DOI:10.19651/j.cnki.emt.2107000

基于 CNN 和 LSTM 的人脸表情识别模型设计*

程换新 王 雪 程 力 孙胜意
(青岛科技大学 自动化与电子工程学院 青岛 266061)

摘 要:人脸表情能够正确地反映人的内心活动,但由于表情的复杂性和微妙性,准确地识别人脸表情仍然是一大难题。本文设计了一种基于卷积神经网络(CNN)和长短期记忆神经网络(LSTM)的方法让计算机能够识别人脸的表情,损失函数采用 Focal loss。该框架包括 3 个方面:采用两种不同的预处理技术处理光照变化,并保留图像的边缘信息;预处理后的图像被输入到两个独立的 CNN 层用于提取特征;将提取到的特征与 LSTM 层融合。使用 FER2013、JAFFE 和 CK+ 3 个数据集验证模型准确性,并选择 FER2013 数据集制作混合矩阵,结果为该模型在 FER2013 数据集上的准确率相比于目前先进模型提升了 9.65%,在 JAFFE 和 CK+数据集上也表现良好,结果表明所提出的模型具有较强的泛化能力。

关键词:人脸表情;CNN;LSTM;Focal loss
中图分类号:TP183 **文献标识码**:A **国家标准学科分类代码**:520.60

Facial expression recognition model design based on CNN and LSTM

Cheng Huanxin Wang Xue Cheng Li Sun Shengyi
(College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: Facial expressions can correctly reflect people's inner activities, but due to the complexity and subtlety of facial expressions, accurate recognition of facial expressions is still a big problem. This paper designs a method based on convolutional neural network (CNN) and long short term memory (LSTM), so that the computer can recognize the expression of human face the loss function uses Focal loss. The framework includes three aspects: two different preprocessing techniques are used to deal with the illumination change and preserve the edge information of the image. The preprocessed image is input into two independent CNN layers for feature extraction. The extracted features are fused with LSTM layer. Using FER2013, JAFFE and CK+ data sets to verify the accuracy of the model, and select FER2013 data set to make a mixed matrix. The results show that the accuracy of the model on FER2013 data set is improved by 9.65% compared with the current advanced model, and it also performs well on JAFFE and CK+ data sets. The results show that the proposed model has strong generalization ability.

Keywords: facial expression; CNN; LSTM; Focal loss

0 引 言

随着计算机和人工智能技术的迅猛发展,人类对人机智能交互需求日益强烈。如果计算机能过通过对象面部表情的变化来获取其情感的变化,那么计算机就能够更好地为人类服务。然而在实际获取到的表情数据中,经常由于光照、头部姿态以及图像分辨率等因素影响模型检测的准确性。因此,准确的人脸表情识别依然是现在研究的热点。

人脸特征提取是人脸表情识别的核心步骤,传统的表情特征提取主要采用数学方法,这种方法需要大量的专业知识才能设计出相对满意的提取器。Gupta 使用 SVM 的方法在 CK+数据集上取得 93.7%的准确性[1]。但目前基于深度学习的模型能有效地提取表情特征,使计算机深度理解人脸表情图像传达的意义。深度学习通过反向传播和优化损失函数迭代更新权重值,在学习了大量的样本之后可以提取到更深层次、更加抽象的特征,而且,所提取的特征更能表征人脸表情的本质。因此,近年来众多的学者运用深度学习方法识别人脸表情,取得了较满意的效果。

近年来,应用深度学习方法能够在图像序列的情感识别中取得良好的效果。主要是应用卷积神经网络[2],并且

收稿日期:2021-06-21
* 基金项目:国家海洋局重大专项(国海科字[2016]494 号 No.30)资助

在面部表情识别方面取得了显著的成就[3-8]。由于神经网络通过数据集进行训练,所以它的性能通常比传统方法好。但是考虑到实时性时,深度学习难以处理时间和空间信号从而获得更好的情感识别性能。许多二维 CNN 无法识别时间信息,在这种情况下,研究人员研发出了一种将时间和空间特征综合的方法,即 CNN-LSTM[9]。受到该框架的启发,本文采用 CNN-LSTM 模型用于从数据集和实时环境中获取的图像序列。

如今,CNN[10]在图像分类和计算机视觉方面都有突破。2012 年 Krizhevsky 等[11]提出了 AlexNet 的深度 CNN,该网络能够实现对超过百万张图片的大型数据集进行分类,它由几个卷积层、1 个最大池化层、1 个 ReLU 激活函数的全连接层和 1 个 softmax 层组成。类似的,其他比较流行的 CNN 网络还有 VGGNet[12]、Google Net[13]和 ResNet[14]等,这些网络在同样的数据库上也展现出来较好的性能。

使用 CNN 网络提取特征时会忽略时间信息,当处理实时视频信息时会不精确。因此,学者们研发处理几种能够同时学习时间特征和空间特征的深度学习网络。Fan 等[15]采用 CNN-RNN 混合模型进行情感识别,并将混合模型特征映射与 3D-CNN 相结合,在同时具有音频和视频输入的情况下展现了更好的性能。Donahue 等[9]将 CNN 与 LSTM 相结合的神经网络模型,用于联合学习空间和时间特征来完成不同的目标识别任务。

Hadsell 等[16]提出了 contrastive loss 损失函数,其目的是增大类间差异的同时减少类内差异。Schroff 等[17]提出 triplet loss 损失函数,能够更好的对细节进行区分,但 triplet loss 收敛速度慢,导致模型性能下降。Byoung[18]提出了中心损失函数 center loss,让样本绕类内中心均匀分布,最小化类内差异,但计算效率太低。为了解决样本之间不平衡问题,Lin 等[19]提出 Focal loss 损失函数,通过聚焦参数 γ 使模型更多的关注难分类样本,提高模型分类性能,但是不能解决标注样本问题。本文在使用 CNN-LSTM 模型的基础上,将 Focal loss 应用于人脸表情识别中,相对于传统交叉熵损失函数,应用 Focal loss 能够提高模型的准确率。

1 模型介绍

本文所使用的模型包括 3 个阶段:首先,其一是使用预处理技术处理光照方差,另一个是增强每个图像的边缘。其次,使用双层 CNN 层生成特征图,加权直方图均衡为输入 1,边缘增强为输入 2,分别送给两个 CNN。最后,将两层 CNN 的特征与 LSTM 融合,映射到一个全局池化层,通过该层减少特征数量。因此,可通过 Softmax 层得到预测每个表情的概率。图片预处理过程如图 1 所示。

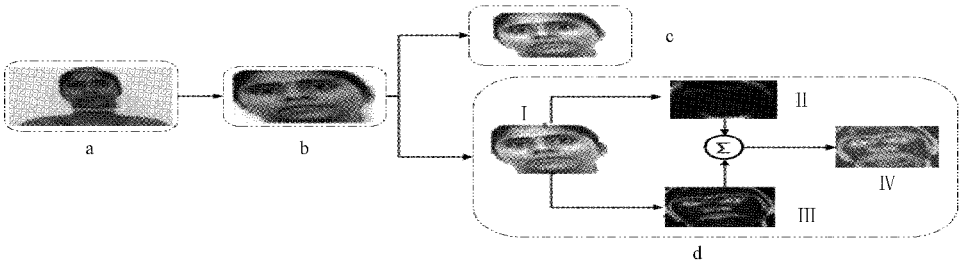

图 1 图片预处理

其中每个部分采用的技术为:a 脸部识别框,b 灰度转换并调整大小,c 加权直方图均衡化,d 边缘检测;在 d 中:I 图像锐化技术,II 几何距离处理技术,III 距离技术,IV 应用边缘增强技术(使用 II 和 III)。本文所提出的 CNN-LSTM 模型体系结构如图 2 所示。

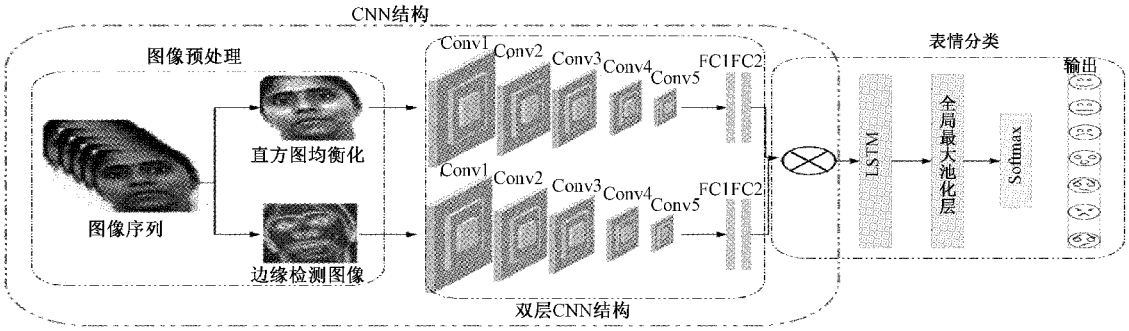

图 2 CNN-LSTM 结构

1.1 图像预处理

为了统一输入数据,首先检测人脸区域,然后,裁剪并调整检测到的人脸大小 96×96 像素,调整后的概率密度函数计算式为:

$$P(G_k)=\frac{N_k}{N} \tag{1}$$

其中，G_k 为绘图图像，N_k 为 G_k 出现的次数，N 为图像中像素的总个数。

接下来通过变换函数 $T(\cdot)$ 提高图像的质量，加权直方图均衡化为：

$$P_w(G_k)=T\{P(G_k)\} \tag{2}$$

$$P_w(G_k)=\begin{cases}P_w(G_k), & P(G_k)>\tau_1\\ 0, & P(G_k)<\tau_2\end{cases} \tag{3}$$

其中，τ_1 和 τ_2 为上下限，定义如下，其中 β 为权重函数，且 $\beta<1$ 且非零。

$\tau_1 \rightarrow \beta \max\{P(G_k)\}$

$\tau_2 \rightarrow G_k<1$

因此，直方图均衡化的计算公式如下：

$$P(G_k)=\left[\frac{P(G_k)-\tau_2}{\tau_1-\tau_2}\right]\times\tau,\tau_1<\tau<\tau_2 \tag{4}$$

使用上述公式以处理不同光照条件的图像，将其作为 CNN 的输入之一。其次，对同一组图像进行边缘增强，使用欧氏距离和倒角距离增强边缘像素。距离是用最近的距离值标记图像中的像素值，欧氏距离采用平均距离，它们的公式如下：

$$D_{ou}(p,q)=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \tag{5}$$

$$D_{dao}(M,N)=\frac{1}{p}\sum_{m_i\in M}\min_{n_j\in N}|m_i-n_j| \tag{6}$$

1.2　特征提取

CNN 网络结构通常包含有卷积层、激活函数、池化层和全连接层。本文所用网络中，使用了批次调整和 ReLU 激活函数在卷积层和池化层中间，因此该网络能够有效地学习图像的特征。ReLU 激活函数具体公式如下：

$$\mathrm{ReLU}(x)=\max(0,x) \tag{7}$$

本文所使用的 CNN 由 5 个卷积层，滤波器大小为 5×5，池化层选择最大池化，经过 5 次特征提取后，提取到的特征发送到全连接层，与 LSTM 层集成在一起。具体每层参数如表 1 所示。

1.3　Focal loss 损失函数

表情识别任务中，多分类任务中的损失函数会由于样本的不均衡导致分类性能差。因此，Lin 等[19]针对这个问题，对标准交叉熵损失函数的基础上进行改进，提出了聚焦损失函数(Focal loss)，公式如下：

$$FL=-\alpha(1-p_i)^{\gamma}\ln p_i \tag{8}$$

其中，平衡参数 α 的作用是控制不平衡样本对总损失的权重，平衡不同类别样本的数量。聚焦参数 γ 是一个大于 0 的超参数，用来控制易于分类和难于分类样本的权重。若样本被错分则 p_i 是无穷小的数，因此 $(1-p_i)^{\gamma}$ 接近 1，当正确分类即 p_i 为 1 时，$(1-p_i)^{\gamma}$ 接近 0。Focal loss 通过控制调制系数使得可应用于困难样本中，通过平衡参数 α 实现对不同类别样本数量达到平衡的目的。

表 1　CNN-LSTM 结构各层参数

层	滤波器大小	步长	大小
输入	—	—	96×96
Conv1	32	1	5×5
最大池化层 1	—	2	3×3
Conv2	32	1	5×5
最大池化层 2	—	2	3×3
Conv3	64	1	5×5
最大池化层 3	—	2	3×3
Conv4	128	1	5×5
最大池化层 4	—	2	3×3
Conv5	256	1	5×5
最大池化层 5	—	2	3×3
FC1	512	—	—
FC2	7	—	—

2　实验分析

2.1　实验环境

本实验所用计算机处理器为 AMD Ryzen 7-1700 型号，显卡采用 GTX 1080 TITAN，内存为 16 GB。操作系统为 Windows10 64 位，神经网络部分使用开源的 Keras 模块搭建，软件编程环境为 python3.0。

2.2　人脸表情数据集

为了证明提出体系结构的有效性，在 CK+、FER2013 和 JAFFE 的面部表情数据库上做了实验，各个数据库中 6 种表情的图片数量如表 2 所示。

表 2　FER2013、JAFFE、CK+中 6 种表情数量

数据集	愤怒	厌恶	恐惧	开心	悲伤	惊讶
FER2013	4 953	547	5121	8 989	6 077	4 002
JAFFE	30	32	31	31	30	30
CK+	135	177	75	207	84	249

2.3　实验结果与分析

1）使用直方图均衡化和边缘增强两种方法对 FER2013、JAFFE 和 CK+进行预处理，并将处理后的图片使用本文所提出的模型验证准确性，两种预处理模型准确率对比如表 3 所示，看到使用边缘增强得到的准确率高 3%～8%，并且使用边缘增强的预处理方法后在 FER2013 数据集达到 80.14%的准确率。

表 3　使用不同预处理方法的准确率　%

模型	直方图均衡化	边缘增强
FER2013	77.32	**80.14**
JAFFE	70.14	**78.25**
CK+	71.56	**79.67**

2)为了验证在交叉数据集上的有效性,根据 FER2013 数据集制作了混合矩阵。如表 4 所示,其中,列代表预测的类别,行代表真实的类别,对角线数据为预测的准确率,其余为预测错误的概率。

表 4 FER2013 混合矩阵 %

表情	愤怒	厌恶	恐惧	开心	悲伤	惊讶
愤怒	**91.3**	2.6	0.9	0	0	0
厌恶	2.7	**90.1**	1.2	0	0	0
恐惧	0	0	**88.2**	4.8	1.35	5.6
开心	1.1	0	0.9	**92.6**	0	0
悲伤	0	0	0.99	0	**91.1**	0
惊讶	0.3	0	0	0	0.7	**89.2**

3)为了测试本文模型的识别能力,设计了基于真实人脸表情的仿真实验,对每帧画面进行表情识别,如图 3 所示。实验结果表明,在真实条件下,本文算法具有较好的泛化性。

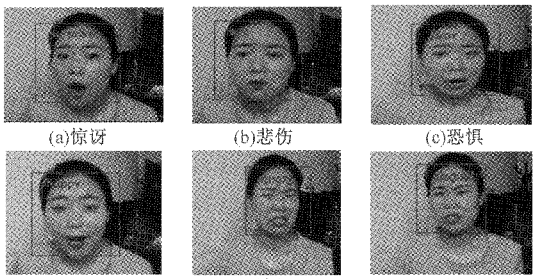

图 3 各种表情识别测试效果

4)最后,与国际上已有的几种方法在 FER2013 数据集上的准确率相比较。与其他方法比较,本文所提出的 CNN-LSTM 模型准确率相比于目前准确率最高的算法高出 9.65%,具体数据如表 5 所示。

表 5 不同方法在 FER2013 数据集上识别的准确率

%

模型	准确率
CNN+Improve_Softmax[20]	70.91
IcRL[21]	72.36
CPC[22]	71.35
DNNRL[23]	70.60
SHCNN[24]	69.10
VGG+Focal loss[25]	72.49
本文算法	**82.14**

3 结 论

本文提出了一种新的人脸表情识别的方法,通过使用边缘增强和直方图均衡化两种预处理方法,处理不同环境中各种光照,由表 3 可以得出通过边缘增强的预处理准确率更高。此外,本文将 CNN 与 LSTM 两个模块结合使用,应用 CNN 模块学习空间特征,应用 LSTM 模块学习时间特征,并用 Focal loss 损失函数代替传统损失函数使用,使得本文模型在分类效果上更好,通过表 5 的实验结果也能说明这一点,该模型在训练样本相对于局部信息较小的情况下更具有竞争力。在未来的研究中,希望可以实现将表情识别转向室外,能够应用在更复杂的环境中。

参考文献

[1] GUPTA S. Facial emotion recognition in real-time and static images[C]. IEEE, 2018: 553-560.

[2] ROSKA T, CHUA L O. The CNN universal machine: an analogic array computer [J]. IEEE Transactions on Circuits & Systems II Analog & Digital Signal Processing, 2015, 40(3): 163-173.

[3] BYEON Y H, KWAK K C. Facial expression recognition using 3D convolutional neural network[J]. International Journal of Advanced Computer Science & Applications, 2014, 5(12): 107-112.

[4] LOPES A T, AGUIAR E D, OLIVEIRA-SANTDS T. A facial expression recognition system using convolutional network[C]. Proc. Int. Conf. SIBGRAPI Graphics, Patterns and Images, 2015: 273-280.

[5] SULTAN G, PALUMBO A. Facial expression recognition using convolutional neural network[C]. Proc. Int. Conf. Vision, Image and Signal Processing (ICVISP), 2021, DOI: 10.13140/RG.2.2.30784.46086.

[6] YANG H, YIN L. CNN based 3D facial expression recognition using masking and landmark features[C]. International Conference on Affective Computing & Intelligent Interaction, 2017: 556-560.

[7] 祝勇俊, 刘文波, 郑祥爱,等. 基于 FOCUSS 改进算法的图像稀疏重构[J]. 电子测量技术, 2020,43(4): 126-131.

[8] 姚品, 万旺根. 基于深度学习和属性特征的行人再识别算法[J]. 电子测量技术, 2020, 43(12): 70-74.

[9] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description [M]. Elsevier, 2015: 677-691.

[10] WERBOS L, WERBOS P. Self-organization in CNN-based object nets[C]. Cellular Nanoscale Networks and Their Applications (CNNA), 2010, DOI: 10.1109/CNNA.2010.5430292.

[11] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks [C]. Advances in Neural Information Processing Systems, Curran Associates Inc, 2012:

1097-1105.

[12] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014, ArXiv:1409.1556.

[13] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[J]. IEEE Computer Society, 2014: 1-9.

[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[J]. IEEE, 2016: 770-778.

[15] FAN Y, LU X J, LI D, et al. Video-based emotion recognition using CNN-RNN and C3D hybrid networks[J]. Proc. Int. Conf. Multi-modal Interaction, 2016: 445-450.

[16] HADSELL R, CHOPRA S, LECUN Y. Dimensionality reduction by learning an invariant mapping [C]. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE, 2006: 1735-1742.

[17] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 815-823.

[18] BYOUNG K. A brief review of facial emotion recognition based on visual information[J]. Sensors, 2018, 18(2), DOI: 10.3390/s18020401.

[19] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980-2988.

[20] ZHOU J, JIA X, SHEN L, et al. Improved softmax loss for deep learning-based face and expression recognition[J]. Cognitive Computation and Systems, 2019, 1(4): 97-102.

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

[22] CHANG T, WEN G, HU Y, et al. Facial expression recognition based on complexity perception classification algorithm[J]. ArXiv Preprint, 2018, ArXiv:1803.00185.

[23] JEON J, PARK J C, JO Y J, et al. A real-time facial expression recognizer using deep neural network[C]. Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, 2016: 1-4.

[24] MIAO S, XU H, HAN Z, et al. Recognizing facial expressions using a shallow convolutional neural network[J]. IEEE Access, 2019, 7: 78000-78011.

[25] 崔子越，皮家甜，陈勇，等. 结合改进 VGGNet 和 Focal Loss 的人脸表情识别[J]. 计算机工程与应用, 2007, DOI: 10.3778/j.issn.1002-8331.2007-0492.

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。
E-mail: 3555184923@qq.com

王雪，研究生，主要研究方向为人工智能、图像识别等。
E-mail: 2388861238@qq.com

程力，工学博士，研究员，主要研究方向为大数据分析、人工智能、互联网网络安全等。
E-mail: chengli@ms.xjb.ac.cn

孙胜意，研究生，主要研究方向为人工智能、图像识别等。
E-mail: 1622844945@qq.com