应平均池化层调整其大小为1×1，经过降维、升维及拼接操作，得到语义局部特征$\boldsymbol{M}_{\mathrm{s}}$，其大小为1024×7（分别对应人脸解析结果的7类语义特征：左眉毛、右眉毛、左眼睛、右眼睛、上嘴唇、下嘴唇和嘴巴内部）。$\boldsymbol{M}_{\mathrm{s}}$经过自适应最大池化层得到尺寸为1024×7的向量$\boldsymbol{f}_{\mathrm{s}}$。$\boldsymbol{f}_{\mathrm{g}}$与$\boldsymbol{f}_{\mathrm{s}}$按通道拼接，生成全局语义复合特征$\boldsymbol{f}_{\mathrm{gs}}$。最后经过分类器进行分类。当训练到第60轮时，分类损失不再下降，平均准确率不再上升，此时网络性能达到了最优。如表1所示，在JAFFE数据集和KDEF数据集上的平均准确率分别为88.09%和83.67%，相比Baseline实验分别高了1.90%和2.55%。由实验结果可以看出，两个数据集上的准确率都比Baseline高，说明加入人脸的语义信息是有效的。语义区域提取分支学习人脸的语义特征，在像素级上学习相同表情的相似和不同表情的差异，更有利于网络识别表情。

3.4.3 局部-全局特征融合实验（解冻部分层训练策略）

在3.4.2实验的基础上，加入解冻部分层训练策略，即在上分支中，只训练部分层参数。当训练到第60轮时，分类损失不再降低，平均准确率不再上升，此时得到了最优的网络参数。如表1所示，经过训练之后JAFFE数据集和KDEF数据集的平均准确率分别为93.81%和88.78%，比3.4.2实验分别高了5.72%和5.11%。由实验结果可以看出，两个数据集上的准确率均比冻结上分支全部参数时高。原因是表情分类损失进行反向传播时，不断更新上分支的部分层参数，增强了语义特征对表情识别的适应性。如图3所示，分别为JAFFE数据集和KDEF数据集的混淆矩阵。

表 1　两个数据集的消融实验准确率

Tab. 1　Accuracy of ablation experiments on two datasets

Dataset	Baseline/%	Baseline+FCN/%	Baseline+FCN (the unfreeze partial layers training strategy)/%
JAFFE	86.19	88.09	93.81
KDEF	81.12	83.67	88.78

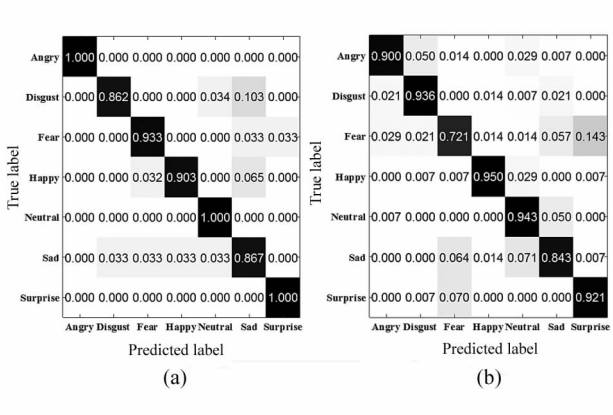

图 3　局部-全局特征融合实验(解冻部分层训练策略)JAFFE 和 KDEF 数据集的混淆矩阵:(a) JAFFE;(b) KDEF

Fig. 3　The confusion matrices for JAFFE and KDEF datasets on local-global feature fusion experiment (training strategy of unfreeze partial layers): (a) JAFFE; (b) KDEF

3.5　对比实验

本文的实验结果与现有的几种实验结果进行比较,结果见表 2 和表 3。

表 2　JAFFE 数据集的对比实验准确率

Tab. 2　Accuracy of comparative experiments on JAFFE dataset

Method	Accuracy/%
SFL[12]	78.64
KECANet+DAGSVM[13]	88.35
LBP+ORB Features[14]	88.50
Multi-modal[15]	91.80
Our	93.81

表 3　KDEF 数据集的对比实验准确率

Tab. 3　Accuracy of comparative experiments on KDEF dataset

Method	Accuracy/%
EDR-PCANet[16]	80.61
LDBP +SVM[17]	83.51
Deep-Emotion[18]	86.33
TLCNN+FOS[a][19]	88.27
Our	88.78

在 JAFFE 数据集上,将本文实验结果与文献[12]—[15]进行比较。SUN 等[12]提出了基于先验知识图的自适应特征学习方法(self-adaptive feature learning,SFL),去学习和提取感兴趣的特征,该方法不仅考虑了全局信息,也计算了局部特征。本文不仅将局部与全局特征融合,还计算了人脸表情的语义特征,本文的准确率比其高了 15.17%。CHEN 等[13]提出了一种采用核熵成分分析网络(kernel entropy component analysis network,KECANet)、二进制哈希和分块直方图提取图像特征,有向无环图支持向量机(directed acyclic graph support vector machine,DAGSVM)进行分类的模型。但没有将局部特征与全局信息很好地联系起来,本文的准确率比它高了 5.46%。NIU 等[14]提取了图像的 LBP 特征和 ORB 特征(oriented FAST and rotated BRIEF features),并进行融合。但该方法需要人工提取大量特征,且这些特征比较表层,本文的准确率比它高了 5.31%。WEI 等[15]提出了融合低级经验特征和高级自学习特征的模型。但在特征融合时采取了线性融合,此操作会丢失一些原始信息,本文的准确率比它高了 2.01%。

在 KDEF 数据集上,将本文实验结果与文献[16]—[19]进行比较。SUN 等[16]利用 EDR-PCA-Net(extended dictionary representation - principal component analysis network)模型来识别人脸表情。相比上述模型,本文引入了对识别表情有意义区域的语义特征,识别率比其高了 8.17%。

SANTRA 等[17]提取图像的局部主导二进制模式(local dominant binary patterns,LDBP)特征,并使用支持向量机(support vector machine,SVM)来分类多视角人脸表情。本文使用了较深的 DenseNet121 和 FCN-8S 网络提取特征,模型能学到更深层的抽象特征,且鲁棒性更强。本文方法的识别率比它高了 5.27%。MINAEE 等[18]为了更关注人脸表情的关键区域,提出了注意力卷积神经网络。本文引入了人脸表情的局部语义特征,比其准确率高了 2.45%。ZHOU 等[19]将人脸图像输入到迁移

学习卷积神经网络(transfer learning convolutional neural network, TLCNN),并利用FOS(face-occupancy-based feature selection)方法来选择特征图,以此提高分类的鲁棒性。本文将局部特征和全局信息融合计算,使网络对任务的了解更全面,识别准确率比ZHOU等[19]提高了0.51%。

4 结 论

本文提出了一种融合局部语义与全局信息的人脸表情识别网络。首先将FCN-8S网络迁移至人脸表情数据集得到语义解析,并提取其中对表情识别有意义的区域及其语义特征。然后与输入图像经过DenseNet121网络得到的全局特征融合得到语义局部特征。此特征经过最大池化层后,与被平均池化的全局特征拼接构造全局语义复合特征,并通过分类器进行分类。本文提出的方法在JAFFE和KDEF数据集上进行了实验,识别准确率分别为93.81%和88.78%。通过消融实验可以看出,人脸表情的语义特征和解冻部分层训练策略在表情识别任务中的重要性。由于本文注重于网络的设计,只采用了简单的交叉熵损失函数,后续将会使用更适合表情分类任务的损失函数。

参考文献:

[1] LIU Y Y, ZENG J B, SHAN S G, et al. Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition[C]//IEEE International Conference on Automatic Face and Gesture Recognition, May 15-19, 2018, Xi'an, China. New York: IEEE, 2018: 458-465.

[2] YOLCU G, OZTEL I, KAZAN S, et al. Facial expression recognition for monitoring neurological disorders based on convolutional neural network[J]. Multimedia Tools and Applications, 2019, 78(22): 31581-31603.

[3] CAI J, MENG Z B, KHAN A S, et al. Island loss for learning discriminative features in facial expression recognition[C]//IEEE International Conference on Automatic Face and Gesture Recognition, May 15-19, 2018, Xi'an, China. New York: IEEE, 2018: 302-309.

[4] WEN Y D, ZHANG K P, LI Z F, et al. A discriminative feature learning approach for deep face recognition[C]// European Conference on Computer Vision, October 08-16, 2016, Amsterdam, Netherlands. Berlin, Heidelberg: Springer-Verlag, 2016: 499-515.

[5] LYONS M J, BUDYNEK J, AKAMATSU S. Automatic classification of single facial images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(12): 1357-1362.

[6] LUNDQVIST D, FLYKT A, ÖHMAN A. The Karolinska directed emotional faces (KDEF)[M]. Stockholm, Sweden: Department of Clinical Neuroscience, Psychology Section, Karolinska Institute, 1998.

[7] SHELLHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[8] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[OL]. (2015-04-10) [2022-11-20]. https://arxiv.org/abs/1409.1556.

[9] JUNG H, LEE S, YIM J, et al. Joint fine-tuning in deep neural networks for facial expression recognition[C]//IEEE International Conference on Computer Vision, December 11-18, 2015, Santiago, Chile. New York: IEEE, 2015: 2983-2991.

[10] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Hawaii, America. New York: IEEE, 2017: 2261-2269.

[11] LE V, BRANDT J, LIN Z, et al. Interactive facial feature localization[C]//European Conference on Computer Vision, October 07-13, 2012, Florence, Italy. Berlin, Heidelberg: Springer-Verlag, 2012: 679-692.

[12] SUN Z, CHIONG R, HU Z P. Self-adaptive feature learning based on a priori knowledge for facial expression recognition[J]. Knowledge-Based Systems, 2020, 204: 106124.

[13] CHEN X M, KE L, DU Q, et al. Facial expression recognition using kernel entropy component analysis network and DAGSVM[J]. Complexity, 2021, 2021: 6616158.

[14] NIU B, GAO Z X, GUO B B. Facial expression recognition with LBP and ORB features[J]. Computational Intelligence and Neuroscience, 2021, 2021: 8828245.

[15] WEI W, JIA Q X, FENG Y L, et al. Multi modal facial expression feature based on deep neural networks[J]. Jour-

nal on Multimodal User Interfaces,2020,14(1):17-23.

[16] SUN Z,CHIONG R,HU Z P. An extended dictionary representation approach with deep subspace learning for facial expression recognition[J]. Neurocomputing,2018,316:1-9.

[17] SANTRA B,MUKHERJEE D P. Local dominant binary patterns for recognition of multi-view facial expressions [C]//Indian Conference on Computer Vision, Graphics and Image Processing,December 18-22,2016,Guwahati, India. New York:ACM,2016.

[18] MINAEE S,MINAEI M,ABDOLRASHIDI A. Deep-emotion: facial expression recognition using attentional convolutional network[J]. Sensors,2021,21(9):3046.

[19] ZHOU Y Q,SHI B E. Action unit selective feature maps in deep networks for facial expression recognition[C]//International Joint Conference on Neural Networks,May 14-19,2017, Anchorage, AK. New York: IEEE, 2017: 2031-2038.

作者简介:

苏　雯　(1992—),女,博士,讲师,硕士研究生助理导师,主要从事语义分割和单目深度估计方面的研究.