



文章编号:1001-9081(2019)09-2568-07

DOI:10.11772/j.issn.1001-9081.2019030540

基于多尺度核特征卷积神经网络的实时人脸表情识别

李旻择¹, 李小霞^{1,2*}, 王学渊^{1,2}, 孙维¹

(1. 西南科技大学 信息工程学院, 四川 绵阳 621010; 2. 特殊环境机器人技术四川省重点实验室(西南科技大学), 四川 绵阳 621010)

(* 通信作者电子邮箱 664368504@qq.com)

摘要:针对人脸表情识别的泛化能力不足、稳定性差以及速度慢难以满足实时性要求的问题,提出了一种基于多尺度核特征卷积神经网络的实时人脸表情识别方法。首先,提出改进的 MobileNet 结合单发多盒检测器(MSSD)轻量化人脸检测网络,并利用核相关滤波(KCF)模型对检测到的人脸坐标信息进行跟踪来提高检测速度和稳定性;然后,使用三种不同尺度卷积核的线性瓶颈层构成三条支路,用通道合并的特征融合方式形成多尺度核卷积单元,利用其多样性特征来提高表情识别的精度;最后,为了提升模型泛化能力和防止过拟合,采用不同的线性变换方式进行数据增强来扩充数据集,并将 FER-2013 人脸表情数据集上训练得到的模型迁移到小样本 CK+ 数据集上进行再训练。实验结果表明,所提方法在 FER-2013 数据集上的识别率达到 73.0%,较 Kaggle 表情识别挑战赛冠军提高了 1.8%,在 CK+ 数据集上的识别率高达 99.5%。对于 640×480 的视频,人脸检测速度达到每秒 158 帧,是主流人脸检测网络多任务级联卷积神经网络(MTCNN)的 6.3 倍,同时人脸检测和表情识别整体速度达到每秒 78 帧。因此所提方法能够实现快速精确的人脸表情识别。

关键词:人脸表情识别;卷积神经网络;人脸检测;核相关滤波;迁移学习

中图分类号: TP391.4 **文献标志码:**A

Real-time facial expression recognition based on convolutional neural network with multi-scale kernel feature

LI Minze¹, LI Xiaoxia^{1,2*}, WANG Xueyuan^{1,2}, SUN Wei¹

(1. School of Information Engineering, Southwest University of Science and Technology, Mianyang Sichuan 621010, China;

2. Key Laboratory of Special Environmental Robotics in Sichuan Province (Southwest University of Science and Technology), Mianyang Sichuan 621010, China)

Abstract: Aiming at the problems of insufficient generalization ability, poor stability and difficulty in meeting the real-time requirement of facial expression recognition, a real-time facial expression recognition method based on multi-scale kernel feature convolutional neural network was proposed. Firstly, an improved MSSD (MobileNet + Single Shot multiBox Detector) lightweight face detection network was proposed, and the detected face coordinates information was tracked by Kernel Correlation Filter (KCF) model to improve the detection speed and stability. Then, three linear bottlenecks of three different scale convolution kernels were used to form three branches. The multi-scale kernel convolution unit was formed by the feature fusion of channel combination, and the diversity feature was used to improve the accuracy of expression recognition. Finally, in order to improve the generalization ability of the model and prevent over-fitting, different linear transformation methods were used for data enhancement to augment the dataset, and the model trained on the FER-2013 facial expression dataset was migrated to the small sample CK+ dataset for retraining. The experimental results show that the recognition rate of the proposed method on the FER-2013 dataset reaches 73.0%, which is 1.8% higher than that of the Kaggle Expression Recognition Challenge champion, and the recognition rate of the proposed method on the CK+ dataset reaches 99.5%. For 640×480 video, the face detection speed of the proposed method reaches 158 frames per second, which is 6.3 times of that of the mainstream face detection network MTCNN (MultiTask Cascaded Convolutional Neural Network). At the same time, the overall speed of face detection and expression recognition of the proposed method reaches 78 frames per second. It can be seen that the proposed method can achieve fast and accurate facial expression recognition.

Key words: Facial Expression Recognition (FER); Convolutional Neural Network (CNN); face detection; Kernel Correlation Filter (KCF); transfer learning

0 引言

人的丰富情感信息。在当下的人工智能时代,人机交互在日常生活

中越来越普及,想要让机器更好地理解人类,人脸表情的识别是必不可少的途径。人脸表情识别(Facial Expression

收稿日期:2019-04-03;修回日期:2019-06-07;录用日期:2019-06-10。

基金项目:国家自然科学基金资助项目(61771411);四川省科技计划项目(2019YJ0449);西南科技大学研究生创新基金资助项目(18ycx123)。

作者简介:李旻择(1992—),男,四川南充人,硕士研究生,CCF 会员,主要研究方向:深度学习、计算机视觉; 李小霞(1976—),女,四川安岳人,教授,博士,主要研究方向:模式识别、计算机视觉; 王学渊(1974—),男,四川绵阳人,副教授,博士,主要研究方向:图像处理; 孙维(1995—),男,四川达州人,硕士研究生,主要研究方向:图像处理。



Recognition, FER)是计算机视觉、人工智能等领域的重要研究方向,它是一个有趣且具有挑战性的问题,在教育、医疗、心理分析等领域均具有重要的研究价值与意义。

1971年,著名的心理学家 Ekman^[1]将人脸部表情划分为六类基本表情:愤怒、厌恶、恐惧、高兴、悲伤和惊讶,并提出表情可以通过观察面部信号来识别。例如“高兴”是看到令人高兴的事情或者听到好的消息,一般会通过嘴角抬高、眼睛变小来表达。此后的研究均是在这六种基本表情基础上展开的,用于人脸表情识别的各种特征提取算法和分类器被相继开发出来。典型的表情特征提取方法有局部二值模式(Local Binary Pattern, LBP)^[2]、方向梯度直方图(Histograms of Oriented Gradients, HOG)^[3]、Gabor 小波变换^[4]、尺度不变的特征变换(Scale Invariant Feature Transform, SIFT)^[5]、主动外观模型(Active Appearance Model, AAM)^[6]等,典型的表情分类方法有隐马尔可夫模型(Hidden Markov Model, HMM)法^[7]、支持向量机(Support Vector Machine, SVM)^[8]、局部线性嵌入(Local Linear Embedding, LLE)^[9]、K 最近邻(K-Nearest Neighbors, KNN)算法^[10]等。这些研究大多都是人工提取特征,因此效果优劣依赖于前期的特征提取,人为干扰因素较多,且泛化能力不足。

2012年,Krizhevsky 等^[11]在 ILSVRC-2012 中使用 AlexNet 卷积神经网络(Convolutional Neural Network, CNN)取得了惊人成绩,其识别率远超其他人工特征的传统方法。随后深度神经网络使得人脸表情识别得到了进一步的发展,用于人脸表情识别的各种数据集也日益增多,常见的有 JAFFE^[12]、Extensive Cohn-Kanade(CK+)^[13]、FER-2013^[14]、SFEW2.0^[15]等。2013年,Tang^[16]提出将 CNN 与 SVM 相结合,并且放弃了普通 CNN 所使用的交叉熵损失最小化方法,而是用标准的铰链损失来最小化基于边际的损失。他的方法在私有测试集上实现了 71.2% 的识别率,获得了 FER-2013 人脸表情识别挑战赛的冠军。2017 年,Al-Shabi 等^[17]通过合并 CNN 和 SIFT 特征,建立了一个混合 CNN-SIFT 分类器,使得小样本数据也能够有较好的识别效果,在 CK+ 数据集上的识别率达到

了 99.4%。人脸表情的识别率虽然在逐步升高,但识别速度却很低,想要实际应用还很难实现。2014 年,Fang 等^[18]提出了一种新的人脸表情自动分析框架,选择具有峰值表情的帧来提取突出信息,实现了 3.5 帧/s(Frame Per Second, FPS)的识别速度。2016 年,Jeon 等^[19]使用 HOG 特征来检测人脸,CNN 来提取特征,在 FER-2013 数据集上实现了 70.7% 的识别率,6.5 帧/s 的识别速度。2017 年,Nehal 等^[20]提出了一种智能层次支持向量机,用多级的 SVM 来减少混淆表情间的相互关系,获得了 10.8 帧/s 的识别速度。

用深度神经网络来进行人脸表情识别,虽然能够减少人为干扰因素、提高稳定性,但是要拥有较高的识别率,网络模型一般都比较大,参数量太大导致速度很慢,使得其难以满足实时性需求。针对这个问题,以单发多盒检测器(Single Shot multibox Detector, SSD)网络^[21]为基础,改进的 MobileNet-SSD(MSSD)轻量化人脸检测网络被提出,来进行人脸的检测,结合核相关滤波(Kernel Correlation Filter, KCF)算法^[22]进行人脸的跟踪,可大幅提高人脸的检测速度并且提高多角度和遮挡的人脸检测的稳定性;然后利用迁移学习方法,将两个数据集进行联合训练,并且使用多尺度核特征 CNN 来进行人脸表情特征提取与识别,进一步提高识别率;最后将以上两个模型相融合,实现快速精确的实时人脸表情识别。

1 实时人脸表情识别系统概述

一个完整的实时人脸表情识别系统包括:人脸检测与定位、表情特征提取和表情分类。针对实际应用中需要兼顾识别速度与精度的问题,首先将改进的 MSSD 人脸检测网络和 KCF 快速跟踪模型相结合,进行人脸目标的快速稳定检测;然后进行人脸的切片,再将人脸切片(即单纯人脸图片)输入已经训练好的多尺度核特征 CNN 进行表情识别;最后,将以上两个网络进行串联融合,整个过程形成了检测-跟踪-识别模式,构成了一个完整的实时人脸表情识别系统。图 1 是实时人脸表情识别系统总体流程。

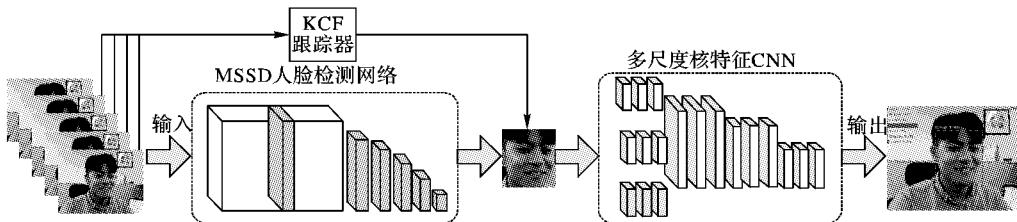


图 1 实时人脸表情识别系统总体流程

Fig. 1 Overall process of real-time facial expression recognition system

2 快速稳定的人脸检测

2.1 MSSD 人脸检测网络

目标检测网络一般由基础网络进行特征提取,元结构进行分类回归和边界框回归。以 SSD 目标检测网络为基础,首先将原基础网络 VGG-16^[23]改为轻量化网络 MobileNet^[24],然后将其中的第 7 个深度可分离卷积层(浅层特征)与最后 5 层(深层特征)的特征图进行融合,改进为 MSSD 网络,网络模型如图 2 所示。MobileNet 中最大的亮点就是深度可分离卷积,它由深度卷积和点卷积组成,极大地加快了训练与识别

的速度,因此采用深度可分离卷积来构建网络。在 MSSD 网络中,输入端通过 1 个卷积核大小为 3×3 、步长为 2 的标准卷积层,再经过 13 个深度可分离卷积层,后面输出端连接了 4 个卷积核分别为 1×1 、 3×3 交替组合的标准卷积层和 1 个最大池化层,考虑到池化层会损失一部分有效特征,因此在网络的标准卷积层中使用了步长为 2 的卷积核替代池化层。

网络浅层特征的感受野较小,拥有更多的细节信息,对小目标的检测更具优势,因此 MSSD 人脸检测网络采用浅层与深层特征融合的方式。经实验分析,将第 7 层的浅层特征与深层特征融合时效果最好,因此网络采用第 7、15、16、17、18、



19层的融合特征。网络先将这六层的特征图分别重新调整为一维向量,再进行串联融合,实现多尺度人脸检测。

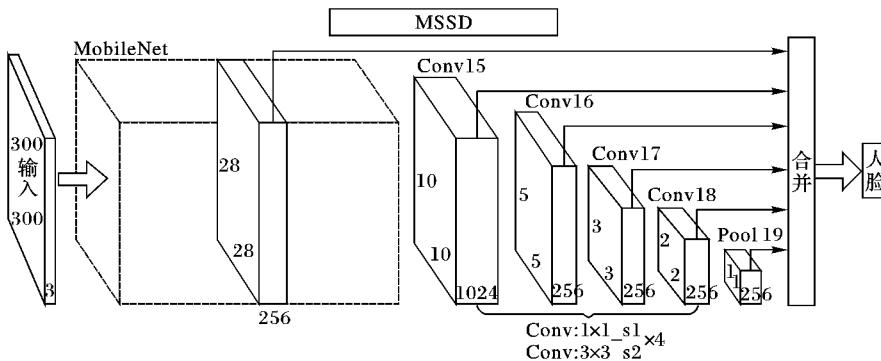


图2 MSSD 网络模型
Fig. 2 MSSD network model

2.2 结合跟踪模型的人脸检测

为了进一步地提高检测速度,将人脸检测网络和跟踪模型相结合,形成检测-跟踪-检测的模式。这样的结合方式不仅有效地提高了人脸检测的速度,还可处理多角度、有遮挡的人脸检测问题。跟踪模型是基于统计学习的跟踪算法 KCF,该算法主要使用了轮转矩阵对样本进行采集,然后使用快速傅里叶变换对其进行加速运算,这使得该算法的跟踪效果和速度都大大提升。先利用 MSSD 模型对人脸进行检测,并进行 KCF 跟踪模型更新;然后,将检测到的人脸坐标信息输入跟踪模型 KCF 中,以此作为人脸基础样本框并采用检测 1 帧跟踪 10 帧的策略来进行跟踪;最后,为了防止跟踪丢失,再次进行 MSSD 模型更新,重新对人脸进行检测。图 3 为结合跟踪的人脸检测流程。

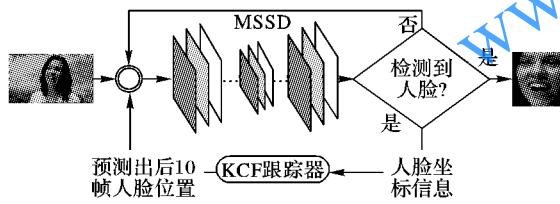


图3 结合跟踪的人脸检测流程
Fig. 3 Face detection flow chart combined with tracking

3 多尺度核特征人脸表情识别网络

3.1 深度可分离卷积

Howard 等^[24]在 2017 年提出 MobileNet,对标准卷积进行了分解,分为了深度卷积和点卷积两个部分,共同构成深度可分离卷积,标准卷积核与深度可分离卷积核的对比如图 4 所示。

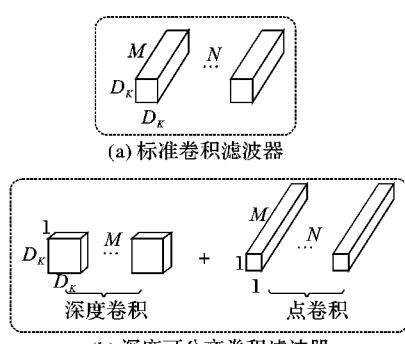


图4 两种卷积核对比
Fig. 4 Comparison of two convolution kernels

假设输入特征图尺寸为 $D_F \times D_F$,通道数为 M ,卷积核大小为 $D_K \times D_K$,卷积核个数为 N 。

对于同样的输入和输出,标准卷积过程计算量为: $D_K \times D_K \times M \times N \times D_F \times D_F$,深度可分离卷积过程计算量为: $D_K \times D_K \times 1 \times M \times D_F \times D_F + 1 \times 1 \times M \times N \times D_F \times D_F$ 。

通过以上可知深度可分离卷积方式与标准卷积方式的计算量比例为:

$$\begin{aligned} & (D_K \times D_K \times 1 \times M \times D_F \times D_F + \\ & 1 \times 1 \times M \times N \times D_F \times D_F) / \\ & (D_K \times D_K \times M \times N \times D_F \times D_F) = \\ & (1/N) + (1/D_K^2) \end{aligned} \quad (1)$$

对于卷积核大小为 3×3 的卷积过程,计算量可减少至原来 $1/9$ 。可见这样的结构使其极大地减少了计算量,有效提高了训练与识别的速度。

3.2 多尺度核卷积单元

多尺度核卷积单元主要以深度可分离卷积为基础,分支中采用了 MobileNetV2^[25]的线性瓶颈层结构并对其进行了改进,将其中的非线性激活函数改为 PReLU^[26],图 5 是改进的线性瓶颈层(bottleneck_p)结构。

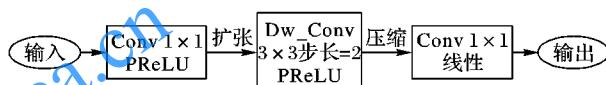


图5 改进的线性瓶颈层
Fig. 5 Improved linear bottleneck

深度卷积(图中为 Dw_Conv)作为特征提取部分,点卷积(图中为 Conv 1×1)作为瓶颈层进行通道数的缩放,并且输出端的点卷积采用的是线性结构,因为该处点卷积是用于通道数的压缩,若再进行非线性操作,则会损失大量有用特征。图 6 是多尺度核卷积单元结构图,它包含了三条分支,每个分支均采用步长为 2 的改进的线性瓶颈层结构。通过三个不同深度卷积核大小的分支并联形成的多尺度核卷积单元,融合了不同卷积核大小提取的多样性特征,进而有效地提高人脸识别的识别率。

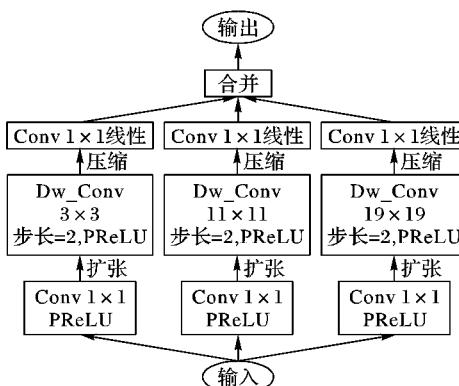


图6 多尺度核卷积单元
Fig. 6 Multi-scale kernel convolution unit

为了说明多尺度核特征的有效性以及卷积核大小的选取,用表 1 所示网络结构进行了 10 组对比实验。表 1 是在 FER-2013 上的多尺度核特征有效性评估结果。实验 1 是将多尺度核卷积单元改为核大小为 3×3 的标准卷积进行的实



验,实验2~6是将多尺度核卷积单元的三条支路均使用同一大小的卷积核进行的实验,实验7~10是改变多尺度核卷积单元三条支路的卷积核大小进行的实验。实验1~6表明网络使用适当卷积核大小的单一尺度核卷积单元比不使用的识别率更高;实验2~6表明具有单一尺度核卷积单元的网络使用 3×3 卷积核的效果比其他卷积核大小更好;实验2~10表明除了实验9的情况外,多尺度核卷积单元比单一尺度核卷积单元更有效,同时实验9的情况说明了多尺度核卷积单元的三个卷积核不能都取比较大的尺寸。

通过以上分析,多尺度核卷积单元的核大小选取了 3×3 、 11×11 、 19×19 三种最优尺度,使用多尺度核卷积比标准卷积的识别率提升了3.2%。

表1 FER-2013上的多尺度核特征有效性评估

Tab. 1 Effectiveness evaluation of multi-scale kernel feature on FER-2013

实验序号	类型	卷积核大小	识别率/%
1	标准卷积	3×3	69.8
2		3×3	70.9
3		7×7	70.7
4	单一尺度	11×11	70.7
5		15×15	69.5
6		19×19	69.4
7		$3 \times 3, 7 \times 7, 11 \times 11$	71.8
8	多尺度	$7 \times 7, 11 \times 11, 15 \times 15$	72.4
9		$11 \times 11, 15 \times 15, 19 \times 19$	70.5
10		$3 \times 3, 11 \times 11, 19 \times 19$	73.0

在多尺度核卷积单元中,除了用于压缩的点卷积不使用非线性激活函数外,其他卷积层均使用PReLU激活函数。式(2)、式(3)分别是激活函数ReLU^[27]和PReLU的表达式, i 表示不同通道。

$$\text{ReLU}(x_i) = \begin{cases} x_i, & x_i > 0 \\ 0, & x_i \leq 0 \end{cases} \quad (2)$$

$$\text{PReLU}(x_i) = \begin{cases} x_i, & x_i > 0 \\ a_i x_i, & x_i \leq 0 \end{cases} \quad (3)$$

ReLU激活函数是将所有负值都设为0,其余保持不变。当训练过程中有较大梯度经过ReLU时,会引起输入数据产生巨大变化,会出现大多数输入是负数的情况,这种情况下会导致神经元永久性失活,梯度永远为0,无法继续进行网络权重的更新。然而在PReLU中修正了数据的分布,使得一部分负值也能够得以保留,很好地解决了ReLU中存在的问题,并且式(3)中的参数 a_i 可以通过训练得到,能够根据数据的变化而变化,灵活性与适应性更强。

通过以上分析,将不同激活函数对多尺度核特征人脸表情识别效果进行了对比,表2是不同激活函数在FER-2013数据集上的识别率,可知使用PReLU比ReLU的识别率高1.8个百分点,因此选择PReLU作为激活函数。

表2 不同激活函数在FER-2013上的识别率

Tab. 2 Recognition rate of different activation functions on FER-2013

激活函数	识别率/%	激活函数	识别率/%
ReLU ^[27]	71.2	ELU ^[29]	72.5
LeakyReLU ^[28]	71.4	PReLU ^[26]	73.0
ReLU6 ^[25]	71.8		

3.3 多尺度核特征网络

用于人脸表情识别的多尺度核特征网络结构如表3所示。表中multi_conv2d、bottleneck_p(1~5)分别表示3.2节介绍的多尺度核卷积单元和改进的线性瓶颈层。网络的输入首先经过一个多尺度核卷积单元(multi_conv2d),采用6倍的扩张系数,每个分支采用16个卷积核进行卷积,输出通道数为16,步长为2,再将三分支特征进行融合,输出通道数变为48;然后经过12个改进的线性瓶颈层,每层的深度卷积核大小均使用 3×3 ,并且在训练期间进行数据的批量归一化;最后会通过一个卷积核大小为 1×1 、步长为1的标准卷积层和一个核大小为 3×3 的平均池化层;输出端的分类器设计采用了全卷积神经网络的分类策略,使用了步长为1、核大小为 1×1 、输出通道数为7(7类表情)的标准卷积层来替代全连接层,加快表情识别速度。

表3 多尺度核特征网络结构

Tab. 3 Network structure of multi-scale kernel feature

输入尺寸	操作	扩张系数	输出通道数	重复次数	步长
$48 \times 48 \times 1$	multi_conv2d	6	48	1	2
$24 \times 24 \times 48$	bottleneck_p1	1	32	1	1
$24 \times 24 \times 32$	bottleneck_p2	6	64	3	2
$12 \times 12 \times 64$	bottleneck_p3	6	96	4	2
$6 \times 6 \times 96$	bottleneck_p4	6	160	3	2
$3 \times 3 \times 160$	bottleneck_p5	6	320	1	1
$3 \times 3 \times 320$	conv2d 1 × 1		1280	1	1
$3 \times 3 \times 1280$	avg_pool 3 × 3		1280	1	
$1 \times 1 \times 1280$	conv2d 1 × 1		7	1	1
$1 \times 1 \times 7$	Reshape		7	1	

4 实验结果及分析

实验配置如下:

中央处理器(Central Processing Unit, CPU):Inter Core i7-7700K,主频为4.20 GHz,内存为16 GB;图像处理器(Graphic Processing Unit, GPU):GeForce GTX 1080Ti,显存为12 GB。

4.1 数据集

实验中用到了三种数据集:WIDER FACE^[30]、CK+^[13]和FER-2013^[14]。

WIDER FACE数据集为人脸检测基准数据集,共包含了32203张图像,并对393703个面部进行了标记,具有不同的尺寸、姿势、遮挡、表情、光照以及化妆的人脸。所有的图像被分为61类,每类随机选择40%作为训练集、10%作为验证集、50%作为测试集,即训练集12881张、验证集3220张、测试集16102张。

CK+人脸表情数据集包括123个人,593个图像序列,每个图像序列的最后一张都有动作单元标签,而其中327个图像序列有表情标签,被标注为七类表情标签:愤怒、鄙视、厌恶、恐惧、高兴、悲伤和惊讶。但是在其他的表情数据集中没有鄙视这类表情,为了和其他数据集能够相互兼容,因此去掉了鄙视这类表情。

FER-2013是Kaggle人脸表情识别挑战赛提供的一个人



脸表情数据集。该数据集总共包含 35 887 张表情图像,分为 7 类基本表情:愤怒、厌恶、恐惧、高兴、悲伤、惊讶和中性。FER2013 已被挑战赛举办方分为了三部分:训练集 28 709 张、公共测试集 3 589 张和私有测试集 3 589 张。在训练时将公共测试集作为验证集,私有测试集作为最终指标判断的测试集,该数据集包含了不同年龄、不同角度的人脸表情,并且分辨率也相对较低,很多图片还有手、头发和围巾等的遮挡,非常具有挑战性,很符合真实环境中的条件。

4.2 数据增强

为了增强人脸表情识别模型对噪声和角度变换等干扰的稳定性,对实验数据集进行了数据增强,对每张图像都使用了不同的线性变换方式进行增强,如图 7 所示。进行数据增强的变换有随机水平翻转、比例为 0.1 的水平和竖直方向偏移、比例为 0.1 的随机缩放、在 $(-10, 10)$ 进行随机转动角度、归一化为零均值和单位方差向量,并对变换过程中出现的空白区域按照最近像素点进行填充。



图 7 数据增强效果

Fig. 7 Data enhancement effect

4.3 人脸检测实验结果

对于结合跟踪的 MSSD 人脸检测网络,先将 MSSD 的基础网络 MobileNet 在 ImageNet^[31]1000 分类的大型图像数据库上进行预训练;然后再将预训练好的模型迁移到 MSSD 网络中,用人脸检测基准数据库 WIDER FACE 进行微调;最后用 WIDER FACE 的测试集进行测试。图 8 是测试集中部分图片检测结果,可知 MSSD 人脸检测网络对多尺寸、多角度和遮挡等均具有较好的检测效果,稳定性强。



图 8 WIDER FACE 测试结果示例

Fig. 8 WIDER FACE test result examples

在检测速度方面,使用大小为 640×480 的视频进行测试,取视频的前 3 000 帧来计算平均处理速度,并与主流的人脸检测网络模型进行了对比实验。表 4 是不同方法人脸检测速度对比结果。MSSD 网络人脸检测速度为 63 帧/s,再结合 KCF 跟踪器,速度可达 158 帧/s。多任务级联卷积神经网络 (MultiTask Cascaded Convolutional Neural Network, MTCNN) 是主流的人脸检测网络,本文方法的检测速度是它的 6.3 倍,优势非常明显。

表 4 不同方法人脸检测速度对比

Tab. 4 Comparison of face detection speeds by different methods

方法	平均速度/FPS	方法	平均速度/FPS
Faceness ^[32]	10	MSSD	63
MTCNN ^[33]	25	MSSD + KCF	158
SSD ^[21]	37		

4.4 人脸识别实验结果

人脸表情识别实验主要是在 FER-2013 和 CK+ 两个数据上进行训练和测试,在训练过程中均采用随机初始化权重和偏置,批量大小为 16,初始学习率为 0.01,并且采用了训练自动停止策略,即出现过拟合现象时,训练经过 20 个循环后自动停止并保存模型。

模型训练过程使用 FER-2013 的训练集(28 709 张)进行训练,公共测试集(3 589 张)作为验证集来调整模型的权重参数,最后用私有测试集(3 589 张)进行最后的测试。然后与目前先进的表情识别网络进行了对比。表 5 第一部分是不同方法在 FER-2013 上的识别率对比结果。可知本文方法优于其他主流方法,达到了 73.0% 的识别率,比 Kaggle 人脸表情识别挑战赛冠军 Tang^[16] 的识别率提高了 1.8 个百分点,同时识别速度达到了 154 帧/s。

在 CK+ 数据集上的实验采用了迁移学习方法,将模型在 FER-2013 上训练得到的权重参数作为预训练结果,然后在 CK+ 上进行微调,并采用 10 折交叉验证对模型性能进行评估。表 5 第二部分是不同方法在 CK+ 数据集上的识别率对比,本文方法取得了 99.5% 的最高识别率。

表 5 不同数据集上的识别率对比

Tab. 5 Comparison of recognition rate on different datasets

数据集	方法	识别率/%
FER-2013	MobileNetV2 ^[25]	69.9
	Jeon ^[19]	70.7
	InceptionV4 ^[34]	70.8
	Tang ^[16]	71.2
	Guo ^[35]	71.3
	Yan ^[36]	72.0
CK+	本文方法	73.0
	Fernandez ^[37]	90.3
	Song ^[38]	93.2
	MobileNetV2 ^[25]	98.3
	InceptionV4 ^[34]	98.8
	Zhang ^[39]	98.9
	Connie ^[17]	99.4
	本文方法	99.5

表 6 和表 7 分别是在 FER-2013 和 CK+ 两个数据集上的识别结果混淆矩阵。在数据集 FER-2013 中,高兴的识别率最高为 90.0%,其次是惊讶和厌恶,对恐惧和悲伤的识别率相对较低。从表 7 可看出造成这两者识别率较低的原因是这两类表情容易相互混淆。为了更直观地对这两类表情进行分析,图 9 给出了 FER-2013 中的恐惧和悲伤两类表情图像,可知在该数据集中恐惧和悲伤两类表情极易混淆,人工都很难进行准确判断。在数据集 CK+ 中,其数据集较小并且没有 FER-2013 中那么多的标签噪声,同时又全是清晰的正面表情



照片,因此本文方法在该数据集中除了厌恶之外的各类表情识别率均为100%,仅将厌恶表情中的3%识别为了愤怒,整体识别率高达99.5%。

表6 FER-2013识别结果混淆矩阵

Tab. 6 Confusion matrix of FER-2013 recognition result

真实表情 类别	预测表情类别						
	愤怒	厌恶	恐惧	高兴	悲伤	惊讶	中性
愤怒	0.66	0.02	0.10	0.03	0.11	0.02	0.06
厌恶	0.11	0.76	0.04	0.01	0.02	0.04	0.01
恐惧	0.08	0.00	0.63	0.02	0.13	0.08	0.05
高兴	0.01	0.00	0.01	0.90	0.02	0.02	0.03
悲伤	0.08	0.00	0.14	0.02	0.64	0.01	0.11
惊讶	0.02	0.00	0.08	0.04	0.02	0.82	0.03
中性	0.07	0.00	0.06	0.06	0.15	0.01	0.65

表7 CK+识别结果混淆矩阵

Tab. 7 Confusion matrix of CK+ recognition result

真实表情 类别	预测表情类别					
	愤怒	厌恶	恐惧	高兴	悲伤	惊讶
愤怒	1.00	0.00	0.00	0.00	0.00	0.00
厌恶	0.03	0.97	0.00	0.00	0.00	0.00
恐惧	0.00	0.00	1.00	0.00	0.00	0.00
高兴	0.00	0.00	0.00	1.00	0.00	0.00
悲伤	0.00	0.00	0.00	0.00	1.00	0.00
惊讶	0.00	0.00	0.00	0.00	0.00	1.00



(a) 恐惧表情示例



(b) 悲伤表情示例

图9 FER-2013中的易混表情对比

Fig. 9 Confusing expression contrast in FER-2013

5 结语

针对人脸表情识别的泛化能力不足、稳定性差以及速度难以达到实时性要求的问题,提出了一种基于多尺度核特征卷积神经网络的实时稳定人脸表情识别方法。用检测加跟踪的模式进行人脸检测,实现了158帧/s的快速稳定人脸检测,而且多尺度核特征表情识别网络在FER-2013和CK+数据集上分别达到了73.0%和99.5%的高识别率。整个系统采用轻量化网络结构,总体处理速度高达78帧/s。精度和速度都能满足实际需求。在后续的研究中,可以利用反卷积等方法可视化各层特征,结合高低层有效特征进一步提高网络的精度。另外,可以采用更加接近真实环境的表情数据集进行训练,并且增加疼痛之类的表情类别,使得理论研究能够与实际相结合,将该方法使用在医疗监护等的实际场景中。

参考文献 (References)

- [1] EKMAN P. Contacts across cultures in the face and emotion [J]. Journal of Personality and Social Psychology, 1971, 17(2): 124–129.
- [2] ZHAO X, ZHANG S. Facial expression recognition based on local binary patterns and kernel discriminant isomap [J]. Sensors, 2011, 11(10): 9573–9588.
- [3] KUMAR P, HAPPY S L, ROUTRAY A. A real-time robust facial expression recognition system using HOG features [C]// CAST 2016: Proceedings of the 2016 International Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 289–293.
- [4] 刘帅师, 田彦涛, 万川. 基于Gabor多方向特征融合与分块直方图的人脸表情识别方法[J]. 自动化学报, 2011, 37(12): 1455–1463. (LIU S S, TIAN Y T, WAN C. Facial expression recognition method based on gabor multi-orientation features fusion and block histogram [J]. Acta Automatica Sinica, 2011, 37(12): 1455–1463.)
- [5] BERRETTI S, del BIMBO A, PALA P, et al. A set of selected SIFT features for 3D facial expression recognition [C]// ICPR 2010: Proceedings of the 2010 20th International Conference on Pattern Recognition. Piscataway, NJ: IEEE, 2010: 4125–4128.
- [6] CHEON Y, KIM D. Natural facial expression recognition using differential-AAM and manifold learning [J]. Pattern Recognition, 2009, 42(7): 1340–1350.
- [7] 尹星云, 王洵, 董兰芳, 等. 用隐马尔可夫模型设计人脸表情识别系统[J]. 电子科技大学学报, 2003, 32(6): 725–728. (YIN X Y, WANG X, DONG L F, et al. Design of recognition for facial expression by hidden markov model [J]. Journal of University of Electronic Science and Technology of China, 2003, 32(6): 725–728.)
- [8] VAPNIK V N, LERNER A Y. Recognition of patterns with help of generalized portraits [J]. Avtomatika I Telemekhanika, 1963, 24(6): 774–780.
- [9] ROWEIS S T. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323–2326.
- [10] HART P E. The condensed nearest neighbor rule [J]. IEEE Transactions on Information Theory, 1968, 14(3): 515–516.
- [11] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]// NIPS '12: Proceedings of the 25th International Conference on Neural Information Processing Systems. North Miami Beach, FL, USA: Curran Associates, 2012: 1097–1105.
- [12] LYONS M J, AKAMATSU S, KAMACHI M G, et al. Coding facial expressions with Gabor wavelets [C]// AFGR 1998: Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition. Piscataway, NJ: IEEE, 1998: 200–205.
- [13] 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]// CVPRW 2010: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2010: 94–101.
- [14] GOODFELLOW I J, ERHAN D, CARRIER P L, et al. Challenges in representation learning: a report on three machine learning contests [J]. Neural Networks, 2013, 64: 59–63.
- [15] DHALL A, GOECKE R, LUCEY S, et al. Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark [C]// ICCVW 2011: Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops. Piscataway, NJ: IEEE, 2011: 2106–2112.
- [16] TANG Y. Deep learning using linear support vector machines [EB/OL]. [2018-12-21]. <https://arxiv.org/pdf/1306.0239.pdf>.
- [17] AL-SHABI M, CHEAH W P, CONNIE T. Facial expression recognition using a hybrid CNN-SIFT aggregator [EB/OL]. [2018-



- 08-17]. <https://arxiv.org/ftp/arxiv/papers/1608/1608.02833.pdf>.
- [18] FANG H, PARTHALÁIN N M, AUBREY A J, et al. Facial expression recognition in dynamic sequences: an integrated approach [J]. *Pattern Recognition*, 2014, 47(3): 1271–1281.
- [19] JEON J, PARK J-C, JO Y J, et al. A real-time facial expression recognizer using deep neural network [C]// IMCOM '16: Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication. New York: ACM, 2016: Article No. 94.
- [20] NEHAL O, NOHA A, FAYEZ W. Intelligent real-time facial expression recognition from video sequences based on hybrid feature tracking algorithms [J]. *International Journal of Advanced Computer Science and Applications*, 2017, 8(1): 245–260.
- [21] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector [C]// Proceedings of the 2016 European Conference on Computer Vision, LNCS 9905. Berlin: Springer, 2016: 21–37.
- [22] HENRIQUES J F, CASEIRO R, MARTINS, et al. High-speed tracking with kernelized correlation filters [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(3): 583–596.
- [23] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2019-01-10]. <https://arxiv.org/pdf/1409.1556.pdf>.
- [24] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications [EB/OL]. [2018-12-17]. <https://arxiv.org/pdf/1704.04861.pdf>.
- [25] SANDLER M, HOWARD A, ZHU M, et al. Inverted residuals and linear bottlenecks: mobile networks for classification, detection and segmentation [EB/OL]. [2018-12-16]. <https://arxiv.org/pdf/1801.04381v2.pdf>.
- [26] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification [EB/OL]. [2018-12-06]. <https://arxiv.org/pdf/1502.01852.pdf>.
- [27] JARRETT K, KAVUKCUOGLU K, RANZATO M, et al. What is the best multi-stage architecture for object recognition? [C]// ICCV 2009: Proceedings of the IEEE 12th International Conference on Computer Vision. Piscataway, NJ: IEEE, 2009: 2146–2153.
- [28] LIEW S S, KHALIL-HANI M, BAKHTERI R. Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems [J]. *Neurocomputing*, 2016, 216(C): 718–734.
- [29] DJORK-ARNÉ C, UNTERTHINER T, HOCHREITER S. Fast and accurate deep network learning by Exponential Linear Units (ELUs) [EB/OL]. [2019-01-22]. <https://arxiv.org/pdf/1511.07289.pdf>.
- [30] YANG S, LUO P, LOY C C, et al. WIDER FACE: a face detection benchmark [C]// CVPR 2016: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 5525–5533.
- [31] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database [C]// CVPR 2009: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248–255.
- [32] YANG S, LUO P, LOY C C, et al. From facial parts responses to face detection: a deep learning approach [C]// ICCV 2015: Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2015: 3676–3684.
- [33] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks [J]. *IEEE Signal Processing Letters*, 2016, 23(10): 1499–1503.
- [34] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning [C]// AAAI 2017: Proceedings of the 31st AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2017: 23–38.
- [35] GUO Y, TAO D, YU J, et al. Deep neural networks with relativity learning for facial expression recognition [C]// ICMEW 2016: Proceedings of the 2016 IEEE International Conference on Multimedia and Expo Workshops. Piscataway, NJ: IEEE, 2016: 1–6.
- [36] YAN J, ZHENG W, CUI Z, et al. A joint convolutional bidirectional LSTM framework for facial expression recognition [J]. *IEICE Transactions on Information and Systems*, 2018, 101(4): 1217–1220.
- [37] FERNANDEZ P D M, PEÑA F A G, REN T I, et al. FERAtt: facial expression recognition with attention net [EB/OL]. [2019-02-08]. <https://arxiv.org/pdf/1902.03284.pdf>.
- [38] SONG X, BAO H. Facial expression recognition based on video [C]// AIPR 2017: Proceedings of the 2016 IEEE Applied Imagery Pattern Recognition Workshop. Washington, DC: IEEE Computer Society, 2016, 1: 1–5.
- [39] 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.

This work is partially supported by the National Natural Science Foundation of China (61771411), the Sichuan Science and Technology Project (2019YJ0449), the Graduate Innovation Fund of Southwest University of Science and Technology (18yex123).

LI Minze, born in 1992, M. S. candidate. His research interests include deep learning, computer vision.

LI Xiaoxia, born in 1976, Ph. D., professor. Her research interests include pattern recognition, computer vision.

WANG Xueyuan, born in 1974, Ph. D., associate professor. His research interests include image processing.

SUN Wei, born in 1995, M. S. candidate. His research interests include image processing.