·开发研究与工程应用·　　文章编号：1000-3428(2023)02-0254-09　　文献标志码：A　　中图分类号：TP391

基于人脸关键特征提取的表情识别

冉瑞生，翁稳稳，王　宁，彭顺顺

（重庆师范大学 计算机与信息科学学院，重庆 401331）

摘　要：自然场景下人脸表情由于受遮挡、光照等因素影响，以及表情局部变化细微，导致现有人脸表情识别方法准确率较低。提出一种人脸表情识别的新方法，以ResNet18为主干网络，利用残差连接模块加深网络结构，以提取更多深层次的表情特征。通过引入裁剪掩码模块，在训练集图像上的某个区域进行掩码，向训练模型中增加遮挡等非线性因素，提升模型在遮挡情形下的鲁棒性。分别从特征图的通道和空间两个维度提取表情的关键特征，并分配更多的权重给表情变化明显的特征图，同时抑制非表情特征。在特征图输出前加入Dropout正则化策略，通过在训练中随机失活部分神经元，达到集成多个网络模型的训练效果，提升模型泛化能力。实验结果表明，与L2-SVMs、IcRL、DLP-CNN等方法相比，该方法有效提高了表情识别准确率，在2个公开表情数据集Fer2013和RAF-DB上的识别准确率分别为74.366%和86.115%。

关键词：注意力机制；残差网络；人脸表情识别；裁剪掩码；Dropout正则化

开放科学（资源服务）标志码（OSID）：

中文引用格式：冉瑞生，翁稳稳，王宁，等.基于人脸关键特征提取的表情识别[J].计算机工程，2023，49(2)：254-262.

英文引用格式：RAN R S，WENG W W，WANG N，et al.Expression recognition based on the extraction of key facial features[J].Computer Engineering，2023，49(2)：254-262.

Expression Recognition Based on the Extraction of Key Facial Features

RAN Ruisheng，WENG Wenwen，WANG Ning，PENG Shunshun

（School of Computer and Information Science，Chongqing Normal University，Chongqing 401331，China）

【Abstract】The accuracy of existing facial expression recognition methods is typically low owing to the influence of occlusion，illumination，and other factors and to subtle local variations in facial expressions in natural scenes.This study presents a new method for facial expression recognition.A ResNet18 model is adopted as the backbone network，and a residual connection module is employed for a deeper network structure to extract deeper expression features.First，a cutout module is introduced.By masking a certain area of each image in the training set，the model learns to consider several nonlinear factors，such as occlusion，thus improving its robustness to these conditions.The key features of expressions are extracted from the channel and space of the graph，and more weights are assigned to feature maps with obvious expression changes to suppress non-expression features.Finally，prior to the output of the feature map，a Dropout regularization strategy is implemented to randomly deactivate some neurons and integrate multiple network models to improve the generalization ability of the model.The experimental results show that the proposed method exhibits improved accuracy in expression recognition tasks compared with L2-SVMs，IcRL，DLP-CNN，and other methods. Recognition accuracy values of 74.366% and 86.115% are achieved on two public expression datasets，Fer2013 and RAF-DB，respectively.

【Key words】attention mechanism；residual network；facial expression recognition；cropping mask；Dropout regularization

DOI：10. 19678/j. issn. 1000-3428. 0063715

0　概述

面部表情是人体语言的一部分，是对心理情感的一种表露形式，是情感传递的重要方式。美国传播学家MEHRABIAN通过实验提出，在情绪的表达中，面部表情所占比重高达55%[1]。由此可见，人脸表情识别（Facial Expression Recognition，FER）是非常具有研究价值的课题。1971年，心理学家EKMAN把基本表情划分为6种，分别为开心、伤心、惊讶、害怕、生气和厌恶[2]，尽管不同人之间有所差异，但这些表达情感的方

基金项目：重庆市技术创新与应用发展专项面上项目（cstc2020jscx-msxmX0190）；重庆市教委科学技术研究计划项目（KJZD-K202100505，KJQN202100515）。

作者简介：冉瑞生（1976—），男，教授、博士，主研方向为机器学习、计算机视觉；翁稳稳、王　宁，硕士研究生；彭顺顺（通信作者），讲师、博士。

收稿日期：2022-02-07　　**修回日期：**2022-03-09　　**E-mail：**pss@cqnu.edu.cn

式是人类共有的。

传统的表情识别方法主要是通过人工设计特征并结合分类模型达到表情识别的目的。局部二值模式（Local Binary Pattern，LBP）[3]、Gabor小波变换[4]以及尺度不变特征变换（Scale-Invariant Feature Transform，SIFT）[5]等常被用来提取特征，再结合支持向量机（Support Vector Machine，SVM）[6]等分类模型来识别表情。在早期，这些经典的特征提取算法在一些表情数据集上取得了不错的效果[7-8]，但也有很多缺点，主要表现为人工设计特征非常复杂、耗时和性能较低，对于实验室环境下的表情数据集，表情变化单一且不受自然环境干扰，不同表情之间的差异明显，因此基于人工设计特征的方法可以取得不错的效果。但自然场景下的人脸表情受到光照、遮挡、不同种族、年龄、性别等因素的影响，表情特征复杂，传统的表情识别方法效果很差。

随着深度学习的崛起，卷积神经网络（Convolutional Neural Network，CNN）凭借自身强大的特征提取能力被广泛应用在计算机视觉领域，执行图像分类、目标检测等任务。人脸表情识别也属于图像分类任务的一种，因此许多经典的卷积神经网络模型，如LeNet[9]、VGG[10]、ResNet[11]等常被作为基础网络用在人脸表情识别任务上，并在此基础上进行改进优化，从而达到提升模型识别准确率的目的。例如，文献[12]通过深度学习网络来提取特征，并采用L2正则化和支持向量机相结合的方式替代Softmax函数，提升了模型人脸表情识别准确率。文献[13]提出一种新的学习方法IcRL，通过提取独立的表情特征来学习不同类别表情之间的相互关系，并扩大类间距离与类内距离之比。文献[14]基于残差网络ResNet18，将过滤器响应正则化、批量正则化、实例正则化和组正则化进行组合，并分别嵌入网络之中，平衡和改善特征数据分布，提升模型性能。文献[15]提出一种新的深度位置保持卷积神经网络DLP-CNN，目的是通过增强保留局部性来提高深层特征的判别能力，同时最大化类间离散度。文献[16]提出一种基于注意力机制的卷积神经网络（ACNN），可以感知人脸的遮挡区域，并关注最具鉴别性的未遮挡区域，针对不同的关注区域，提出基于局部的ACNN（PACNN）和基于全局人脸区域的ACNN（GACNN）。文献[17]提出一种新颖的深度嵌入方法，该方法的目的是设计学习判别性表情特征同时表示大量类内变化的表情特征，通过最小化样本与其最近的子类中心之间的距离来形成局部紧凑的表示空间结构，最终提升模型性能。

也有针对人脸面部遮挡等因素设计的表情识别方法，如文献[18]提出一种新颖的生成对抗网络用于遮挡表情识别，在加权重建损失、三元组损失和对抗损失的三重约束下，生成器自然地补充了表情图像中的遮挡，再利用2个判别器来区分图像真假以及完成表情分类。文献[19]通过结合残差网络和VGG16网络，提出基于改进VGG16的20层卷积神经网络，并采用混合特征融合策略将Gabor滤波器与改进网络并行化，通过实验验证了方法的优势。

由此可见，为了让表情识别模型在自然场景下具有良好的鲁棒性以及较高的识别率，必须让模型具有提取复杂特征的能力（如表情局部变化细微的特征以及面部表情遮挡的区域），以及能够提取反映表情变化的关键特征，抑制非表情特征。

本文以残差网络作为主要的特征提取网络，在网络输入端通过引入裁剪掩码模块（Cutout）[20]，扩充训练数据的复杂性。在残差单元的最前端引入关键特征表征模块，即使用卷积块注意力模块（Convolutional Block Attention Module，CBAM）[21]，在空间和通道维度上提取表情的关键特征。最后在网络输出端引入Dropout正则化技术[22]，提升模型的泛化能力。将本文方法在两个公开数据集Fer2013[23]和RAF-DB[15]上进行实验，以验证该方法的有效性。

1　相关技术

本文提出的人脸表情识别方法使用残差网络ResNet18作为基础网络，由裁剪掩码、关键特征表征以及Dropout正则化这3部分构成，本文方法的流程如图1所示。

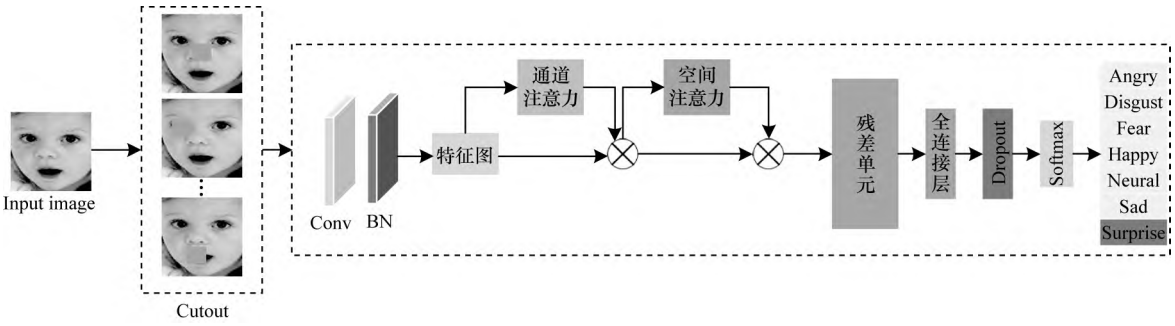

图1　本文方法的流程

Fig.1　Procedure of method in this paper

由图1可知，裁剪掩码单元能够模拟遮挡数据，将输入图像中一部分保持原样，另外一部分随机擦除一个矩形区域，从而增强数据集。在使用残差网络粗略提取特征后，将其送入残差单元之前在模型中加入关键特征表征模块，关键特征表征模块主要由通道注意力和空间注意力构成，用于提取人脸的关键特征，

让残差单元学习到更加精细的特征。最后在特征图输出前，使用Dropout正则化策略，达到组合不同训练模型的目的，提升模型泛化能力。

1.1 关键特征表征

近年来，注意力机制被广泛应用于各种视觉任务中，注意力机制的核心思想是帮助网络选择视觉区域中最重要的特征，并集中关注它。最常使用的有通道注意力机制以及混合注意力机制（空间和通道结合）。其中通道注意力机制最具代表性的网络是通道注意力网络（Squeeze and Excitation Networks，SENet）[24]，通过计算输入特征图每个通道的权值，让网络学习更多重要的特征，从而提升模型性能，最终在图像分类任务上取得了显著效果。在许多任务中，空间位置信息具有不同的作用，尤其在表情识别中，嘴巴、眼睛等区域的重要性程度明显更大。为弥补通道注意力的缺陷，混合注意力机制又增加了空间注意力，从而在特征提取时也关注特征图上的空间位置。CBAM[21]是混合注意力机制最具代表性的网络，通过串联通道注意力和空间注意力，在图像分类任务中相比SENet，取得了较好的效果。本文关键特征表征模块就是利用混合注意力机制[21]进行特征选择，使特征表达更加准确。

1.1.1 通道注意力

将特征图 $F \in \mathbb{R}^{C\times H\times W}$ 输入通道注意力模块中（其中：C 为通道数；H 为特征图高度；W 为特征图宽度）。首先使用最大池化和平均池化对输入特征图进行压缩，得到2个特征向量 $\boldsymbol{F}^{C}_{\max}$ 和 $\boldsymbol{F}^{C}_{\text{avg}}$，分别表示最大池化特征和平均池化特征。然后将 $\boldsymbol{F}^{C}_{\max}$ 和 $\boldsymbol{F}^{C}_{\text{avg}}$ 送入包含一个隐藏层的多层感知机（Multi-Layer Perceptron，MLP）里，得到2个 $1\times1\times C$ 的通道注意力特征图，其中为了减少参数量，隐藏层的神经元个数为 C/r（r 是压缩比例）。将多层感知机输出的两个通道注意力特征图进行元素求和并通过激活函数Sigmoid，最终得到具有特征聚合性的通道注意力图 $M_{\text{c}} \in \mathbb{R}^{C\times1\times1}$。将输入特征与最终得到的通道注意力图相乘即可得到经过通道注意力表征过后的新特征，具体计算式如式（1）所示：

$$M_{\text{c}}(\boldsymbol{F})=\sigma(\text{MLP}(\text{AvgPool}(\boldsymbol{F}))+\text{MLP}(\text{MaxPool}(\boldsymbol{F})))= \sigma(W_1(W_0(\boldsymbol{F}^{C}_{\text{avg}}))+W_1(W_0(\boldsymbol{F}^{C}_{\max}))) \tag{1}$$

其中：σ 代表Sigmoid激活函数；$W_0 \in \mathbb{R}^{C/\text{r}\times C}$ 和 $W_1 \in \mathbb{R}^{C\times C/r}$ 分别为多层感知机的权重，在参数 W_0 后加入ReLU激活函数，向模型中加入更多非线性因素。通道注意力模块结构如图2所示。

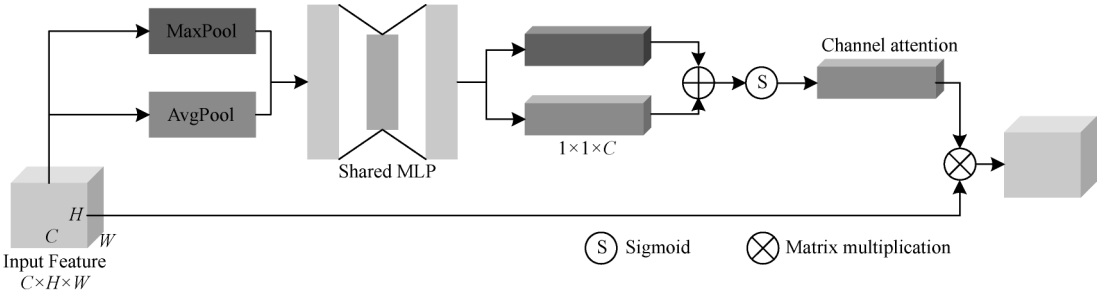

图2 通道注意力模块的结构

Fig.2 Structure of channel attention module

1.1.2 空间注意力

输入一个特征图 $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$，沿通道方向分别使用最大池化和平均池化生成两个二维特征图 $\boldsymbol{F}^{S}_{\max} \in \mathbb{R}^{1\times H\times W}$ 和 $\boldsymbol{F}^{S}_{\text{avg}} \in \mathbb{R}^{1\times H\times W}$，采用通道维度级联的方式将这两个特征图进行合并，生成新的特征图 $\boldsymbol{F}^{S}_{[\text{max,avg}]} \in \mathbb{R}^{2\times H\times W}$。然后使用一个 7×7 大小，填充设置为3的卷积核，作用于新的特征图 $\boldsymbol{F}^{S}_{[\text{max,avg}]}$，并通过Sigmoid激活函数后生成最终的空间注意力图 $M_S \in \mathbb{R}^{H\times W}$。与输入的特征图 $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ 进行相乘即可获得空间位置的关键特征表征图。具体计算式如式（2）所示：

$$M_S(\boldsymbol{F})=\sigma(f^{7\times7}([\text{AvgPool}(\boldsymbol{F});\ \text{MaxPool}(\boldsymbol{F})]))= \sigma(f^{7\times7}([\boldsymbol{F}^{S}_{\text{avg}};\ \boldsymbol{F}^{S}_{\max}])) \tag{2}$$

其中：σ 代表Sigmoid激活函数；$f^{7\times7}$ 代表卷积核大小为 7×7 的标准卷积层。空间注意力模块的结构如图3所示。

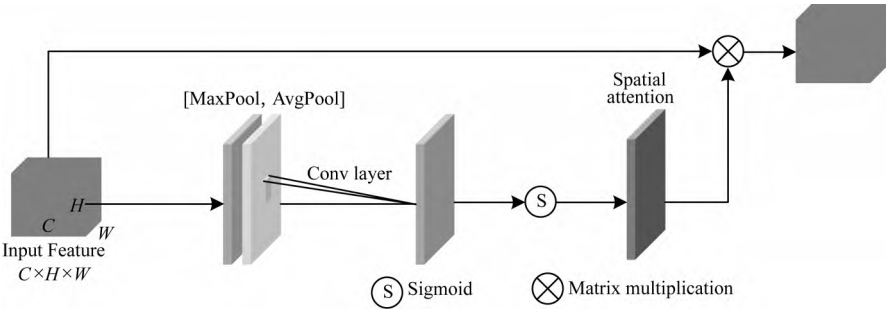

图3 空间注意力模块的机构

Fig.3 Structure of spatial attention module

1.2　裁剪掩码

为提升模型的泛化能力，常常需要对原始输入数据做数据增强处理，例如翻转、镜像变换等操作，但这些操作都是在原数据基础上进行简单的线性变换，并不能带来更多的数据复杂性。因此为了让模型训练过程中学习更多接近真实环境下的人脸表情，本文引入了裁剪掩码模块（Cutout）[20]。

文献[20]对图像进行裁剪掩码有两种思路，最早是通过可视化技术获取输入图像的重要视觉特征，然后把这一部分进行掩码。但是在实验过程中，作者发现该方法与直接随机掩码图像中一部分特征的差别并不大，而且前者还引入了额外的重要特征计算，因此舍弃了这种方法，并且论文中也指出裁剪掩码的区域大小比裁剪的形状更重要。因此本文使用裁剪掩码模块（Cutout），借鉴文献[20]中的第2种思路，在输入图像中进行随机裁剪，掩码形状只需是正方形。具体操作是利用固定大小的正方形对图像进行遮挡，在正方形范围内，所有值都被设置为0或者其他纯色值。裁剪掩码的算法过程主要有以下4步。

步骤1　输入参数n_holes、l_{length}和i_{img}。其中第1个参数为掩码单元的个数，第2个参数为掩码正方形像素边长，第3个参数为输入图像像素矩阵。

步骤2　根据输入图像img获取图像的高（H）和宽（W），并生成一个二维矩阵$\boldsymbol{m}_{\text{mask}}^{H\times W}$，其中元素全部赋值为1。

步骤3　根据n_holes值进行遍历，生成掩码矩阵。计算式如式（3）所示：

$$\begin{aligned}&x_i=\text{Rand}(0,W)\\&y_i=\text{Rand}(0,H)\\&X_i=\min(\max(0,x_i\pm l_{\text{length}}/2),W)\\&Y_i=\min(\max(0,y_i\pm l_{\text{length}}/2),H)\\&\text{mask}[X_1:Y_1,X_2:Y_2]=0\end{aligned}\tag{3}$$

步骤4　把输入图像的像素矩阵与掩码矩阵进行矩阵点乘获得最终图像，计算式如式（4）所示：

$$\boldsymbol{i}_{\text{Cutout}}^{\text{img}}=\boldsymbol{i}_{\text{img}}\otimes\boldsymbol{m}_{\text{mask}}^{H\times W}\tag{4}$$

1.3　Dropout正则化

在使用深度卷积神经网络时，为防止网络过拟合，往往需要使用大量数据进行训练，但在实际中大数据集的标注需要大量时间和资源。为解决这一问题，文献[22]提出一种Dropout正则化策略，通过阻止特征检测器的共同作用来提高神经网络的性能。以下主要介绍Dropout正则化在训练和测试阶段的大致过程。

训练时，首先随机（临时）删除网络中一部分隐藏神经元，输入输出神经元保持不变，其次把输入特征通过修改后的网络前向传播，然后把计算出的损失结果通过网络反向传播回去，并在没有删除的神经元上根据优化策略更新连接参数，被随机删除的神经元不会参与本次前向传播的计算。如图4所示为标准神经网络（左）与加入Dropout后的神经网络（右）的前向传播结构。测试时，为保证模型输出结果的稳定性，并且让测试数据和训练数据总体一致，需要在测试阶段时乘以丢弃权重p，即测试时权重必须进行缩放，测试时的权重参数为$W_{\text{test}}^{(l)}=p\times W^{(l)}$。

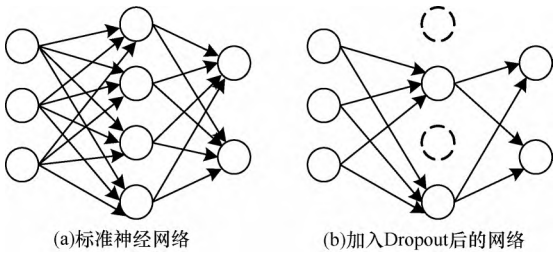

图4　神经网络的结构对比

Fig.4　Structure comparison of neural network

2　本文方法

为提高在自然场景下的人脸表情识别率，本文提出基于人脸关键特征提取的表情识别方法。首先经过裁剪掩码模块得到真实场景下具有遮挡因素的表情图像，其次利用关键特征表征模块来帮助网络提取更加精细的表情特征，然后结合Dropout正则化策略帮助网络融合多次训练结果，提升模型泛化能力。其中基础网络使用残差网络ResNet18，本文使用2种结构的残差单元来提取特征，如图5所示。使用两种残差模块是为了保持特征尺寸一致，残差模块1提取特征前后特征图尺寸没有发生变化，残差模块2的捷径连接需要让输入特征尺寸和输出特征尺寸一致，因此使用步长为2的卷积。

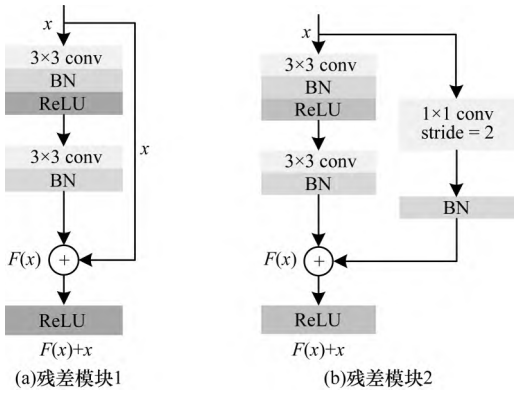

图5　不同残差单元结构

Fig.5　Different residual unit structures

本文所提模型主要由输入层裁剪掩码、特征提取层（一个标准卷积、一个通道注意力和空间注意力、一个残差模块1和3个残差模块2）、全局平均池化、全连接层以及丢弃权重p=0.5的Dropout层组成，最后使用Softmax Loss进行表情分类损失的计算。

在现实场景中，遮挡问题一直以来是一个难题。为使模型能够处理更多具有遮挡人脸表情的数据，提升对遮挡人脸表情识别的能力，本文引入了裁剪掩码单元（Cutout），通过模拟现实场景中人脸表情遮挡数据，提升模型学习能力。图6所示为输入图像进行裁剪掩码后的示例图，在模型训练过程中裁剪区域是随机产生的。本文将裁剪掩码模块中n_holes（掩码个

数)和l_{length}(掩码正方形边长)分别设置为1和16。

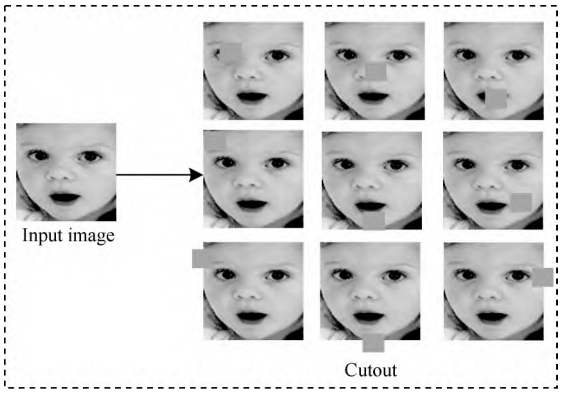

图6 裁剪掩码示意图

Fig.6 Schematic diagram of clipping mask

在使用卷积操作提取特征的过程中,特征图的通道数会逐渐增加,而其中每个通道的特征对于关键信息的贡献是不一样的。有的通道存在大量的关键特征,而有的通道有用信息少,因此会产生冗余特征,导致模型性能降低。为解决该问题,本文采用通道注意力机制,使用不同的池化策略并行计算,压缩特征图所产生的权重,并与输入特征图点乘,从而给予每个特征通道不同的权重。

在提取人脸表情特征时,人脸五官的位置也具有一定的空间关系,而不同五官的特征对于表情的影响程度不同,如嘴巴区域肌肉的变化比鼻子、眉毛等区域的特征更多,但传统卷积操作对于空间位置的特征提取使用相同的方法。为解决该问题,本文引入空间注意力机制,使网络可以学习到特征图空间位置之间的关系。将通道注意力与空间注意力融合后的结构如图7所示。

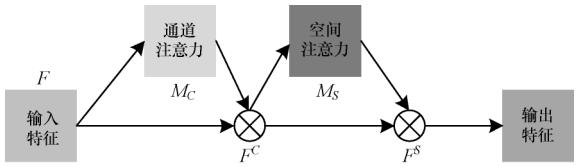

图7 混合注意力结构

Fig.7 Hybrid attention structure

通过融合通道注意力和空间注意力模块,可以同时获取关键特征通道和特征间的位置关系,从而使模型提取的表情特征表征更加准确。输入特征$\boldsymbol{F}\in\mathbb{R}^{C\times H\times W}$,经过通道注意力模块后得到新的特征$\boldsymbol{F}^C$,再把该特征输入到空间注意力模块中得到最终的关键特征表征$\boldsymbol{F}^S$,具体特征计算式如式(5)所示:

$$\boldsymbol{F}^C=M_C(\boldsymbol{F})\otimes\boldsymbol{F}$$
$$\boldsymbol{F}^S=M_S(\boldsymbol{F}^C)\otimes\boldsymbol{F}^C \tag{5}$$

其中:$\otimes$表示矩阵同位素点乘运算。

通过引入裁剪掩码和关键特征表征模块后,模型提取的特征更具代表性,表情区分度更高。但在人脸表情识别中,各类表情的种类单一且数量较少,使用深度学习模型训练时容易造成过拟合,且表情识别精度也会降低。因此,本文在裁剪掩码和关键特征表征模块后,在网络末端又加入了Dropout策略,这样可以起到2个作用:1)训练时前向传播随机失活部分神经元,有利于加快模型训练速度,并减少相邻神经元间的过渡依赖,有效解决网络过拟合问题;2)在多次迭代训练时,随机失活神经元不同,可以达到类似训练不同模型的效果,多个结果相互修正,最终提升模型识别准确率。本文方法的网络结构如图8所示。

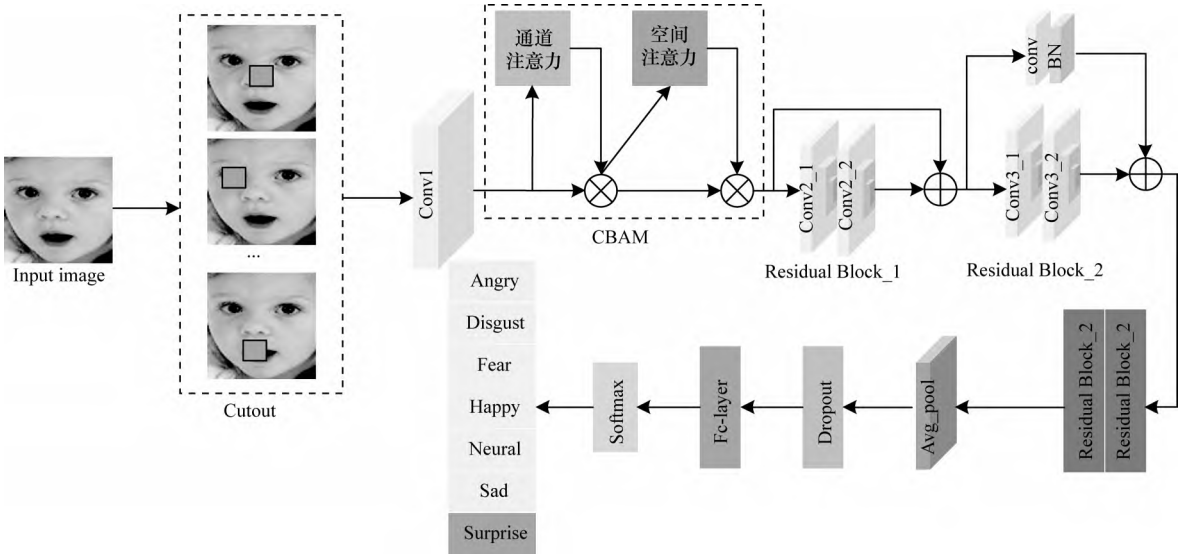

图8 本文方法的网络结构

Fig.8 Network structure of method in this paper

3 实验结果与分析

本文实验使用深度学习框架PyTorch,版本1.10.0,编程语言为Python3.7,操作系统为Ubuntu 18.04.5,显卡型号为NVIDIA RTX3090。实验中保持超参数一致,使用随机梯度下降算法(Stochastic

Gradient Descent，SGD）对交叉熵损失优化，动量设置为0.09，衰减系数设置为0.000 5，初始学习率设置为0.01，在训练80次后学习率开始衰减，总共迭代次数（epoch）设置为300，将每次训练完成后测试集上准确率最高的参数作为模型精度。在实验过程中为了使模型训练达到最优，对训练集数据进行了数据增强。本文使用10-crop数据增强的手段，使数据量得到扩充，其具体做法是将尺寸为48×48像素的原始图像进行裁剪，分别从图像正中间、右上角、右下角、左上角和左下角进行裁剪，最后生成5张尺寸为44×44像素的图像，然后把得到的图像进行镜像操作，使训练数据被扩充为原来的10倍。

3.1　表情数据集介绍

本文实验使用2个公开人脸表情数据集进行评估，分别是Fer2013和RAF-DB数据集，均为真实场景下的数据库，均包含7种基本表情，包括惊讶、害怕、厌恶、开心、伤心、生气和自然。

Fer2013[23]数据集是2013年Kaggle比赛使用的人脸表情数据集，图像均是使用谷歌人脸识别接口从网上获取，人脸角度较多且有遮挡，涵盖不同年龄段的人，且男性和女性各占一定比例，符合自然条件下的表情分布。其主要由35 886张不同表情图像组成，其中训练集有28 708张，验证集和测试集各3 589张，每张图像的大小是48×48像素。

RAF-DB[15]是一个真实世界人脸表情数据集，该数据集从互联网上下载了大约30 000张面部图像，图像大小均为100×100像素。本文手动将图像缩放到48×48像素。数据库包含单标签子集和双标签子集两个不同子集。单标签子集包括7类基本情绪和边界框，该数据集中的受试者年龄从0～70岁不等，包括52%的女性，43%的男性，还有5%的不确定。对于种族分布，高加索人占77%，非裔美国人占8%，亚洲人占15%。并且数据集中的大量图像具有遮挡、姿态等变化，符合自然场景下的表情分布。本文主要使用7类基本表情，共15 339张图像作为实验数据集，其中包括12 271张训练集图像和3 068张测试集图像。

图9是本文所使用数据库的示例图像，第1排为Fer2013数据库示例图，第2排为RAF-DB数据库示例图。从图9可知，这2个数据集中存在大量遮挡、光照、性别等变化的图像。此外，两个数据集上每种表情数量的统计数据如图10所示。

图9　本文数据集示例

Fig.9　Examples of datasets in this paper

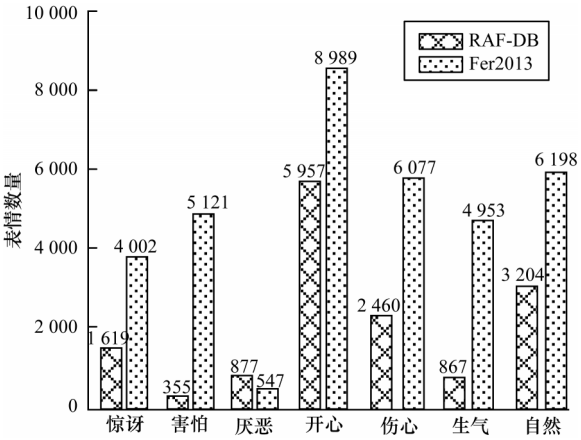

图10　Fer2013和RAF-DB数据集的表情数据分布

Fig.10　Distribution of expression data of Fer2013 and RAF-DB datasets

3.2　结果分析

为验证本文方法的可靠性，在Fer2013与RAF-DB数据集上进行了实验验证，并与当前先进的人脸表情识别方法进行比较。此外，为说明本文方法对各类表情的识别效果，使用最终训练模型在测试集上生成的混淆矩阵进行分析。最后，为验证本文提出的各个模块的有效性，进行了消融实验。

3.2.1　网络模型有效性验证

本文所提模型在2个公开数据集上的混淆矩阵结果如图11所示。图11（a）是在Fer2013数据库上的实验结果，从中可以看出对于“高兴”和“惊讶”2个特征变化明显的表情，模型准确率达到了较高水平，分别为93%和84%。图11（b）是在RAF-DB数据库上的实验结果，该数据集上的图像质量较Fer2013数据集好，因此整体模型识别准确率相对较高，实验结果中有4种表情识别率都超过了80%，模型能较好地识别各种表情。

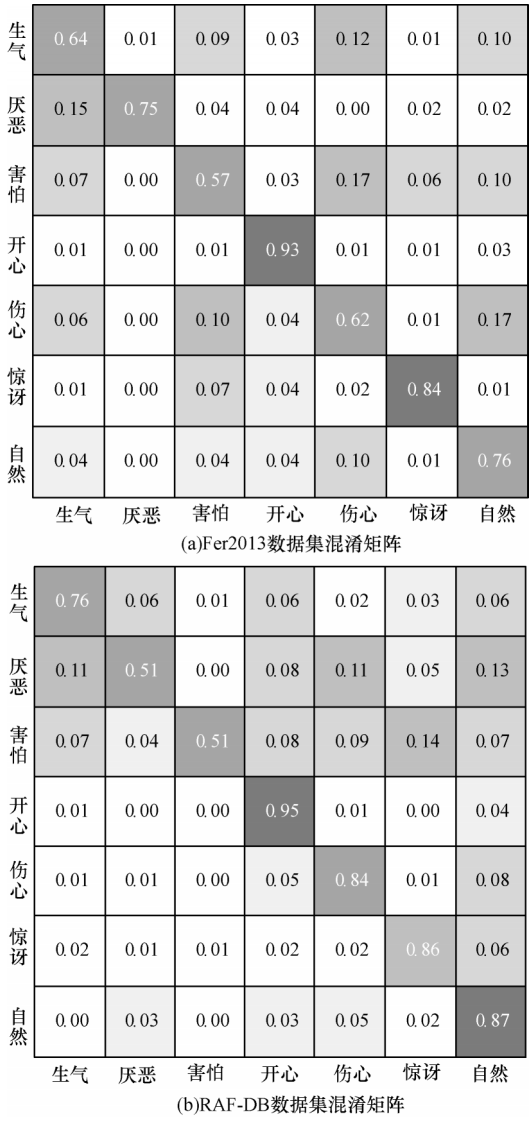

图11 2个公开数据集上的混淆矩阵结果

Fig.11 Confusion matrix results on the two public datasets

3.2.2 与现有方法的对比

为进一步证明本文所提方法的识别性能，在Fer2013与RAF-DB数据集上与当前已有的先进方法进行对比，实验结果如表1和表2所示。从表中可知，各种方法在本文所使用的数据集中都取得了较高的识别率。其中文献[25]提出的一种注意力分层双线性池化残差网络，采用有效的通道注意力机制显式地建模各通道的重要程度，并引入双线性池化层来捕获层间部分特征关系，该方法在Fer2013数据集上取得了73.840%的识别准确率。文献[26]提出的一种双通道遮挡感知神经网络模型，分别使用VGG和ResNet网络来学习遮挡表情特征和全脸特征，将两种特征融合后在RAF-DB数据集上取得了86%的识别准确率。本文方法是在残差网络提取全局特征的前提下，通过引入通道注意力和空间注意力来提取图像浅层的关键特征，为模型增加了更多精细化特征。另外引入的裁剪掩码是通过随机掩码输入图像，手动向网络中添加非线性因素，迫使模型在真实环境数据集上学习更多遮挡表情特征。最后使用Dropout正则化，使模型融合学习参数，提升模型的泛化性。本文方法在Fer2013和RAF-DB数据集上分别取得了74.366%和86.115%较高的识别准确率，与对比方法相比，准确率最高，验证了本文方法的有效性。

表1 不同方法在Fer2013数据集下的识别准确率对比

Table 1 Comparison of recognition accuracy of different methods under Fer2013 dataset %

方法	准确率
L2-SVMs[12]	71.200
IcRL[13]	72.360
文献[14]方法	73.558
AHBPRN[25]	73.840
本文方法	74.366

表2 不同方法在RAF-DB数据集下的识别准确率对比

Table 2 Comparison of recognition accuracy of different methods under RAF-DB dataset %

方法	准确率
DLP-CNN[15]	84.130
GACNN[16]	85.070
Subclass CL[17]	84.260
.文献[26]方法	86.000
本文方法	86.115

3.2.3 消融实验

为测试本文所引入的裁剪掩码和关键特征表征模块CBAM对网络的有效性，进行了交叉对比实验，实验结果如表3所示。以融合了Dropout正则化策略的残差网络ResNet18作为基础网络(Base)，分别向模型中加入裁剪掩码以及关键特征表征模块CBAM后作对比实验。从表3可知，在基础网络中单独加入裁剪掩码和关键特征表征模块都能提升网络性能，当2个模块同时加入网络时，准确率提升最为显著，在Fer2013和RAF-DB数据集上比基础网络分别提升了约1.34和0.99个百分点。由此可以推断在加入关键特征表征模块后可以使基础模型提取的表情特征更加精细化，从而提升模型识别率。可见本文引入的各个模块对于基础网络都是有效的，并且能够共同促进网络性能的提升。

表3 不同模块的识别准确率对比

Table 3 Comparison of recognition accuracy of different modules %

模块	准确率	
	Fer2013数据集	RAF-DB数据集
Base	73.029	85.137
Base+Cutout	74.115	85.267
Base+CBAM	74.115	85.561
Base+Cutout+CBAM	74.366	86.115

3.2.4　遮挡表情验证

为了让本文模型学习到人脸表情被遮挡的情形，从而更接近真实环境下的人脸表情，本文引入了裁剪掩码模块。此外，为验证本文方法在具有遮挡情形下的人脸表情识别效果，本文在CK+数据集[7]上利用裁剪掩码模块[20]随机在人脸图像上添加遮挡，从而模拟具有遮挡的表情。然后用本文方法进行表情识别。图12给出了各种表情的遮挡图像，以及用本文方法预测的结果。图13给出了本文方法在CK+数据集上实验所得的混淆矩阵。从图12、图13可以看出，遮挡住人脸较关键的部位如嘴巴、眼睛等之后，本文模型仍能准确识别出图像的真实表情。此外，本文方法在CK+数据集上取得93.939%的准确率，说明本文方法对于遮挡图像仍然具有较高的识别率，具有一定鲁棒性。

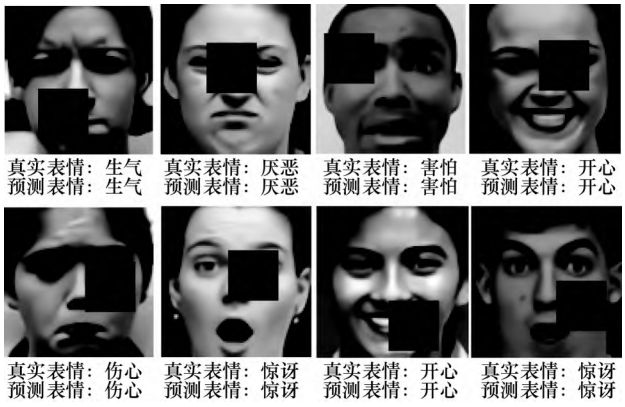

图12　本文方法预测遮挡表情的结果示例

Fig.12　Example of the results of method in this paper to predict occlusion expressions

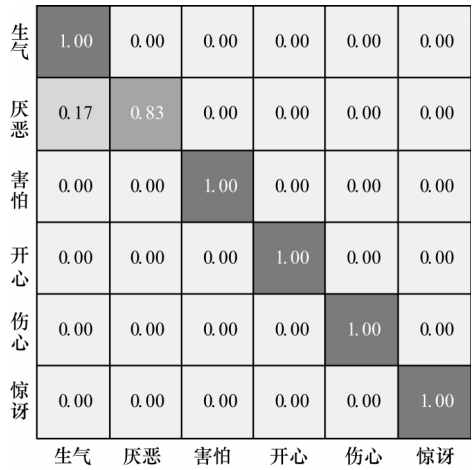

图13　本文方法在CK+数据集下的混淆矩阵

Fig.13　Confusion matrix of method in this paper under CK+ dataset

本文方法也存在一定的局限性。从图11的混淆矩阵可以看出，本文模型对个别表情的识别准确率较低，且存在表情相互识别错误的情形。比如在Fer2013数据集上，“害怕”表情的识别准确率较低。这主要是由于Fer2013数据集上存在大量低质量图像以及非人脸图像。Fer2013数据集是评估实验中最难的数据集，该数据集上人工正常识别率仅为65%左右。此外，相似表情易发生混淆也是原因之一，例如“害怕”和“伤心”、“厌恶”和“生气”表情在现实中并非单一发生，生气的情绪会产生厌恶，害怕会导致伤心，因此利用静态表情图像进行识别是较难的。另外从图10可知，部分表情数量较少（如厌恶等），因此模型很难学到相关表情特征，导致识别准确率较低。

本文也分析了表情识别失败的案例。图14给出了本文方法在RAF-DB数据集下识别失败的案例。经分析可知，这可能是由于部分图像的质量太低；有些图像的表情表达特别隐晦，容易造成误判；有些图像中表情明显的区域被完全遮挡，模型无法提取到特征。这时可能需要结合人脸姿态、手势等进行表情判别。

图14　本文方法在RAF-DB数据集下识别失败的案例

Fig.14　Identifies failure case of method in this paper under RAF-DB dataset

4　结束语

针对自然场景下人脸表情受遮挡、光照等因素影响，以及表情局部变化细微，导致现有人脸表情识别准确率较低的问题，提出一种基于人脸关键特征提取的表情识别方法。通过引入裁剪掩码模块，使模型能有效提取遮挡表情特征。在此基础上使用关键特征表征模块使模型在通道和空间维度上引导网络学习更多关键特征，提高模型区分表情局部细微变化的能力及鲁棒性。最后在网络末端加入Dropout正则化，有效缓解过拟合，提升模型的识别性能。在两个自然场景下的人脸表情数据集Fer2013和RAF-DB上的实验结果表明，本文方法与L2-SVMs、IcRL、DLP-CNN等方法相比，表情识别准确率得到有效提升。但该方法存在部分表情识别率较低、个别表情之间误判的问题，下一步将在保证识别准确率的前提下，通过研究动态序列的人脸表情识别，提升人脸表情识别方法在自然场景下的识别准确率及在低质量图像等情形下的鲁棒性。

参考文献

[1] MEHRABIAN A, RUSSELL J A. An approach to environmental psychology[M]. Cambridge, USA: MIT Press, 1974.

[2] EKMAN P. Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique[J]. Frontiers in Cell and Developmental Biology, 1994, 115(2): 268-287.

[3] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.

[4] ZHANG Z Y, MU X M, GAO L. Recognizing facial expressions based on Gabor filter selection[C]// Proceedings of the 4th International Congress on Image and Signal Processing. Washington D. C., USA: IEEE Press, 2011: 1544-1548.

[5] LOWE D G. Object recognition from local scale-invariant features[C]//Proceedings of the 7th IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 1999: 1150-1157.

[6] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[7] 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]//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. Washington D. C., USA: IEEE Press, 2010: 94-101.

[8] LYONS M, AKAMATSU S, KAMACHI M, et al. Coding facial expressions with Gabor wavelets[C]//Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition. Washington D. C., USA: IEEE Press, 1998: 200-205.

[9] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[10] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2021-12-01]. http://arXiv:1409.1556.

[11] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 770-778.

[12] TANG Y C. Deep learning using linear support vector machines[EB/OL]. [2021-12-01]. http://arXiv:1306.0239.

[13] CHEN Y Z, HU H F. Facial expression recognition by inter-class relational learning[J]. IEEE Access, 2019, 7: 94106-94117.

[14] 兰凌强,李欣,刘淇缘,等. 基于联合正则化策略的人脸表情识别方法[J]. 北京航空航天大学学报,2020,46(9):1797-1806.
LAN L Q, LI X, LIU Q Y, et al. Facial expression recognition method based on a joint normalization strategy[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1797-1806. (in Chinese)

[15] LI S, DENG W H, DU J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 2584-2593.

[16] LI Y, ZENG J B, SHAN S G, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE Transactions on Image Processing, 2019, 28(5): 2439-2450.

[17] LUO Z M, HU J N, DENG W H. Local subclass constraint for facial expression recognition in the wild[C]// Proceedings of the 24th International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 3132-3137.

[18] LU Y, WANG S G, ZHAO W T, et al. WGAN-based robust occluded facial expression recognition[J]. IEEE Access, 2019, 7: 93594-93610.

[19] CHEN Y J, LIU S G. Deep partial occlusion facial expression recognition via improved CNN[M]//Advances in Visual Computing. Berlin, Germany: Springer, 2020: 451-462.

[20] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout[EB/OL]. [2021-12-01]. http://arXiv:1708.04552.

[21] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 3-19.

[22] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[EB/OL]. [2021-12-01]. https://arxiv.org/abs/1207.0580.

[23] DHALL A, GOECKE R, LUCEY S, et al. Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark[C]//Proceedings of IEEE International Conference on Computer Vision Workshops. Washington D. C., USA: IEEE Press, 2011: 2106-2112.

[24] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 7132-7141.

[25] 张爱梅,徐杨. 注意力分层双线性池化残差网络的表情识别[J]. 计算机工程与应用,2020,56(23):161-166.
ZHANG A M, XU Y. Attention hierarchical bilinear pooling residual network for expression recognition[J]. Computer Engineering and Applications, 2020, 56(23): 161-166. (in Chinese)

[26] 王军,赵凯,程勇. 基于遮挡感知卷积神经网络的面部表情识别模型[J]. 计算机工程,2021,47(10):242-251.
WANG J, ZHAO K, CHENG Y. Facial expression recognition model based on convolutional neural network with occlusion perception[J]. Computer Engineering, 2021, 47(10): 242-251. (in Chinese)

编辑 赖玉玲