基于多特征融合的人脸表情识别算法

吕 鹏,单剑锋
(南京邮电大学 电子信息与光学工程、微电子学院,江苏 南京 210023)

摘 要:由于稠密网络(DenseNet)模型具有独特的特征提取和传输方式,使其面对小数据集时在缓解网络过拟合的同时,可以取得不错的分类效果。但是传统的 DenseNet 模型具有较深的网络结构,可能造成特征冗余和硬件内存的负担。针对该问题,研究了一种相对浅层的稠密网络,通过压缩稠密网络的深度并增加每个模块中卷积核的数量来高效提取表情图像的隐性特征。考虑到该稠密网络在提取特征时也舍弃了部分图像信息以及单一特征可能难以表达人脸表情图像的全部信息,利用 LDN(Local Directional Number Pattern,LDN)算法提取表情图像的梯度方向纹理信息,与稠密网络提取的隐式特征进行特征融合,共同进入 Softmax 层进行表情分类。该算法在 CK+和 Jaffe 数据集上进行仿真实验,获得了不错的识别率,在一定程度上证实了算法的有效性。

关键词:人脸表情识别;深度学习;稠密网络;浅层网络;特征融合
中图分类号:TP391.41 **文献标识码**:A **文章编号**:1673-629X(2022)10-0151-05
doi:10.3969/j.issn.1673-629X.2022.10.025

Facial Expression Recognition Algorithm Based on Multi-feature Fusion

LYU Peng,SHAN Jian-feng
(School of Electronic Information and Optical Engineering,Microelectronics,Nanjing University of Posts and Telecommunications,Nanjing 210023,China)

Abstract:Because DenseNet model has a unique feature extraction and transmission mode,it can not only alleviate network overfitting but also achieve excellent classification effect when facing small data sets. However,the traditional DenseNet model has a deep network structure,which may cause feature redundancy and hardware memory burden. To solve this problem,we study a relatively shallow dense network and extract the recessive features efficiently by compressing the depth of the dense network and increasing the number of convolution kernels in each module. The local directional number pattern (LDN) algorithm is used to extract gradient direction texture information of facial expression images,considering that the dense network abandoned part of image information and a single feature could not express all information of facial expression images. Feature fusion is carried out with implicit features extracted from dense network,and they are jointly entered into Softmax layer for facial expression classification. The algorithm is simulated on CK+ and Jaffe datasets,and high recognition rate is obtained,which proves the effectiveness of the algorithm to a certain extent.

Key words:facial expression recognition;deep learning;DenseNet;shallow network;feature fusion

0 引 言

面部表情在人类的社会交往中起着至关重要的作用。1971 年,心理学家 Ekman 与 Friesen 提出六种人类主要的基本表情,每种表情代表一种独特的心理活动,分别为愤怒、厌恶、恐惧、高兴、悲伤、惊讶。随着人工智能(AI)技术的不断发展,人机交互、智能控制等技术的研究变得越来越流行,面部表情识别[1](FER)是其中一个重要的视觉信息,如果机器能借此预测人类的情绪,就可以做出相应的行为来满足人类的需求。典型的表情识别系统包括人脸图像采集、特征提取、训练和识别。大多数人脸图像特征对噪声和光照变化非常敏感。因此,能够容忍噪声和光照变化的特征有助于生成鲁棒的表情识别系统。

传统的特征提取方法大致分为两种:基于几何特征的特征提取方法,即通过对嘴巴、眼睛、眉毛等具有显著特征的位置进行定位,测量确定其大小、距离、形状等特征,再进行分类。基于 AUs 的表情识别算法即通过检测一些预定义的 AUs,然后根据 FACS 将它们

收稿日期:2021-09-13 **修回日期**:2022-01-14
基金项目:江苏省教育科学“十三五”规划 2020 年度课题(B-a/2020/01/01)
作者简介:吕 鹏(1995-),男,硕士研究生,研究方向为图像处理、深度学习;单剑锋,博士,副教授,研究方向为无线通信系统中的信号处理技术、智能信息处理、目标识别等。

的组合编码为特定的表情。由于 AU 的定义在语义上有歧义,在实际应用中很难实现对 AU 的准确检测。基于外观的传统特征提取方法,从面部整体或局部的图像过滤器来提取面部外观变化,即通过对图像整体或局部的纹理进行检测并提取,再进行分类。诸如,Gabor[2]小波、Haar 小波、LBP[3]、LDP[4]等方法,在一定程度上提高了表情识别的正确率,但是对光照和噪声的鲁棒性并不高,在人脸图像的平滑区域提取基于边缘的局部特征会产生对噪声敏感的不稳定模式,对分类结果产生负面影响。LDN[5]算法通过提取表情图像的梯度方向信息,得到可以区分具有相同微结构但强度不同的编码,该算法对光照和噪声有不错的鲁棒性。

随着深度学习的快速发展,人脸表情识别进入新的阶段。G. Wang 等[6]通过改进 LeNet-5 网络,增加了卷积层和池化层,将低级特征与高级特征相结合,用可训练卷积核学习其隐性特征,一定程度上提高了识别率;H. Yang 等[7]通过 cGAN 网络生成原始人脸表情对应的中性表情,利用过滤在中间层的残余表情成分进行分类。该算法在一定程度上减少了在表情分类的过程中同一个人的不同表情被误分为同一类的概率;J. Li 等[8]提出一种 Faster R-CNN 算法,用卷积网络提取表情隐性特征之后,利用 RPNs 生成高质量的区域建议并通过 R-CNN 进行检测,最后进行分类,一定程度上避免传统面部表情识别中复杂的显性特征提取过程和低层次数据操作的问题。以上大部分方法是提取单一特征或者对不同层次的单一特征进行识别分类,难免会对表情的部分细节特征有所遗漏,导致难以详细地描述表情图像信息。该文在研究浅层 DenseNet 模型的基础上,将 LDN 特征与稠密特征进行融合,利用特征的多样性提高表情识别率。

1 相关理论

1.1 DenseNet

K. He 等[9]提出 ResNet 网络,通过残差块增加网络深度的同时也在一定程度上缓解梯度消失的问题。G. Huang 等[10]提出的 DenseNet 网络借鉴了 ResNet 网络的思想,设计了 Dense Block 模块,通过对比式(1)ResNet 中残差模块与式(2)DenseNet 中 Dense Block 模块得出两者的本质区别,其目的在于将输入与该层之前的网络层输出叠加作为该层的输入,增加各卷积层特征的利用率。

$$x_\ell = H_\ell(x_{\ell-1}) + x_{\ell-1} \tag{1}$$

$$x_\ell = H([x_0, x_1, \cdots, x_{\ell-1}]) \tag{2}$$

其中,x_ℓ 为第 ℓ 层输出,H 表示非线性变换,$[x_0, x_1, \cdots, x_{\ell-1}]$ 表示从 0 到 $\ell-1$ 层的特征作合并操作。

每个 Dense Block 中都有多个 BN-ReLU-Conv(3×3)层,同一个 Dense Block 中在不改变的输入尺寸的情况下,尽可能提取图像的隐性特征。假设每一个非线性变换 H 的输出为 K 个特征,那么第 ℓ 层网络的输入便为 $K_\ell + (\ell-1) \times K$ 个输入特征,其中 K 称为增长率。为了防止特征图过多导致过拟合和计算复杂度的问题,在 Conv(3×3)之前增加 1×1 的卷积层进行降维,即 BN-ReLU-Conv(1×1)-BN-ReLU-Conv(3×3),在每两个 Dense Block 之间增加过渡层,即 BN-ReLU-Conv(1×1)+2X2AvgPooling 在减小图像尺寸的同时再次降维,最终输入到 Softmax 层进行分类。

BN 层为批量标准化层,将每个批量数据标准化,用来加快模型学习和收敛速度,防止梯度爆炸、消失和过拟合的问题。ReLU 为激活函数,用以增加非线性因素,提高模型的表达能力。此外,面对全连接层有增加 Training 以及 testing 的计算量而降低了训练速度和参数量过多导致过拟合的缺点,DenseNet 中利用全局平均池化层(GAP)来代替全连接层。GAP 使得特征图与最终的分类间转换更加简单自然,不像全连接层需要大量训练调优的参数,降低了空间参数会使模型更加健壮,抗过拟合效果更佳。

1.2 LDN 算法

A. Ramirez Rivera 等提出 LDN(Local Directional Number)算法,以一种紧凑的方式对人脸纹理的方向信息(即纹理的结构)进行编码,通过与方向算子进行卷积,产生比现有方法更有分辨力的编码。如图 1 所示,借助提取方向信息的 kirsch 算子来计算每个微图像的结构,并使用突出的方向索引(方向号)和符号来编码这些信息。这能够区分具有不同强度转换的相似结构模式。

$$M_0 = \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \quad M_1 = \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \quad M_2 = \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix} \quad M_3 = \begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix}$$

$$M_4 = \begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix} \quad M_5 = \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix} \quad M_6 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} \quad M_7 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}$$

图 1 kirsch 算子

将人脸表情图像与 M_0 到 M_7 各个算子卷积得到不同方向的边界响应,找出特定方向上具有高值的边界响应,生成编码。这里以高值的边界响应作为主要区域进行编码。正响应最高和负响应最低分别为编码方案中的 MSB 和 LSB,其表达式为:

$$\mathrm{LDN}(x,y) = 8 \times i_{(x,y)} + j_{(x,y)} \tag{3}$$

其中,(x,y) 为编码局域的中间像素,$i_{(x,y)}$ 为最大正

响应(MSB)的方向数,$j_{(x,y)}$ 为最小负响应(LSB)的方向数,其表达式为:

$$i_{(x,y)} = \text{argmax}i\{\text{Ⅱ}^i(x,y)\,|0 \leqslant i \leqslant 7\} \tag{4}$$

$$j_{(x,y)} = \text{argmin}i\{\text{Ⅱ}^j(x,y)\,|0 \leqslant j \leqslant 7\} \tag{5}$$

其中,Ⅱ^i 为原始图像的卷积运算:

$$\text{Ⅱ}^i = \text{Ⅰ} * \text{M}_i \tag{6}$$

其中,M_i 为 kirsch 第 i 个掩模,得到整体的 LDN 编码和直方图 v_i,但是由于该编码方式缺乏位置信息,需要将图像分成多个 cell,通过顺序级联每个 cell 的 LDN 编码和直方图,可以得到人脸表情图像的全局纹理信息,如公式(7)所示。其中 s^i 为第 i 个 cell。

$$H^i(c) = \sum_{\substack{\text{LDN}(x,y)=c \\ (x,y)\in s^i}} v, \forall c \tag{7}$$

LDN 算法对噪声和光照有较好的鲁棒性,可以有效地描述人脸表情的方向纹理信息。

2　文中模型

针对单一特征可能会丢失部分表情图像信息的问题,设计一种多特征融合的并行网络,如图 2 所示。上侧通道通过参考 DenseNet 结构,设计了三个 Dense Block 级联的浅层网络结构,用以提取人脸表情特征,下侧通道通过 LDN 算法提取表情图像的方向信息,将两种方法提取的特征进行融合,最后一起送入 SoftMax 进行表情分类。

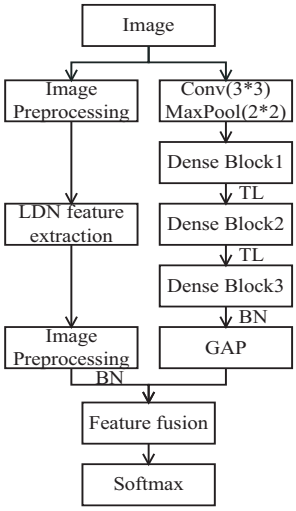

图 2　文中网络模型

左侧通道通过预处理表情图像与 kirsch 算子卷积得到每张图像的各个方向信息,利用公式计算最高正响应和最低负响应,得到每张图像的局部方向数字编码,再将编码输入到全连接层,经过 BN 层后与右通道提取的稠密特征级联。

右侧通道是改进的 Dense Block 结构(如图 3 所示),特征图通过瓶颈层(bottleneck layer)H_i(i = 1,2,3)后在通道上进行叠加。图 4 为瓶颈层结构,每个 Dense Block 结构中包含 3 个 bottleneck layer 结构级联,每个瓶颈层中包含 64 个 1 * 1 卷积核和 32 个 3 * 3 卷积核,右侧通道网络总层数为 20 层,在利用稠密网路结构特点的同时大大缩短稠密网络的层数,缩短每个 Dense Block 深度的同时增加每层卷积核来增加特征图的数量尽可能提取图像的隐性特征。为防止特征图在之后的卷积层中叠加过多造成特征冗余,在每两个 Dense Block 之间增加过渡层(Transition Layer,TL 结构),压缩系数为 0.5,用来缩小图像尺寸和防止特征冗余。

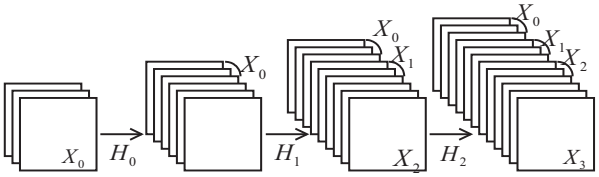

图 3　Dense Block i (i = 1,2,3) 结构

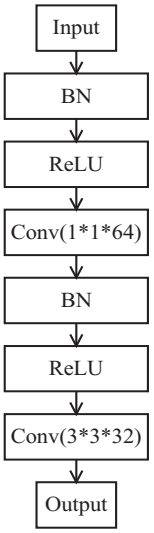

图 4　bottleneck layer

3　仿真实验

3.1　数据集预处理

CK+[11] 数据集由 123 名参与者的 593 个表情图像序列组成,包含 7 种基本表情:生气、蔑视、厌恶、恐惧、高兴、悲伤、惊讶。从每个序列中选取 3 幅最具表情代表性的人脸图像。通过人脸检测选取人脸部分,并将图像裁剪为 48 * 48 大小的人脸表情图像,如图 5 所示。

图 5　CK+数据集

Jaffe 数据集是由 10 位日本女性在实验室条件下根据指示做出的 7 种表情,包括生气、厌恶、恐惧、高

兴、悲伤、惊讶、中性,共 213 张图像,如图 6 所示。通过对每张图像镜像反转并在左上角、右上角、右下角、左下角、中心方位裁剪为 42 * 42 的图像,将数据集扩大 10 倍,再送入模型进行训练。

生气 厌恶 恐惧 高兴 中性 悲伤 惊讶

图 6 Jaffe 数据集

3.2 实验结果分析

将数据集分成 5 份,令其中 4 份作为训练集,1 份作为测试集,并将 5 次测试结果取平均作为最终的识别率。为了增加模型的鲁棒性,对训练集部分作数据增强:随机水平翻转、随机角度旋转、随机水平或者垂直方向平移、随机缩放等操作。模型采用 Adam 优化器,进行 150 轮训练,初始的学习率为 0.001,经过 100 轮训练之后,学习率衰减 10 倍继续进行训练。CK+数据集的批数量为 64,由于 Jaffe 数据集的样本数量相对较少,批数量为 16。

表 1 为 CK+数据集测试集的混淆矩阵。从表 1 中可以得出高兴、厌恶、恐惧和惊讶的表情识别率最高,蔑视的表情识别率相对较低,模型将某些蔑视的表情误分类成恐惧,通过观察 CK+数据集中的蔑视类和恐惧类表情,原因可能是试验人员在发出蔑视和恐惧表情时带有相似特征,如嘴巴紧闭等,增加了两类表情的混淆度。

表 1 CK+数据集混淆矩阵

	生气	蔑视	厌恶	恐惧	高兴	悲伤	惊讶
生气	0.93	0.00	0.00	0.02	0.00	0.04	0.00
蔑视	0.00	0.83	0.00	0.17	0.00	0.00	0.00
厌恶	0.00	0.00	1.00	0.00	0.00	0.00	0.00
恐惧	0.00	0.00	0.00	1.00	0.00	0.00	0.00
高兴	0.00	0.00	0.00	0.00	1.00	0.00	0.00
悲伤	0.04	0.00	0.00	0.00	0.00	0.96	0.00
惊讶	0.00	0.00	0.00	0.00	0.00	0.00	1.00

表 2 CK+数据集算法对比

算法	CK+识别率/%
FRR—CNN[12]	92.06
LDTP[13]	93.58
IACNN[14]	95.37
改进 DenseNet	94.57
DenseNet+LDN	96.00

表 2 为针对 CK+数据集的不同算法之间识别率的对比,文献[12]提出了特征冗余缩减卷积神经网络(FRRCNN),通过在同一层的特征映射之间呈现更具辨别力的图像特征来减少冗余,文献[13]通过利用方向信息和三元模式有效地编码情绪相关特征,文献[14]提出了一种表情敏感对比损失方法来度量表情相似度并且提出了一种身份敏感的对比损失算法,用于从身份标签中学习身份相关信息,实现身份不变表达式识别,都取得了不错的识别率。从表 2 中可以看出,利用稠密网络改进的卷积神经网络取得了 94.57% 的识别率,但是由于数据量小,单一特征容易丢失表情图像信息的问题,表情识别率并不算高。提出的稠密特征与 LDN 特征融合算法在一定程度上弥补了单一特征的不足,提高了表情的识别率,证明该算法有一定的有效性。

表 3 Jaffe 数据集混淆矩阵

	生气	厌恶	恐惧	高兴	悲伤	惊讶	中性
生气	0.91	0.04	0.02	0.00	0.04	0.00	0.00
厌恶	0.00	0.92	0.08	0.00	0.00	0.00	0.00
恐惧	0.00	0.00	0.96	0.00	0.00	0.04	0.00
高兴	0.00	0.00	0.00	0.97	0.00	0.03	0.00
悲伤	0.00	0.00	0.04	0.00	0.96	0.00	0.00
惊讶	0.00	0.00	0.02	0.00	0.00	0.98	0.00
中性	0.00	0.00	0.00	0.00	0.00	0.02	0.98

表 3 为 Jaffe 数据集的混淆矩阵,从表中得出惊讶和中性的表情识别率相对较高,达到 98%。生气和厌恶表情的识别率相对较低,分别被误分为悲伤和恐惧。对比欧美人组成的 CK+数据集,亚洲女性组成的 Jaffe

数据集中面部表情幅度较小,而且生气会伴随悲伤,恐惧伴随厌恶,具有相似的特征,因此增加了表情的混淆度。

表 4　Jaffe 数据集算法对比

算法	Jaffe 识别率/%
Gabor+NN[15]	89.00
PDM+SVM[16]	93.16
ARLCP[17]	94.41
改进 DenseNet	91.07
DenseNet+LDN	95.43

表 4 为 Jaffe 数据集的不同算法之间识别率对比。文献[15]采用 Cabor 小波变换提取图像特征再利用神经网络进行分类,用单一方法很难提取表情图像的全部特征,并且该数据集相对较小,使得提取的图像特征更少,因此该方法的识别率并不高。文献[16]通过训练一个点分布模型(PDM)来提取人脸特定区域的几何特征,再利用 SVM 进行分类。但是该方法依赖特定区域的定义,而且部分表情的变化对特定区域的改变很难察觉。文献[17]的 ARLCP 算法是利用边缘响应的符号、幅度和方向信息提取表情图像相关特征,但单一方法难以提取表情的全部信息,因此改进的 DenseNet 方法取得的识别率不高,通过结合 LDN 算法提取的特征,经过 5 折交叉验证,并将 5 次测试结果取平均值后取得 95.43% 的识别率。

表 5　参数量对比

数据集	改进 DenseNet	DenseNet+LDN	VGG
CK+	268 695	694 839	1.8 m
JAFFE	268 695	595 479	1.8 m

表 5 为模型参数量的对比,其中 VGG 模型深度为 10 层,参数量为 180 万左右。DenseNet 具有独特的特征提取和传递方式,对比相同深度的卷积神经网络,其模型总体的参数复杂度会低很多。而改进的 DenseNet 缩短了原模型的深度,进一步减少了参数复杂度。在其基础上结合 LDN 特征提取算法,最终提高了模型的人脸表情识别率。虽然在一定程度上增加了参数复杂度,但相比 VGG 模型的参数量,总体增加的参数复杂度并不大。

4　结束语

针对在小数据集上单一特征可能丢失表情图像部分信息,该文提出一种多特征融合的并行网络,将 DenseNet 网络提取的稠密特征与 LDN 算法提取的特征进行融合,再用 Softmax 进行分类。对比改进 DenseNet 网络在 CK+数据集和 Jaffe 数据集上的识别率,有一定的提升,证明了该并行网络的有效性。且该方法采用稠密连接的特点和缩短网络深度后使得网络的参数量较少,但该网络在有效提高表情识别率的同时也增加了算法复杂度,下一步工作可以继续优化网络模型和参数。

参考文献:

[1] 党宏社,王　淼,张选德. 基于深度学习的面部表情识别方法综述[J]. 科学技术与工程,2020,20(24):9724-9732.

[2] 梁华刚,张志伟,王亚茹. 自适应 Gabor 卷积核编码网络的表情识别方法[J]. 计算机工程与应用,2020,56(10):149-156.

[3] 姜　万,周晓彦,徐华南,等. 基于 LBP 与双时空神经网络的微表情识别[J]. 信息与控制,2020,49(6):673-679.

[4] 叶　杨,孙会龙,刘　贞. 一种自适应加权 LDP 的虚拟现实设备表情识别方法[J]. 重庆理工大学学报:自然科学版,2019,33(10):109-114.

[5] RIVERA A R,CASTILLO J R,CHAE J R. Local directional number pattern for face analysis:face and expression recognition[J]. IEEE Transactions on Image Processing,2013,22(5):1740-1752.

[6] WANG G,GONG J. Facial expression recognition based on improved LeNet-5 CNN[C]//2019 Chinese control and decision conference (CCDC). Nanchang:IEEE,2019:5655-5660.

[7] YANG H Y,CIFTCI U,YIN L J. Facial expression recognition by de-expression residue learning[C]//2018 IEEE/CVF conference on computer vision and pattern recognition. Salt Lake City:IEEE,2018:2168-2177.

[8] LI J X,ZHANG D X,ZHANG J J,et al. Facial expression recognition with faster R-CNN[J]. Procedia Computer Science,2017,107:135-140.

[9] HE K M,ZHANG X Y,REN S Q,et al. Deep residual learning for image recognition[C]//2016 IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas:IEEE,2016:770-778.

[10] HUANG G,LIU Z,MAATEN L V D,et al. Densely connected convolutional networks[C]//2017 IEEE conference on computer vision and pattern recognition (CVPR). Honolulu:IEEE,2017:2261-2269.

[11] 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.

[12] XIE S Y,HU H F. Facial expression recognition with FRR-CNN[J]. Electronics Letters,2017,53(4):235-237.

[13] RYU B,RIVERA A R,KIM J,et al. Local directional ternary pattern for facial expression recognition[J]. IEEE Transactions on Image Processing,2017,26(12):6006-6018.

(下转第 181 页)