ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 6,Nov. 2023,pp. 1422－1433
DOI:10. 16337/j. 1004-9037. 2023. 06. 016

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

# 基于多区域检测网络的复杂场景面部表情识别

潘新辰，秦　岭，杨小健

(南京工业大学计算机科学与技术学院,南京 211816)

**摘　要**：面部表情是人类情绪状态的最直观表现,卷积神经网络在面部表情识别上表现出了优异的性能。然而复杂场景下遮挡和姿势变化仍是面部表情自动识别的两个主要问题,它们会显著改变人脸的外观,从而影响最终的识别结果。针对面部表情识别中遮挡和姿势变化的问题,提出了一种基于双注意力和多区域检测网络的面部表情识别方法。双注意力用于提升整体网络的特征提取能力,使网络能够关注更详细的特征信息。多区域检测用于在遮挡和姿态变化的面部表情识别中自适应地捕捉重要的局部区域,抑制遮挡和姿势变化带来的负面影响。最终在AffectNet、RAF-DB和SFEW三种公开的自然场景面部表情数据集上验证了所提方法的有效性。

**关键词**：面部表情识别;卷积神经网络;注意力机制;多区域检测;深度学习

**中图分类号**：TP391.4　　**文献标志码**:A

## Facial Expression Recognition Under Complex Scenes Based on Multi-region Detection Network

PAN Xinchen, QIN Ling, YANG Xiaojian

(College of Computer Science and Technology, Nanjing University of Technology, Nanjing 211816,China)

**Abstract**: Facial expressions are the most intuitive representation of human emotional states, and convolutional neural networks have shown excellent performance in facial expression recognition. However, occlusion and pose changes in complex scenes are still two major problems in automatic facial expression recognition, which significantly changes the appearance of faces and affects the final recognition results. Aiming at the problems of occlusion and pose change in facial expression recognition, a facial expression recognition method based on dual attention and multi-region detection network is proposed. Dual attention is used to improve the feature extraction capability of the overall network, enabling the network to focus on more detailed feature information. Multi-region detection is used to adaptively capture important local regions in facial expression recognition of occlusion and pose changes, and suppress the negative effects of occlusion and pose changes. Finally, the effectiveness of the proposed method is verified on three public natural scene facial expression datasets AffectNet, RAF-DB and SFEW.

**Key words:** facial expression recognition; convolutional neural network; attention mechanism; multi-region detection; deep learning

---

**基金项目**:国家自然科学基金(61672279)。
**收稿日期**:2022-09-05;**修订日期**:2023-01-15

## 引　言

面部表情在人类日常交流中发挥着重要作用,自动面部表情识别在服务机器人、驾驶员疲劳检测和智能辅导等人机交互系统中具有重要的应用价值,在计算机视觉领域引起了越来越多的关注。面部表情识别(Facial expression recognition, FER)传统方法主要利用局部二值模式、方向梯度直方图及Haar-like等方法进行特征提取,然后再利用支持向量机、K近邻及AdaBoost等方法进行分类。近几年,随着深度学习的飞速发展,卷积神经网络在实验室环境下的面部表情数据集上的准确率已经接近100%。但是真实场景下的准确率不尽如人意,主要是由于真实场景下的面部表情图像中含有一定比例的遮挡和姿势变化,导致人脸外观上的显著变化,对特征学习有一定的干扰作用。因此部分研究学者通过网络、多媒体平台等收集并公开了包含面部表情的真实场景图像数据集[1]。为了解决遮挡和姿态变化对FER任务的影响,一些研究通过加分阶段训练[2]、人脸对齐[3]及多网络特征融合[4]等策略有效提升了真实场景下FER的准确性。还有一些研究利用生成对抗网络将遮挡与非遮挡特征进行联合训练,使整个网络模型对遮挡面部表情图像的识别更具鲁棒性[5]。Wang等[6]根据FERPlus、RAF-DB和AffectNet三个数据集设计了6个包含遮挡和姿势变化人脸图像数据集:Occlusion-FERPlus、Pose-FERPlus、Occlusion-AffectNet、Pose-AffectNet、Occlusion-RAF-DB和Pose-RAF-DB,并且进一步提出了一个区域注意网络(Region attention networks, RAN),能够自适应地学习不同面部区域的重要性系数,对遮挡和姿势变化面部表情识别更具鲁棒性。

本文受Faster RCNN网络[7]中候选区域生成网络(Region proposal network,RPN)方法和网络设计[8]的启发,设计并实现了一种基于双注意力和多区域检测的面部表情识别方法。具体为对于一张面部表情图像,特征提取网络首先提取图像的全局特征,多区域检测模块根据全局特征学习得到包含重要信息的局部区域;然后将局部区域在原图中裁剪出来进一步通过特征提取网络提取得到局部特征;最后将局部特征和全局特征相融合用于面部表情的识别。双注意力则用于提升网络的特征提取能力,捕获更详细的特征信息。

## 1　相关工作

### 1.1　面部表情识别

在现实场景下面部表情很可能会被眼镜、帽子或围巾等佩戴物遮挡,并且会以不同的角度被拍摄,因此现实场景下的面部表情识别任务仍具有很大的挑战性。许多研究学者针对遮挡和姿势变化等问题,提出了一些新方法进一步提升了面部表情识别网络模型的性能。Wang等[9]基于对抗学习的思想提出了对姿势和身份较为鲁棒的人脸表情识别方法。该方法输入相同表情但姿势和身份不一样的两张人脸图像,通过对抗学习的方式去除姿势和身份的变化,仅保留表情的特征信息再进行识别,从而达到对姿势和身份都较为鲁棒的目的。Zhang等[10]针对面部表情识别中存在的人脸姿势问题,提出了一种由深度神经网络驱动的特征学习方法。该方法首先从每个面部图像中提取一组与人脸特征点相对应的尺度不变特征转换(Scale-invariant feature transform, SIFT)特征。然后将包含SIFT特征向量的矩阵用作输入数据,将其发送到设计的深度神经网络(Deep neural network,DNN)模型中,学习用于分类的最佳判别特征。Pan等[11]基于对抗学习的思想,提出了一个有效提升含遮挡的人脸表情识别网络。在训练阶段,网络利用Resnet对遮挡和非遮挡人脸分别进行训练得到$y_1$和$y_2$两组特征,随后使用其设计的5个损失函数对网络进行优化。同时考虑到含遮挡的人脸表情图片较少,作者还通过人工合成的方式构建了含遮挡的人脸表情数据。

### 1.2 注意力机制

注意力机制在使用深度学习方法来模仿人类的视觉机制中发挥了重要作用，能够使网络忽略无关的特征信息而更加关注重要的特征信息。在计算机视觉领域，注意力机制主要分为通道注意力机制和空间注意力机制，分别用于学习图像特征通道和空间上的依赖关系，捕获更加详细的特征信息。Hu等[12]通过建模通道间的相关性提出了SE(Squeeze-and-excitation)模块，该方法能够使网络自适应地学习每个特征通道的重要程度，然后根据重要程度去强化有用的特征并抑制对当前任务用处不大的特征。Cao等[13]简化了非局部操作复杂的网络结构减少了计算量和参数量，提出了一种简化版的非局部操作，并结合了SE模块的思想进一步提出了上下文注意块。该方法首先对特征层进行全局建模，然后学习通道间的依赖关系，最后将特征进行融合进一步提高了网络的性能。SA-Net[14]则针对融合了空间注意力和通道注意力后不可避免地会导致计算量提升的问题，提出了一种采用分组单元组合通道和空间注意力机制的方法。该方法首先将特征层按照通道的维度分组为多个子特征层，然后再并行学习在空间与通道维度上的相关性，最后将所有子特征层聚合在一起并使用“通道置换”来实现不同子特征之间的信息通信。

## 2 本文方法

本节首先介绍本文中网络的整体结构，其中包含一个主干网络(Backbone)和多区域检测模块，主干网络中在特征提取器之后添加双注意力机制以获得更详细的特征信息以及平均池化操作，然后分别对多区域检测部分和双注意力机制部分进行详细说明。

### 2.1 整体网络结构

如图1所示，首先将原始图像$\boldsymbol{R}_0 \in \mathbf{R}^{C\times H\times W}$输入至特征提取(Feature extractor)网络中用于提取原始图像的全局特征，$C$、$H$、$W$分别为通道数、高和宽，然后经过双注意力机制进一步细化全局特征，并通过第一个全连接层$\mathrm{FC}_0$计算图像的原始损失$L_{\mathrm{raw}}$。其次将上述输出的全局特征层(Feature map)输入至多区域检测模块(Multi-region detection)来获得每个局部区域$\{\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_A\}$及对应的得分

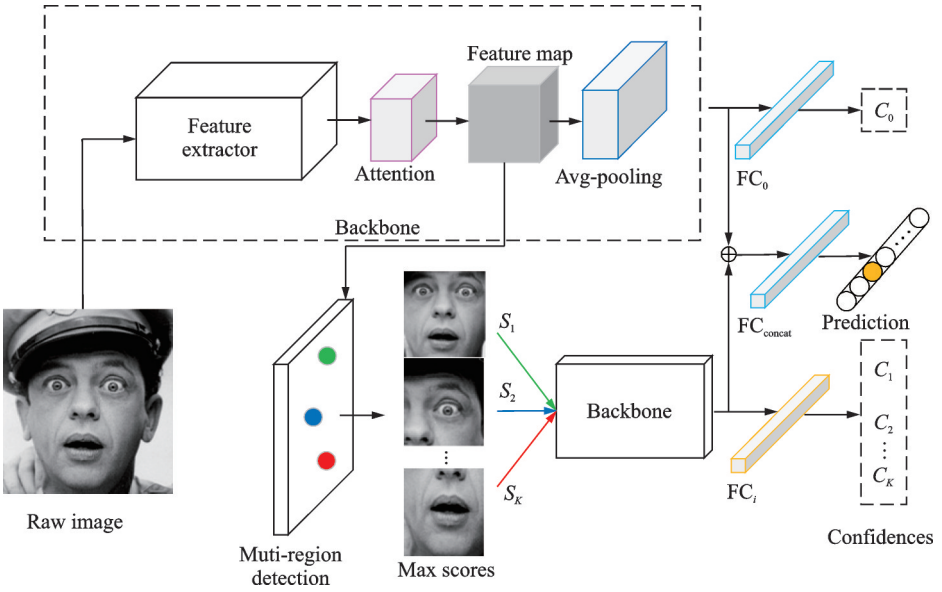


图1 整体网络结构

Fig.1 Overall network structure

$\{S_1, S_2, \cdots, S_A\}$,$A$ 为局部区域的数量,将区域得分排序以获得最高得分的 $K$ 个辨别性最大的局部区域 $\{\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_K\}$,并将这 $K$ 个局部区域在原始图像中裁剪出来上采样至 $224 \times 224$ 大小的尺寸,将其再次输入至主干网络中用于提取局部区域的特征,并使用其对应的全连接层($\mathrm{FC}_i, i \in K$)通过 Cross Entropy Loss 损失函数来计算局部区域损失 $L_i$,同时根据全连接层 $\mathrm{FC}_i$ 的输出使用 Log_softmax 函数来计算每个局部区域相对于类别标签的置信度 $C_i$。最后将图像的全局特征和局部区域特征进行合并作为最终的特征表示,并将结果输入至全连接层 $\mathrm{FC}_{\mathrm{concat}}$,通过 Cross entropy loss 损失函数计算整体的损失 $L_{\mathrm{concat}}$,并得到整个网络的类别分类结果。

## 2.2 多区域检测

在目标检测领域,Faster R-CNN[7]是目前最具代表性的区域检测算法之一。该方法将 RPN 网络和分类网络整合到了一个网络中,首先使用一组基础的卷积、激活和池化层提取图像的特征形成一个全局特征层,用于后续的 RPN 层和全连接层。受 RPN 结构的启发,Yang 等[8]提出一种利用类别标签进行区域检测的方法,在网络学习过程中无需额外的标注框信息,即可完成对网络的训练。

本文受文献[8]中区域检测方法的启发,设计实现了一种用于面部表情识别的多区域检测方法,如图 2 所示。多区域检测的输入为图 1 中主干网络的输出特征层(大小为 $2\,048 \times 14 \times 14$),通过使用具有横向连接自上而下的结构来检测局部区域。具体为分别使用滤波器为 128,卷积核大小为 $3 \times 3$,步长为 1、1、2,填充为 1 的 3 层卷积操作自上而下获得不同尺度的特征层($128 \times 14 \times 14$,$128 \times 7 \times 7$,$128 \times 4 \times 4$),每个卷积后紧接着 ReLU 激活函数。对于每层卷积后的输出,分别使用滤波器为 6、6、9,卷积核大小为 $1 \times 1$,步长为 1,填充为 0 的 3 个卷积操作对特征层进行降维,并提升网络的表达能力。最终获得 3 个尺度分别为 $6 \times 14 \times 14$、$6 \times 7 \times 7$、$9 \times 4 \times 4$ 的特征层,对应 3 个局部区域大小($48 \times 48$,$96 \times 96$,$192 \times 192$),这 3 个特征层上的每个特征值均代表了一个候选区域的得分 $\{S_1, S_2, \cdots, S_A\}$。由于部分候选区域之间有着较大的重叠,为了减少网络的计算量,采用极大值抑制的方法来减少候选区域的数量,通过设置一个阈值(threshold = 0.25)来剔除重复区域大于该阈值的得分较低的候选区域。最后对候选区域的得分进行排序,选出最大的 $K$ 个局部区域 $\{\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_K\}$ 并上采样至 $224 \times 224$ 大小,然后通过特征提取网络进一步学习得到局部特征。

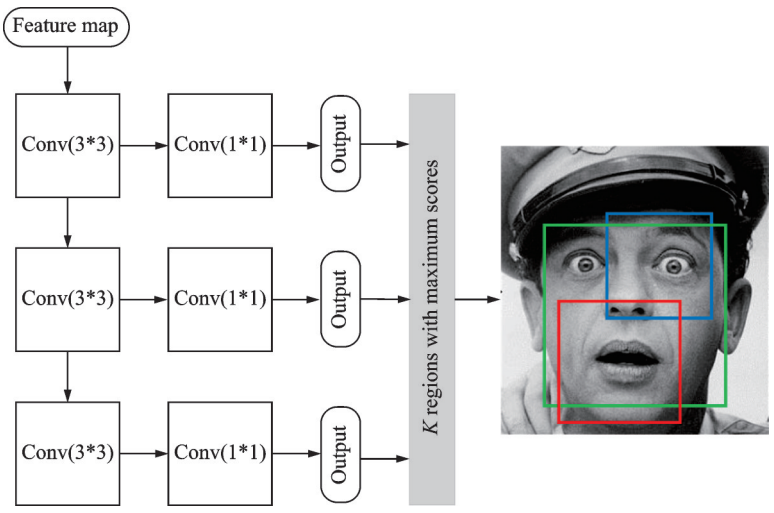


图 2 多区域检测模块

Fig.2 Multi-region detection module

### 2.3 注意力机制

注意力机制已经成为提升卷积神经网络性能的一个重要模块,并且通道注意力机制和空间注意力机制相结合的方法能够更加显著地提升网络的性能,然而这样却不可避免地增加了网络的复杂度和计算量。Yang等[14]采用特征分组和通道置换的方式,提出了一种轻量型双注意力机制方法,本文参考该方法并做了部分改动,注意力模块如图3所示,主要分为特征分组、通道注意力、空间注意力和特征融合4个部分,GAP(Global average pooling)指全局平均池化。

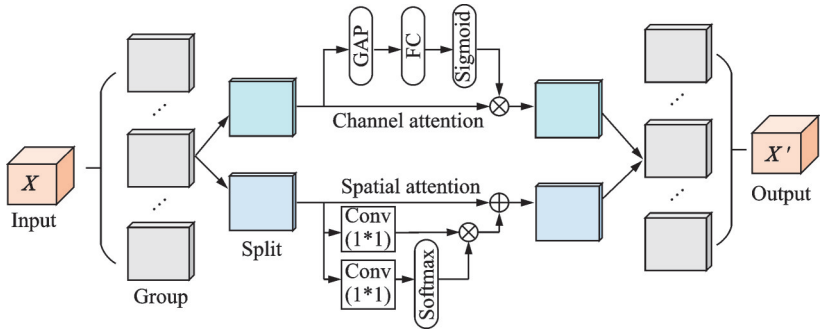


图3 双注意力机制模块

Fig.3 Dual attention mechanism module

#### 2.3.1 特征分组

对于特征提取网络的输出特征层 $\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$,特征分组按照通道的维度 $C$ 将 $X$ 进行拆分为 $G$ 组 $\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_G\}\in\mathbf{R}^{C/G\times H\times W}$。对于每一组子特征层 $\boldsymbol{X}_k,k=1,2,\cdots,G$,都通过两种注意力机制生成不同的权重系数。具体来说,子特征层将被划分为两部分 $\boldsymbol{X}_{k1}、\boldsymbol{X}_{k2}\in\mathbf{R}^{C/2G\times H\times W}$,一部分用于学习通道注意力特征;另一部分用于学习空间注意力特征。

#### 2.3.2 通道注意力

SE模块[15]是一个经典的通道注意力实现方法,然而它会给网络带来过多的参数量,考虑到网络的计算复杂度,为了尽可能地简化注意力模块,本文在实现通道注意力时使用了最简单的GAP、FC和Sigmoid激活函数组合操作。具体为对于输入 $\boldsymbol{X}_{k1}$,首先使用全局平均池化生成通道的整体信息 $\boldsymbol{O}\in\mathbf{R}^{C/2G\times 1\times 1}$,可以通过如下公式计算

$$\boldsymbol{O}=F_{\text{gap}}(\boldsymbol{X}_{k1})=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{X}_{k1}(i,j) \tag{1}$$

然后通过全连接学习一个新的权重系数 $\boldsymbol{W}_1\in\mathbf{R}^{C/2G\times 1\times 1}$ 和偏置 $\boldsymbol{b}_1\in\mathbf{R}^{C/2G\times 1\times 1}$,用来表示各个通道的重要程度,最后再经过Sigmoid函数进行激活,并和输入特征相乘,计算公式为

$$\boldsymbol{X}'_{k1}=\sigma(F_{\text{fc}}(\boldsymbol{O}))=\sigma(\boldsymbol{W}_1\boldsymbol{O}+\boldsymbol{b}_1)\cdot\boldsymbol{X}_{k1} \tag{2}$$

通道注意力实现的过程使用公式可以整体描述为

$$\boldsymbol{X}'_{k1}=\sigma\left(\boldsymbol{W}_1\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{X}_{k1}(i,j)+\boldsymbol{b}_1\right)\cdot\boldsymbol{X}_{k1} \tag{3}$$

#### 2.3.3 空间注意力

不同于通道注意力机制,空间注意更关注各个像素点之间的依赖关系,因此本文利用空间的特征信息来捕捉输入图像中不同局部区域之间的相关性。在实现方面,本文使用Cao等[13]提出的简化的非

局部操作,首先分别使用两个大小为1×1的卷积$\mathrm{Conv}_m$和$\mathrm{Conv}_n$操作,在$\mathrm{Conv}_m$后使用Softmax激活函数,并将两者的结果进行相乘,最后和输入的特征层相加得到空间注意特征,计算公式为

$$x_i'=x_i+\sum_{j=1}^{N}\frac{\exp(W_m x_j)}{\sum_{v=1}^{N}\exp(W_m x_v)}(W_n x_j) \tag{4}$$

式中:$x_i\in \boldsymbol{X}_{k2}$;$N$为特征层$\boldsymbol{X}_{k2}$的像素点个数$W\times H$;$W_m$和$W_n$分别为卷积$\mathrm{Conv}_m$和$\mathrm{Conv}_n$的权重。为了进一步减小计算的复杂度,式(4)可以将$W_n$移到求和公式的外面,改进的公式为

$$x_i'=x_i+W_n\sum_{j=1}^{N}\frac{\exp(W_m x_j)}{\sum_{v=1}^{N}\exp(W_m x_v)}x_j \tag{5}$$

2.3.4 特征融合

所有分组的子特征层$\boldsymbol{X}_k$经过双注意力机制后,重新聚合到一起生成注意力特征层$\boldsymbol{X}'$。最后类似于ShuffleNet v2[16]中的通道置换方法,通过设置一个通道置换操作来确保不同分组间的特征信息能够"相互流转",即将每组的特征进行互相交叉。对于特征层$\boldsymbol{X}'\in\mathbf{R}^{C\times W\times H}$,按照通道数$C$分为$G_{CS}$组,即特征层的维度为$G\times C/G_{CS}\times W\times H$,然后将$G$和$C/G_{CS}$轴转置得到$C/G_{CS}\times G\times W\times H$维度的特征层,最后再将分组进行合并得到原始维度大小($C\times W\times H$)的特征层。

2.3.5 损失函数

对于整个网络中的输入$\{\boldsymbol{R}_0,\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_K\}$,其中,$\boldsymbol{R}_0$为原始图像,$\{\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_K\}$为区域检测并上采样得到的局部区域图像,能够分别得到局部区域的得分$\{\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_K\}$和置信度$\{C_0,C_1,C_2,\cdots,C_K\}$,除了上文中提到的分类损失$L_i$、$L_0$、$L_{\mathrm{concat}}$使用Cross Entropy Loss损失函数来计算外,还需要计算局部区域得分损失$L_S$和置信度损失$L_C$,计算公式为

$$L_S=\sum_{(i,j):C_i<C_j}\mathrm{Relu}\left(\max\left\{S_j-S_i+1,0\right\}\right) \tag{6}$$

$$L_C=-\sum_{i=1}^{K}\ln C_i-\ln C_0 \tag{7}$$

本文最终损失函数如下

$$\mathrm{Loss}=\sum_{(i,j):C_i<C_j}f(S_j-S_i)+\lambda L_S+\mu L_C \tag{8}$$

式中:$f$为交叉熵损失函数;$\lambda$、$\mu$为超参数。

## 3 实验与分析

### 3.1 数据集

本文在3个公开的真实场景面部表情图像数据集上进行了实验来评估所提出的方法,分别为AffectNet[17]、RAF-DB[1]和SFEW[18],这些数据集覆盖了不同尺度的人脸图像以及遮挡和姿势变化等具有挑战性的条件。

AffectNet是迄今为止最大的人脸表情图像数据集。该数据集共包含100多万张图像,使用6种不同语言和1 250个与情绪相关的关键词在3个网络搜索引擎上进行收集得到。其中接近一半的图像是按照7种表情标签进行手工标注的。RAF-DB(Real-world affective faces database)是一个大型的面部表情数据库,包含从互联网上下载的3万多张面部图像。每张图像都由大约40个标记者独立进行表情标注。该数据库中的图像在年龄、性别、种族、头部姿势、光照条件、遮挡(如眼镜、面部毛发、围巾)、后期

处理操作(如各种滤镜和特效)等方面具有很大的差异性。SFEW(Static facial expressions in the wild)通过选择AFEW[19]视频数据中的图像帧来构建,包括了不受约束的面部表情、不同的姿势和遮挡、不同的侧重点、不同的分辨率和现实光照等限制条件。每一张图像都被标注了愤怒、厌恶、恐惧、开心、悲伤、惊讶或平静这7种表情中的一种。

在以上数据集中,AffectNet和SFEW两种数据集只用来测试,RAF-DB用来训练和测试,具体表情类别数量以及训练集和测试集的划分如表1所列。除了上述数据集外,本文还引入了Wang等[6]根据RAF-DB和AffectNet构建的测试数据集Occlusion-RAF-DB、Pose-RAF-DB、Occlusion-AffectNet和Pose-AffectNet用于测试在遮挡和姿势变化面部表情识别上的鲁棒性。这4种数据集的图像数量如表2所示,图4为遮挡和姿势变化面部表情的示例图像。

**表1 数据集类别、训练和测试集划分信息**

**Table 1 Classes and partition information of train and test dataset**

| Dataset | Class | Train | Test |
|---|---|---|---|
| AffectNet | 7 | — | 4 200 |
| SFEW | 7 | — | 808 |
| RAF-DB | 7 | 12 271 | 3 068 |

**表2 遮挡和姿势变化表情数据集**

**Table 2 Occlusion and pose change expression dataset**

| Dataset | Occlusion | | | | Pose | |
|---|---|---|---|---|---|---|
| | 上 | 下 | 左/右 | 眼镜 | >30° | >45° |
| RAF-DB | 126 | 151 | 160 | 298 | 1 248 | 558 |
| AffectNet | 84 | 183 | 128 | 288 | 1 949 | 985 |


图4 遮挡和姿势变化表情示例图像

Fig.4 Example images of occlusion and pose change in facial expression

### 3.2 实验细节

本文使用在MSCeleb-1M[20]人脸表情数据集上预训练的ResNet-50[21]网络模型作为本文方法的基准网络(Baseline),将人脸图像大小调整到448×448,使用动量(Momentum)梯度下降算法(Stochastic gradient descent,SGD)作为优化算法,Batch Normalization作为正则化方法,初始学习率为0.01,并且每20个epoch乘以0.1,权重衰减因子设为1e−4。在与其他方法的对比实验中,参数$K$设置为4,超参数threshold设置为0.25,使用的深度学习框架为Pytorch。

### 3.3 结果对比

表3给出了本文方法和近几年面部表情识别方法在数据集RAF-DB的实验结果对比,结果显示本文方法的识别率达到了87.24%,优于其他方法。这主要是因为区域检测能够有效利用重要的区域信息,抑制遮挡和姿势变化带来的不利信息,并且通过在网络中加入通道和空间注意力机制,使网络能够提取更加详细的面部特征。与RAN方法[6]相比,本文的区域检测能够根据局部特征的重要程度动态地调整所选局部区域的位置,而RAN方法则是以固定位置进行局部区域选取,因此本文方法相比之下可以提高0.34%的准确率。与目前最优的ODAN方法[22]相比,本文方法的准确率提升了0.08%。图5显示了在数

据集RAF-DB上实验结果的混淆矩阵。从图中可以看出,恐惧和厌恶是两种较难识别的表情,恐惧和惊讶容易混淆,12%的恐惧样本图像被识别为惊讶,厌恶的错误识别主要集中在开心和伤心上。

**表3 数据集RAF-DB上的实验结果**

**Table 3 Experimental results on RAF-DB dataset**

| Method | Accuracy/% |
|---|---|
| ResiDen[23] | 76.54 |
| ResNet-PL[11] | 81.97 |
| PG-CNN[24] | 83.27 |
| DLP-CNN[1] | 84.13 |
| ALT[25] | 84.50 |
| gACNN[26] | 85.07 |
| RAN[6] | 86.90 |
| SCN[27] | 87.03 |
| OADN[22] | 87.16 |
| Ours | **87.24** |

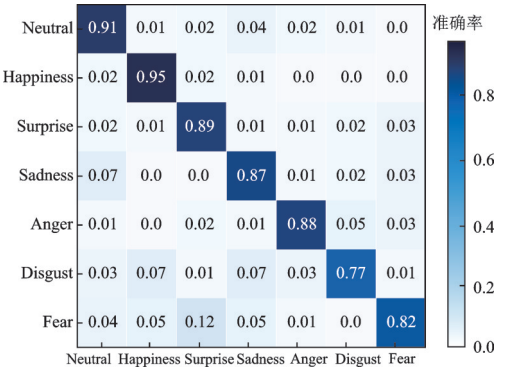


图5 RAF-DB数据集的混淆矩阵

Fig.5 Confusion matrix for RAF-DB dataset

如表4所示,本文方法在AffectNet测试集上获得了61.48%的准确率,相比目前最优的ODAN方法[22],本文方法在准确率上仅降低了0.41%。图6显示了在AffectNet数据集上实验结果的混淆矩阵,开心表情获得了最高的识别准确率为81%,生气、厌恶和恐惧是较难分类的表情,这些表情相比于开心表情,其面部表现比较微妙,并且每个人的表现存在一定的差异性,导致识别较为困难。

**表4 数据集AffectNet上的实验结果**

**Table 4 Experimental results on AffectNet dataset**

| Method | Accuracy/% |
|---|---|
| ResNet-PL[11] | 56.42 |
| PG-CNN[24] | 55.33 |
| DLP-CNN[1] | 54.47 |
| gACNN[26] | 58.78 |
| RAN[6] | 59.50 |
| SCN[27] | 60.23 |
| OADN[22] | **61.89** |
| Ours | 61.48 |

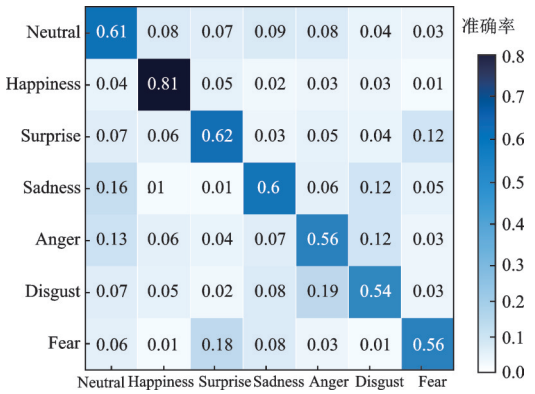


图6 AffectNet数据集的混淆矩阵

Fig.6 Confusion matrix for AffectNet dataset

表5为在数据集SFEW上的实验结果,本文方法获得了57.63%的最高准确率。图7显示了在SFEW数据集上实验结果的混淆矩阵。从图7可以看出,在7种表情中开心是最容易识别的,明显高于其他表情。和前面两个数据集的实验结果对比可以发现,厌恶是最难识别的表情,主要原因在于日常生活中厌恶表情较为少见,每个人对厌恶表情的表现形式不尽相同,并且容易与生气表情混淆。

在Occlusion/Pose-RAF-DB和Occlusion/Pose-AffectNet数据集上的面部表情识别任务中,准确率结果如表6和表7所示,本文方法整体优于RAN方法[6],这主要得益于多区域检测模块的设计。实验结果同样也验证了本文方法在遮挡和姿势变化表情识别任务中的鲁棒性。

**表5 数据集SFEW上的实验结果**

**Table 5 Experimental results on SFEW dataset**

| Method | Accuracy/% |
|---|---|
| Island Loss[28] | 52.52 |
| Identity-aware CNN[29] | 50.98 |
| Multiple deep CNNs[30] | 55.96 |
| DLP-CNN[1] | 54.19 |
| gACNN[26] | 54.47 |
| RAN[6] | 56.40 |
| Ours | **57.63** |

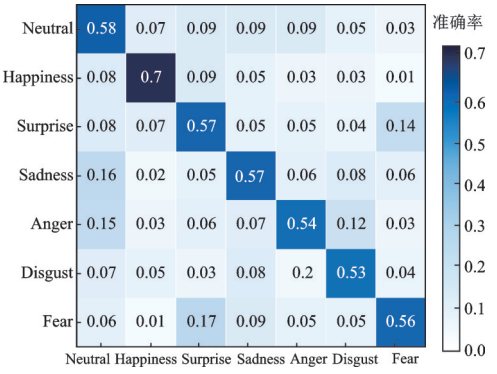


图7 SFEW数据集的混淆矩阵

Fig.7 Confusion matrix for SFEW dataset

**表6 数据集Occlusion/Pose-RAF-DB上的准确率**

**Table 6 Accuracy on Occlusion/Pose-RAF-DB datasets** %

| Method | Occlusion | Pose>30° | Pose>45° |
|---|---|---|---|
| Baseline | 80.19 | 84.04 | 83.15 |
| RAN[6] | 82.78 | **86.74** | 85.20 |
| Ours | **85.16** | 84.79 | **85.45** |

**表7 数据集Occlusion/Pose-AffectNet上的准确率**

**Table 7 Accuracy on Occlusion/Pose-AffectNet datasets** %

| Method | Occlusion | Pose>30° | Pose>45° |
|---|---|---|---|
| Baseline | 49.48 | 50.01 | 48.50 |
| RAN[6] | 58.5 | 53.9 | 53.19 |
| Ours | **61.04** | **56.13** | **55.87** |

### 3.4 消融实验

本文首先针对网络中两个重要的模块(双注意力模块和多区域检测模块)进行消融实验,在ResNet-50基础网络中加入双注意力模块构成网络模型ResNet-50-DA(Dual Attention),在ResNet-50基础网络中加入多区域检测模块构成网络模型ResNet-50-MRD(Multi-region Detection)。除此之外在基础网络中加入特征金字塔网络(Feature Pyramid network,FPN)用作特征提取器构成ResNet-50-FPN网络模型进行实验比较。FPN通过特征金字塔的方式来处理小物体的特征提取以及多尺度变化问题取得不错效果。

表8显示了上述网络模型的实验结果。从表中可以看出,在ResNet-50的基础上,双注意力模块在3种数据集(RAF-DB、AffectNet、SFEW)上的准确率分别提升了1.67%、2.01%、1.68%,多区域检测模块在3种数据集上的准确率分别提升了4.78%、4.19%、3.13%,这验证了本文的双注意力模块和多区域检测模块均能够提升网络的性能。而两者的结合(本文方法)使网络在3种数据集上的准确率分别提升了5.79%、5.14%、4.35%,获得了最高的识别准确率。FPN模型在特征提取方面较双注意力模块有较大的提升,在3个数据集上的准确率分别提升了2.81%、2.03%、1.26%,与多区域检测模块的性能接近,但并不能达到本文两者结合所获得的准确率,因为FPN在小物体检测上效果较好,但本文的遮挡和姿

**表8 双注意力和多区域检测模块的实验结果对比**

**Table 8 Comparison of experimental results of dual attention and multi-region detection modules**

| Method | Accuracy/% | | |
|---|---|---|---|
| | RAF-DB | AffectNet | SFEW |
| ResNet-50[21] | 81.45 | 56.34 | 53.28 |
| ResNet-50-DA | 83.12 | 58.35 | 54.96 |
| ResNet-50-MRD | 86.23 | 60.53 | 56.41 |
| ResNet-50-FPN | 85.93 | 60.38 | 56.22 |
| Ours | **87.24** | **61.48** | **57.63** |

势变化对识别结果影响较大。

另外,本文还单独在RAF-DB数据集上做了一些扩展实验来探究网络模型中局部区域数量$K$、超参数Threshold、损失函数中超参数$\lambda$和$\mu$以及注意力模块对实验结果的影响。由于本文中局部区域的检测对实验结果的提升较大,因此本文通过改变参数$K$的取值来探究最优的局部区域数量。如图8所示,在固定其他参数不变的情况下,当$K$取值为4时,本文方法在表情识别准确率上达到最高。而当选择更多的局部区域时,准确率反而有所下降,这说明过多的局部区域会对网络提取重要的特征信息产生干扰,从而影响最终的分类结果。然后通过固定$K=4$,以0.05为间隔差,在0.1~0.5取多个Threshold的值进行模型训练,验证超参数Threshold对实验结果的影响。如图9所示,超参数Threshold在取0.25时实验结果达到最优。Threshold在0.1~0.25的变化中准确率呈增长趋势,这是因为过小的Threshold会导致所选局部区域覆盖较为分散。而随着Threshold的继续增长,就会导致所选局部区域有较大的重叠。

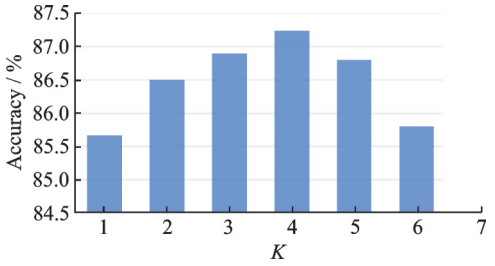


图8 参数$K$对实验结果的影响

Fig.8 Effect of parameter $K$ on experimental results

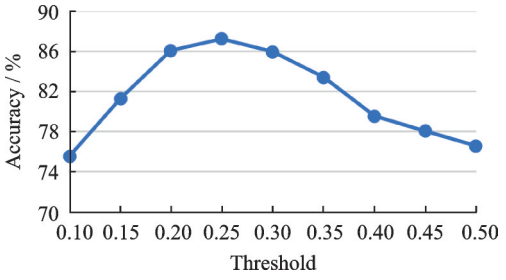


图9 超参数threshold对实验结果的影响

Fig.9 Effect of hyper-parameter threshold on experimental results

本文的整体损失函数包括3个部分:第1部分损失函数如果置信度$C_i<C_j$,那么区域$j$的得分信息$S_j>S_i$;第2部分$L_S$将各个区域和整体图像的特征进行拼接计算其损失;第3部分$L_C$表示所有区域的以及整体图像的交叉熵损失之和。通过超参数$\lambda$和$\mu$控制$L_S$和$L_C$在整体损失函数中的权重,表9展示了在不同权重情况下所使用的整体损失函数对于识别精度的影响,在权重为1时取得最好的效果。因为在图像区域划分计算置信度以及最终合并局部和整体特征时,各部分损失函数的权重不应相差过大,否则整体的识别性能会偏向于区域划分或者特征合并的结果,对框架的整体识别精度产生影响,因此本文使用$\lambda=\mu=1$。

最后,本文探究了注意力模块中通道注意力、空间注意力和通道置换3个组件对实验结果的影响,其中参数$K$的值固定为4,Threshold的值固定为0.25,其他网络设置相同,实验结果如表10所示。从结果可以看出,两种注意力机制的单独应用均能够提升网络的性能,并且两种注意力机制结合应用能够达到更好的效果。而在使用通道置换的情况下,准确率提升了0.22%,这证明了不同分组之间特征信息的交互能够增强整体特征的表示。

**表9 权重$\lambda$和$\mu$对实验结果的影响**

**Table 9 Effect of hyperparameters $\lambda$ and $\mu$ on experimental results**

| $\lambda$ | $\mu$ | Accuracy/% |
|---|---|---|
| 0.6 | 1.0 | 83.56 |
| 0.8 | 1.0 | 85.37 |
| **1.0** | **1.0** | **87.24** |
| 1.0 | 1.2 | 86.12 |
| 1.0 | 1.4 | 84.89 |

**表10 注意力模块对实验结果的影响**

**Table 10 Effect of attention module on experimental results**

| Spatial | Channel | Channel shuffle | Accuracy/% |
|---|---|---|---|
| | | | 86.23 |
| √ | | | 86.67 |
| | √ | | 86.75 |
| √ | √ | | 87.02 |
| √ | √ | √ | 87.24 |

## 4 结束语

本文提出了一种基于双注意力和多区域检测的遮挡和姿势变化面部表情识别。该方法使用通道注意力和空间注意力提升网络的特征提取能力,使用多区域检测捕获有利于面部表情识别的局部区域,然后将局部区域送至网络进一步学习得到局部特征信息,最后将局部特征信息和全局特征信息相融合,用于面部表情识别。在Occlusion-RAF-DB、Pose-RAF-DB、Occlusion-AffectNet和Pose-AffectNet数据集上进行实验,证明了本文方法在遮挡和姿势变化条件下具有较好的鲁棒性。与现有方法相比,本文方法获得了更高的识别准确率,但本文方法整体网络结构较为复杂,对计算性能有较高的要求,这也是未来工作进一步改进的方向。

**参考文献:**


[1] LI S, DENG W. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition [J]. IEEE Transactions on Image Processing, 2018, 28(1): 356-370.

[2] NG H W, NGUYEN V D, VONIKAKIS V, et al. Deep learning for emotion recognition on small datasets using transfer learning[C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. [S.l.]: ACM, 2015: 443-449.

[3] KIM B K, DONG S Y, ROH J, et al. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S.l.]: IEEE, 2016: 48-57.

[4] KIM B K, LEE H, ROH J, et al. Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition[C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. [S.l.]: ACM, 2015: 427-434.

[5] YAN Y, HUANG Y, CHEN S, et al. Joint deep learning of facial expression synthesis and recognition[J]. IEEE Transactions on Multimedia, 2019, 22 (11): 2792-2807.

[6] WANG K, PENG X, YANG J, et al. Region attention networks for pose and occlusion robust facial expression recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 4057-4069.

[7] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[8] YANG Z, LUO T, WANG D, et al. Learning to navigate for fine-grained classification[C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 420-435.

[9] WANG C, WANG S, LIANG G. Identity-and pose-robust facial expression recognition through adversarial feature learning [C]//Proceedings of the 27th ACM International Conference on Multimedia. [S.l.]: ACM, 2019: 238-246.

[10] ZHANG T, ZHENG W, CUI Z, et al. A deep neural network-driven feature learning method for multi-view facial expression recognition[J]. IEEE Transactions on Multimedia, 2016, 18(12): 2528-2536.

[11] PAN B, WANG S, XIA B. Occluded facial expression recognition enhanced through privileged information[C]//Proceedings of the 27th ACM International Conference on Multimedia. [S.l.]: ACM, 2019: 566-573.

[12] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 7132-7141.

[13] CAO Y, XU J, LIN S, et al. GCNet: Non-local networks meet squeeze-excitation networks and beyond[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. [S.l.]: IEEE, 2019. DOI:10.1109/iccvw.2019.00246.

[14] YANG Y B. SA-Net: Shuffle attention for deep convolutional neural networks[C]//Proceedings of the IEEE International Conference on Acoustics. [S.l.]: IEEE, 2021. DOI:10.1109/icassp39728.2021.

[15] KOSSAIFI J, TZIMIROPOULOS G, TODOROVIC S, et al. AFEW-VA database for valence and arousal estimation in-the-wild[J]. Image and Vision Computing, 2017, 65: 23-36.

[16] MA N, ZHANG X, ZHENG H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//

Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 116-131.

[17] MOLLAHOSSEINI A, HASANI B, MAHOOR M H. AffectNet: A database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing, 2017, 10(1): 18-31.

[18] DHALL A, GOECKE R, LUCEY S, et al. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. [S.l.]: IEEE, 2011. DOI: 10.1109/iccvw.2011.6130508.

[19] DHALL A, GOECKE R, LUCEY S, et al. Collecting large, richly annotated facial-expression databases from movies[J]. IEEE Multimedia, 2012, 19(3): 34-41.

[20] GUO Y, ZHANG L, HU Y, et al. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2016: 87-102.

[21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 770-778.

[22] DING H, ZHOU P, CHELLAPPA R. Occlusion-adaptive deep network for robust facial expression recognition[C]//Proceedings of the 2020 IEEE International Joint Conference on Biometrics. [S.l.]: IEEE, 2020: 1-9.

[23] JYOTI S, SHARMA G, DHALL A. Expression empowered residen network for facial action unit detection[C]//Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). [S.l.]: IEEE, 2019: 1-8.

[24] LI Y, ZENG J, SHAN S, et al. Patch-gated CNN for occlusion-aware facial expression recognition[C]//Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR). [S.l.]: IEEE, 2018: 2209-2214.

[25] FLOREA C, FLOREA L, BADEA M S, et al. Annealed label transfer for face expression recognition[C]//Proceedings of the BMVC. Wales, UK: [s.n.], 2019: 104.

[26] LI Y, ZENG J, SHAN S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2439-2450.

[27] WANG K, PENG X, YANG J, et al. Suppressing uncertainties for large-scale facial expression recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 6897-6906.

[28] CAI J, MENG Z, KHAN A S, et al. Island loss for learning discriminative features in facial expression recognition[C]//Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition. Xi'an, China: IEEE, 2018: 302-309.

[29] MENG Z, LIU P, CAI J, et al. Identity-aware convolutional neural network for facial expression recognition[C]//Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition. [S.l.]: IEEE, 2017: 558-565.

[30] YU Z, ZHANG C. Image based static facial expression recognition with multiple deep network learning[C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. [S.l.]: ACM, 2015: 435-442.

**作者简介:**

**潘新辰**(1996-),男,硕士研究生,研究方向:工业信息化。

**秦岭**(1980-),**通信作者**,男,副研究员,研究方向:工业信息化,E-mail:1573749331@qq.com。

**杨小健**(1963-),男,教授,研究方向:工业信息化、计算机控制。

(编辑:张黄群)