DOI:10.19651/j.cnki.emt.2005422

基于注意力机制的 CNN 人脸表情识别*

程换新 成 凯 蒋泽芹
(青岛科技大学 自动化与电子工程学院 青岛 266061)

摘 要：人脸表情识别在人机交互、情感计算等计算机视觉领域具有十分重要的应用前景。针对人脸表情识别的复杂性、多样性、遮挡性、光照等方面的挑战,提出了一种新的端到端网路,并将注意力力机制应用于表情自动识别。新的网络体系结构由特征提取模块、注意力模块、重构模块和分类模块四部分组成。通过 LBP 特征提取图像纹理信息,捕捉人脸的微小运动,提高网络性能。注意力机制可以使神经网络更加关注有用的特征,并结合 LBP 特征和注意力机制对注意力模型进行改进,提高识别精度。将新提出的方法应用于 3 个代表性的数据集,即 JAFFE、CK+和 FER2013,实验结果表明在 3 个数据集上人脸表情识别精度分别达到了 98.95%、98.95%和 79.89%,证明该方法利于提高人脸表情的识别率,具有一定的先进性。

关键词：人脸表情识别;卷积神经网络;特征提取;注意力机制
中图分类号：TP391.9 **文献标识码**：A **国家标准学科分类代码**：520.60

CNN facial expression recognition based on attention mechanism

Cheng Huanxin Cheng Kai Jiang Zeqin
(College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: Facial expression recognition has a very important application prospect in computer vision fields such as human-computer interaction and emotion calculation. Aiming at the challenges of complexity, diversity, occlusion and illumination of facial expression recognition, a new end-to-end network is proposed, and attention mechanism is applied to automatic expression recognition. The new network architecture consists of four parts: feature extraction module, attention module, reconstruction module and classification module. By extracting image texture information from LBP features, the tiny motion of face is captured and the network performance is improved. attention mechanisms can make neural networks pay more attention to useful features. we combine LBP features and attention mechanisms to improve the attention model to improve the recognition accuracy. applying the newly proposed method to two representative expression datasets, namely JAFFE、CK+、FER2013 and the experimental results show that the accuracy of facial expression recognition on three data sets is 98.95%,98.95% and 79.89%, respectively. It is proved that the method is beneficial to improve the recognition rate of facial expression and is advanced.

Keywords: facial expression recognition;convolutional neural network;feature extraction;attention mechanism

0 引 言

面部表情是人们日常交际中表达内心情感最直接的信号之一,一个人一次的身体或精神状态可以通过分析面部表情来获得。因此,面部表情识别在自动驾驶、人机交互、医疗等与面部表情相关的领域具有重要意义,并逐渐成为一个越来越重要的研究方向。由于人脸表情识别的复杂性、多样性、遮挡性、光照等方面的问题,在实际应用中的识别精度仍不尽人意。因此设计一个能够自动、准确地识别不同表情的模型。一般来说,表情识别的过程包括如下 3 个步骤:表情数据的预处理、表情的特征提取、表情分类。通常需要考虑两种特征,即人脸特征和人脸模型特征。面部特征是脸上的特定点,如眼睛、嘴和眉毛;面部模型特征是用来建模面部的特征。因此,人脸表示有多种方法,如用整张脸得到整体表示,用特定的点进行局部表示,将不同的点组合起来得到混合的方法。

当将表情识别作为一个分类问题来处理时,传统的方法通常使用手工制作的特征(如局部二进制模式(LBP))和

收稿日期:2020-11-24
* 基金项目:国家海洋局重大专项项目(国海科字[2016]494 号 No.30)资助

传统的机器学习算法(如支持向量机(SVM))进行分类。这些方法可能对实验室条件下收集的数据集有效,但随着在不可控环境中引入更具挑战性的表达数据集(如 FER2013),它们无法有效地完成这项任务。幸运的是,深度学习自从应用于图像分类问题以来,在方便性和有效性方面取得了突破。注意力机制已广泛应用于各种计算机视觉任务,如显著性检测[1]、人群计数[2]和面部表情识别。该操作通过学习一个中间注意力图,然后在注意力图和特征图上应用元素积来加权不同特征的重要性,从而选出最有用的特征进行分类。在人脸表情识别任务中,对识别有用的特征主要集中在眼睛、鼻子和嘴巴等关键部位。注意力机制增加了这些关键特征的权重,有助于提高表情识别的效果。

提出了一种基于注意力模型的卷积神经网络来识别面部表情。使用 LBP 特征提取优于使用 HOG 和 Gabor 特性提取是因为 LBP 可以实现旋转不变性和灰度不变性适用于不同尺度的纹理特征提取,可以解决位移的不平衡,面部图像的旋转角度和光照条件[3]。此外,LBP 特征可以反映精细面部皮肤纹理的变化。Fernandez 等[4]提出了一种基于注意力模型的端到端网络用于面部表情识别,注意力模块使网络更多地关注重要的有用特征,这使得网络识别更有效率。最后将 LBP 特征提取与面部注意力模型结合起来识别表情。所提出的方法在 3 个面部表情数据集上进行测试,分别是 CK+、JAFFE、FER2013。

1 相关工作

传统的人脸特征提取算法可以分为两类:基于几何的方法,如主动外观模型(AAM)[5];基于外观的方法,如 LBP[6]和 Gabor 表示法[7]。在特征提取后,特征被输入分类器,例如 SVM[8]和 KNN[9],用于识别不同的面部表情。因此,分类器的性能在很大程度上取决于提取特征的质量。

最早的卷积神经网络 LeNet-5[10]来用来识别笔迹,网络只有 7 层,包括 3 个卷积层、2 个采样层和 2 个全连通层,基于这种基本设计的网络的许多变体在深度学习中普遍存在。VGG 网络[11]使用非常小的卷积滤波器来增加架构深度,其中小尺寸滤波器可以使决策函数更具区分性并减少参数数量。网络通常堆叠几个卷积层,然后跟随一个汇集层。当网络中有 16 或 19 个权重层时,该架构可以实现对现有技术配置的显著改进。GoogLeNet[12]是一个 22 层的深度网络,与以前的网络相比,网络的宽度和深度都有所增加。网络的主要结构是“初始”层,它包含几个并行的卷积分支。“初始”层具有不同大小的卷积滤波器,并且输入图像可以在不同比例的特征地图上卷积。在 ResNet[13]中,网络中的输入层和输出层之间增加了跳过连接。这种结构不仅提高了训练速度,提高了模型的训练效果,而且避免了梯度消失和网络退化。

对于面部表情识别任务,大部分作品都是受上述深度网络架构的启发。在对静态图像进行分类时,提出了一种受 GoogLeNet[14]启发的面部表情识别网络。该网络包括两个卷积层,每个卷积层后面是一个最大池层,然后是 4 个初始层。Sun 等[15]提出了一种具有视觉注意力力的面部表情识别网络,在该网络中,从人脸中提取深度卷积特征,从而检测感兴趣区域并用于表情分类。Wu 等[16]提出了一个端到端网络,该网络具有面部表情识别的注意力模型。结果表明,注意力模块提高了分类性能。

在注意力模块中结合了 LBP 特征和卷积特征,借助于提供纹理信息并能反映人脸细微变化的 LBP 特征,可以提高注意力模块的能力,从而提高网络的识别精度。

2 方法设计

2.1 网络结构

新网络由 4 部分组成,即特征提取模块、注意力模块、重构模块和分类模块。该架构从由两个独立的 CNN 组成的特征提取模块开始:一个用于原始图像,另一个用于 LBP 特征图。网络的模型使用纯卷积层作为主干来提取特征。为了防止网络过于复杂,所有层都使用小尺寸卷积滤波器(3×3)。由于 VGG-16 网络具有很强的迁移学习能力和灵活的体系结构,它可以很容易地连接后端以提取更深层次的特征进行分类。为此,VGG-16 的前 10 层被用作网络模型的前端,以提取原始图像的初始特征。对于 LBP 特征图像,使用 VGG-16 的前 10 层来提取更深的特征,然后将维数降低到与原始图像相同。如图 1 所示,基于降维的 CNN 和基于特征提取的 CNN 架构相同。为了降低计算复杂度,在原 VGG-16 网上增加了 3 个 1×1×64 的卷积来减少信道。该网络的特征提取模块是获取初始特征,以便在后续模块中进行进一步处理。然后,网络将从原始图像中提取的特征 F1 与从 LBP 特征图像中提取的特征 F2 融合,然后将融合的特征 F3 添加到关注模块。注意力模块通过增加有用特征的权重来工作,并使网络更多地关注对表情识别至关重要的这些特征。这样,网络可以更高效地识别不同的表情。

在注意力模块之后,使用密集连接的卷积层作为重构模块来调整关注图,以便为分类模块创建增强的特征图。受 DenseNet[17]和 ResNet 的启发,使用密集的 atrous 卷积进行重建。通过在内核掩码的适当位置插入 0,可以用不同的速率扩展 atrous 内核。网络的稠密 atrous 卷积模块由 4 个 3×3 atrous 卷积组成,从较低层到较高层的膨胀率分别为 2、3、4 和 5。与传统的卷积算子相比,atrous 卷积能够在不增加核参数的情况下获得更大的感受视野。注意力模块的特征图包含表情识别的重要信息。对于每个图层,所有前面图层的要素图和要素 F1 连接为输入,其自身的要素图用作所有后续图层的输入。融合后的特征图 F3 通过注意力模块的输出进行元素求和运算。这种体系结构不仅可以提取更深层次的特征,而且有助于缓解重用有用特征

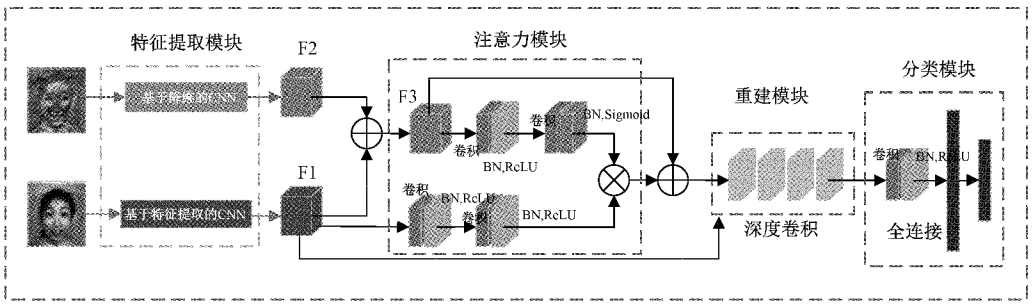

图 1　网络架构

的问题。最后，使用 softmax 的全连接层进行分类。该网络在每一层后使用批量归一化来加快网络的收敛速度，避免过拟合。

2.2　注意力机制

对于分类任务，通过提取图像特征，并根据这些特征之间的差异将图像分为不同的类别。然而，只有有用的特征才有助于分类，不同的特征贡献不同的意义。

注意力机制已经被证明在像素级计算机视觉任务中是有用的，并且被用来测量对不同区域的特征的关注程度。注意力模块包含两个分支，一个是获取特征的主干分支，另一个是整合 LBP 特征以获取注意力图的掩膜分支（mask branch）。如下所示：

$$F_{refine} = F_p \otimes F_m \tag{1}$$

提供掩膜分支中最后一层的输入，注意力图生成如下：

$$F_m = \mathrm{Sigmoid}(W \copyright f_m + b) \tag{2}$$

其中，W 和 b 分别是卷积层的权重和偏差；© 表示卷积运算。Sigmoid 激活函数给出(0,1)个概率分布，使网络区分不同特征的重要性。

2.3　局部二元模式

局部二值模式反映了有助于识别面部表情的基本信息。更具体地说，微妙的变化可以通过用 LBP 提取的特征来反映。此外，LBP 可以实现旋转不变性和灰度不变性，适用于提取不同尺度的纹理特征，可以解决人脸图像中的位移不平衡、旋转角度和光照条件等问题。Kim 等[18]首次引入了在 3×3 窗口中定义的原始 LBP。窗口的中心像素作为阈值，然后与相邻 8 个像素的灰度值进行比较。如果周围像素的值大于中心像素的值，则像素的位置标记为 1；否则，它为 0。这样，3×3 邻域内的 8 个点的比较可以生成 8 位二进制数，通常转化为 10 进制数，即 LBP 码，共有 256 种。通过以上步骤，得到窗口中心像素的 LBP 值，该值反映了区域的纹理信息。它可以定义如下：

$$LBP(x_c, y_c) = \sum_{p=0}^{7} S(i_p - i_c)2^p \tag{3}$$

其中，p 是像素数，(x_c, y_c)是中心像素的坐标，i_c 是中心像是邻域像素的像素值，i_p 是个采样点的像素值，S 是符号函数，定义如下：

$$S(x) = \begin{cases} 1, & x \geqslant 0 \\ 0, & x < 0 \end{cases} \tag{4}$$

原始 LBP 最大的缺点是在固定半径内只有很小的区域被它覆盖，这在纹理大小和频率不同时是不适合的。圆形 LBP 的提出是为了适应不同大小和频率的纹理特征，满足灰度和旋转不变性的需要。它不仅是 3×3 邻域，而且可以是任意邻域，圆形邻域也用正方形邻域代替。圆 LBP 允许半径为 R 的圆形邻域内任意多个像素点，得到半径为 R 的圆形区域内有 P 个采样点的 LBP 算子。它可以定义如下：

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} S(i_p - i_c)2^p \tag{5}$$

其中，p 是采样点数，r 是圆邻域的半径。为了解决二进制特征值编码模式过多的问题，提高统计性能，对 uniform patterns[19] 进行了进一步的扩展。当 bitwise pattern 是圆形时，LBP 最多包括两个从 0～1 或 1～0 的按位转换。它可以定义如下：

$$LBP_{P,R}^{rius2} = \begin{cases} \sum_{p=0}^{P-1} S(i_p - i_c), & U(LBP_{P,R}) \leqslant 2 \\ P+1, & \text{其他} \end{cases} \tag{6}$$

其中，U 是均匀性度量，$rius2$ 代表旋转不变的均匀图案。

3　实验与分析

3.1　实验准备

提出的人脸表情识别框架是用 Keras 实现的。框架以初始化学习率 0.000 001 进行训练，批量为 40。输入数据的大小为 256×256，因为使用大图像作为网络的输入可以使网络尽可能深，并有助于提取更多有用的特征，并每个卷积层之后添加批量归一化。最后在 3 个数据集上评估提出的方法，并将结果与一些最先进的方法进行比较。

3.2　数据集处理

为了保证网络能够达到良好的泛化能力，需要足够的训练数据。然而，大多数公开可用的数据集，如 JAFFE 等，没有足够数量的图像用于训练。数据量小，也会导致过拟合问题。因此，数据扩充对于面部表情识别很重要。标签保持变换是扩大数据集图像数量的最常用方法之一。

采用 4 种方法来放大原始数据的图像数,对图像进行随机旋转、翻转、移动和缩放。旋转范围为 0°～20°,宽度和高度的移动范围设置为 0%～15%。剪切范围和缩放范围都是 0～0.15。

3.3 实验结果与分析

FER2013 数据集包括 28 709 幅训练图像和 3 589 幅测试图像,其包含不同姿势、不平衡照明和遮挡的图片。将提出方法与最近的其他算法进行了评估和比较。表 1 给出了识别结果,表明该方法优于其他 5 种先进的算法。当没有 LBP 或注意力模块时,该数据集中的识别率会降低。

表 1 FER2013 数据集上不同方法的比较

方法	识别率/%
文献[18]	74.32
文献[19]	72.25
文献[20]	71.21
文献[21]	75.32
本文	79.89
本文(不含 LBP 算法)	67.73
本文(不含注意力机制)	73.96

CK+数据集包括 593 个图像序列,这些序列中的 327 个用 6 种面部表情(即,愤怒、厌恶、恐惧、快乐、悲伤和惊讶)标记。每个图像序列从中性脸开始逐渐达到一个峰值表情。对于这 6 种表达式中的每一种,选择最后 3 个具有峰值信息的作为新数据集。将提出方法与一些有代表性的方法在该数据集上进行比较,结果如表 2 所示,这表明该方法优于大多数方法。当移除 LBP 或注意力模块时,识别结果会分别降低。这表明 LBP 和注意力模块可以提高识别精度。

表 2 CK+数据集上不同方法的比较

方法	识别率/%
文献[18]	96.92
文献[19]	96.12
文献[21]	97.25
文献[22]	96.32
本文	98.95
本文(不含 LBP 算法)	97.56
本文(不含注意力机制)	97.01

JAFFE 数据集包括 213 幅姿势表情图像,其中每个图像都有一个表情,是愤怒、厌恶、恐惧、快乐、悲伤、惊讶和中性这几种表情之一。将提出方法与现有方法进行了比较,比较结果如表 3 所示,表明该方法在 JAFFE 数据集上优于其他方法。当不使用 LBP 或注意力力时,识别率会降低。

表 3 JAFFE 数据集上不同方法的比较

方法	识别率/%
文献[18]	91.84
文献[19]	95.74
文献[21]	91.33
文献[23]	95.21
本文	98.32
本文(不含 LBP 算法)	96.51
本文(不含注意力机制)	95.32

4 结 论

提出了一种新的用于面部表情识别的卷积神经网络,该方法融合了 LBP 特征和卷积特征,并结合注意力机制来提高网络的性能。为了防止过拟合和保证网络的泛化能力,在实验中使用的数据集应用了数据扩充。该方法在 JAFFE、CK+和 FER2013 三个著名的人脸表情数据集上进行了评价。实验结果表明,在这些数据集上,该方法优于现有的方法。然而,该方法只适用于 2D 图像。未来,将提出改进的方法,使其适用于视频数据、3D 人脸数据集,并探索更好的机器学习方法来增强网络。

参考文献

[1] ZHAO T, WU X. Pyramid feature attention network for saliency detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 3085-3094.

[2] VARIOR R R, SHUAI B, TIGHE J, et al. Scale-aware attention network for crowd counting[J]. ArXiv Preprint,2019, ArXiv:1901.06026.

[3] MEHTA D, SIDDIQUI M F H, JAVAID A Y. Recognition of emotion intensities using machine learning algorithms: A comparative study[J]. Sensors (Basel, Switzerland),2019,19(8):1897.

[4] FERNANDEZ P D M, PEÑA F A G, REN T I, et al. FERAtt: Facial expression recognition with attentionNet[J]. ArXiv Preprint,2019,ArXiv:1902.03284.

[5] COOTES T, EDWARDS G, TAYLOR C, et al. Active appearance models[C]. IEEE Trans. Pattern Anal. Mach. Intell., 2001,23(6):681-685.

[6] 曹亚媛,郭秀才,程勇.基于视觉模糊的 LBP 鲁棒特征提取与匹配[J].光电子·激光,2021,32(4):361-372.

[7] 李云红,聂梦瑄,苏雪平,等.分区域特征提取的人脸识别算法[J].西北大学学报(自然科学版),2020,50(5):811-818.

[8] 庙传杰,史东承.基于改进的局部二值模式和 SVM 的人脸识别[J].长春工业大学学报,2020,41(3):257-262.

[9] ZHANG K, HUANG Y, DU Y, et al. Facial expression recognition based on deep evolutional spatial-temporal networks[J]. IEEE Transactions on Image Processing, 2017, 26(9): 4193-4203.

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

[11] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014:111-119.

[12] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]. IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2015:1-9.

[13] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[J]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016:770-778.

[14] MOLLAHOSSEINI A, CHAN D, MAHOOR M H. Going deeper in facial expression recognition using deep neural networks [C]. 2016 IEEE Winter Conference on Applications of Computer Vision(WACV), IEEE, 2016: 1-10.

[15] SUN W, ZHAO H, JIN Z. A visual attention based ROI detection method for facial expression recognition [J]. Neurocomputing, 2018, 296: 12-22.

[16] WU M, SU W, CHEN L, et al. Weight-adapted convolution neural network for facial expression recognition in human-robot interaction[M]. IEEE, 2019.

[17] HUANG G, LIU Z, LAURENS V, et al. Densely connected convolutional networks[C]. IEEE Computer Society, IEEE Computer Society, 2016:2261-2269.

[18] 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]. Computer Vision & Pattern Recognition Workshops. IEEE, 2016:1499-1508.

[19] TURAN C, LAM K M, HE X. Soft locality preserving map (SLPM) for facial expression recognition[J]. ArXiv Preprint,2018, ArXiv:1801.03754.

[20] WANG W, SUN Q, CHEN T, et al. A fine-grained facial expression database for end-to-end multi-pose facial expression recognition [J]. ArXiv Preprint, 2019, ArXiv:1907.10838.

[21] OJALA T, HARWOOD I. A comparative study of texture measures with classification based on feature distributions[J]. Pattern Recognition, 1996, 29(1): 51-59.

[22] RODRIGUEZ P, CUCURULL G, GONALEZ J, et al. Deep pain: Exploiting long short-term memory networks for facial expression classification[J]. IEEE Transactions on Cybernetics, 2017(99):1-11.

[23] HAMESTER D, BARROS P, WERMTER S. Face expression recognition with a 2-channel Convolutional Neural Network[C]. International Joint Conference on Neural Networks, IEEE, 2015:1-8.

作者简介

程换新，工学博士，教授，主要研究方向为控制科学与工程、人工智能、图像识别等。

成凯(通信作者)，硕士生，主要研究方向为人工智能、图像识别等。

蒋泽芹，硕士生，主要研究方向为人工智能、图像识别等。E-mail:969206242@qq.com