基于深度学习的人体行为识别综述

邓淼磊,高振东,李 磊,陈 斯
河南工业大学 信息科学与工程学院,郑州 450001

摘 要:人体行为识别旨在对视频监控中的人体行为进行检索并识别,是人工智能领域的研究热点。基于传统方法的人体行为识别算法存在对样本数据依赖大、易受环境噪声影响等不足。为解决此问题,许多适用于不同应用场景的基于深度学习的人体行为识别算法被提出。介绍了人体行为识别任务中传统特征提取方法和基于深度学习的特征提取方法;从性能和应用两方面对基于深度学习的人体行为识别算法进行总结,重点分析了基于3D卷积神经网络、混合网络、双流卷积神经网络和少样本学习(few-shot learning,FSL)的人体行为识别方法及其在UCF101和HMDB51数据集上的表现;在深度学习的基础上,归纳了主流模型迁移方法的优缺点及其有效性;总结了现有基于深度学习的人体行为识别算法存在的不足,并讨论了以元学习(meta-learning)和transformer为代表的FSL算法将成为未来模型主流算法的可能性,同时对未来基于深度学习的人体行为识别算法的发展方向进行展望。

关键词:行为识别;深度学习;双流卷积网络;少样本学习;元学习
文献标志码:A **中图分类号:**TP18; TP391.41 doi:10.3778/j.issn.1002-8331.2201-0096

Overview of Human Behavior Recognition Based on Deep Learning

DENG Miaolei, GAO Zhendong, LI Lei, CHEN Si
College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China

Abstract: Human behavior recognition aims to retrieve and identify the target behavior in surveillance video, which is a research hotspot in the field of artificial intelligence. Human behavior recognition algorithm based on traditional methods has some shortcomings, such as large dependence on sample data and easy to be affected by environmental noise. In order to solve this problem, many human behavior recognition algorithms based on deep learning are proposed for different application scenarios. Firstly, the traditional feature extraction methods and feature extraction methods based on deep learning in human behavior recognition task are introduced. Secondly, the human behavior recognition algorithms based on deep learning are summarized from two aspects of performance and application. The idea of human behavior recognition method based on 3D convolutional neural network, hybrid network, two-stream convolutional neural network and few-shot learning(FSL) and its performance on UCF101 and HMDB51 datasets are analyzed. Thirdly, on the basis of deep learning, the advantages, disadvantages and effectiveness of mainstream model migration methods are summarized. Finally, the shortcomings of existing human behavior recognition algorithms based on deep learning are summarized, and the possibility of FSL algorithm represented by meta-learning and transformer that will become the mainstream algorithm of future models is discussed. At the same time, the future development direction of human behavior recognition based on deep learning is prospected.

Key words: behavior recognition; deep learning; two-stream convolution network; few-shot learning(FSL); meta-learning

随着计算机技术的快速发展,基于视频智能分析的人体行为识别技术得到了广泛的应用。在公共安全领域[1],通过行为识别技术可以检测打架斗殴、持刀抢劫等违反治安法的暴力行为,尽可能减少因此所造成的人员伤害和财产损失;在智慧交通领域[2-3],通过行为识别技术可以自动判别如行人/车辆闯红灯、驾驶员不安全驾驶等交通违法行为,保障人们出行安全;在医疗监护领域[4-5],通过该技术可实现对患者的实时监控和意外跌倒

基金项目:河南省重大公益专项(201300311200)。
作者简介:邓淼磊(1977—),男,博士,教授,硕士生导师,CCF会员,研究方向为信息安全、物联网技术;高振东(1995—),男,硕士研究生,CCF会员,研究方向为人体行为识别;李磊(1979—),男,博士,副教授,CCF会员,研究方向为图像处理、计算机视觉与机器学习,E-mail:leili@haut.edu.cn;陈斯(1995—),男,硕士研究生,CCF会员,研究方向为机器学习。
收稿日期:2022-01-07 **修回日期:**2022-03-09 **文章编号:**1002-8331(2022)13-0014-13

检测等,确保患者能够得到及时治疗和帮助;在安全生产领域[6],可以实现对生产作业全过程的实时监测,对作业生产过程中出现的可能导致安全隐患的行为及时报警,确保作业生产在安全可控范围内进行,保障人员的人身安全和财产安全。由此可见,基于视频分析的人体行为识别技术与人们的生产生活安全息息相关,对其的研究具有深远的意义。

基于视频分析的行为识别任务需建立动作、姿态样本库,并对所设计模型进行训练,以实现对视频中行为的分类。根据特征提取方式的不同,行为识别可划分为传统方法和深度学习方法。传统方法依赖手工对特征提取,由于早期样本库数据量小,场景简单,动作单一,传统方法可以满足一定的需求。但随着视频监控技术的普及,应用场景变得越来越复杂,使用传统方式提取的视频特征在识别准确度上已无法满足实际需求,视频监控的实际价值难以得到充分利用。卷积神经网络(convolutional neural network,CNN)[7]的出现很好地解决了传统方式的局限性,取得了较好的效果。本文通过对行为识别领域所使用技术如3D卷积网络、双流卷积网络、few-shot learning(FSL)等进行分析和总结,对其未来发展进行展望,旨在为行为识别研究提供技术和理论支撑。

1 特征提取方法

特征提取是对待检测对象进行关键信息提取,用来特征表示的过程,特征提取结果的好坏直接影响算法识别率和检测速度。特征提取方法可分为传统方法和深度学习,其过程对比如图1所示。

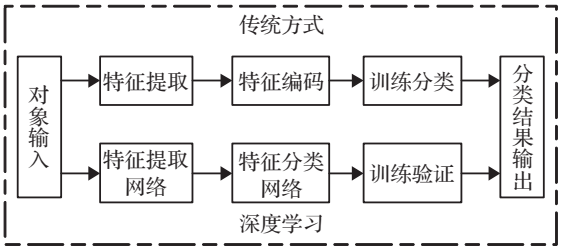

图1 传统和深度学习特征提取过程对比

Fig.1 Comparison of traditional and deep learning feature extraction process

1.1 传统特征提取方式

传统手工特征提取方式可分为两大类:

(1)全局特征提取,即对待测对象进行一次全局信息提取,因此该特征不包含任何的空间征,同时全局特征受噪声区域、视角变化影响较大。其中轮廓剪影(human silhouette)[8-9]、人体关节点(human joint point)[10]最具代表性。

(2)局部特征提取,即对待测对象进行多次局部信息提取,最后将多个特征进行融合。因此局部特征受视角变化、背景噪声影响较小。其中最具代表性方法包括运动轨迹(trajectories)[11]、时空兴趣点采样(space-time interest points)[12]等。

1.2 深度学习提取方式

传统特征提取方式是基于先验模型,并通过关键点提取,生成描述子特征数据。与之不同,基于深度学习的特征提取方式则是利用深度神经网络对待测对象进行深度特征表示的过程,通过对网络模型进行训练,获得网络参数,其具有对样本库数据量依赖小,同时受噪声、角度变化影响小等优势,已成为行为识别领域特征提取的主流方式。虽然深度学习算法在特征提取方面表现优异,但在对行为时空特征提取时也存在部分局限性,为此专家学者进行了深入研究,对于不同问题,提出了针对性的改进措施。针对动态时序特征提取难的问题,研究者们一般使用时间注意力机制使模型聚焦时序特征。例如,文献[13]将预训练的VGG16网络转换为全卷积网络,并用第三池化层提取静态空间特征,同时引入时间注意力机制,学习历史特征和当前特征之间的关系。

针对特征提取过程中受噪声影响导致的特征提取难度大、时序特征信息利用率低的问题,文献[14]提出的基于时间序列膨胀的TS-I3D(time sequential inflated 3 dimensions)模型,利用调频连续波(frequency modulated continuous wave,FMCW)获取视频中每帧的行为信息,同时使用二维快速傅里叶变换计算每个行为的时间域和空间域特征值,然后基于运动参数和频率之间关系得到多普勒图,最后使用小波变换滤除噪声信息,得到最终的特征信息。

为了更好地识别视频中的人体行为,国内外研究学者一直致力于提取可以耦合人体行为的空间外观信息和时间运动信息。研究发现,将原始时空网络拆分为空间卷积和时间卷积两部分,可以更加有效提取行为特征。为解决视频序列对齐问题,文献[15]使用分解的时空卷积网络(factorized spatio-temporal convolutional network,FSTCN)来对异常行为特征进行提取,将给定的视频序列进行多剪辑采样,进一步提高视频监控领域中人体行为识别算法的精确度。此外,文献[16]基于时间域持续灰度值假设以及空间域梯度连续假设,提出的双通道卷积神经网络可以模拟人体大脑视觉神经腹侧和背侧通道,分别提取人体静态空间特征信息和动态运动特征信息。当真贱每个像素均发生改变时,这种利用双通道卷积神经网络提取人体行为特征信息的灰度值和空间梯度保持不变。在对人体行为识别时,考虑到空间连续性和人体行为运动状态的约束,文献[17]利用原始深度数据提取空间深度特征,同时使用帧分化来跟踪人体动作,提取时间深度特征,并将两部分进行融合,以提高模型分类的性能。

人体骨骼特征具有良好的时空性,有利于提高异常行为的判别速度,通过对人体骨骼特征提取,可以更好的获取具有辨识性的时空联合信息。文献[18]通过标注每个视频帧中的骨骼关节点,并将其在量化后按照时间顺序归一化处理,提取出了具有辨识性的空间特征及时间特征信息。目前大部分基于骨骼关节点的特征提取方法,存在参数量大、计算复杂的问题,这促使科研人员深入研究轻量级卷积骨骼特征提取方法,以提高模型的时效性。文献[19]集合多模态融合以及人体骨骼行为识别方法,通过多流信息融合,兼顾了模型识别准确度和时效性。为了有效提取更具有鉴别能力的时空特征,通过结合图卷积LSTM网络和骨骼关节点特征提取方法,不仅能够提取空间信息和时间信息,同时可以进一步表征空间域和时间域的共性关系。文献[20]将增强图卷积LSTM网络(AGC-LSTM)用于骨骼行为识别,提高了对于高层语义的表示能力,同时模型利用时间层次结构增强顶层时间域特征。此外,将混合卷积网络用于行为特征的提取同样取得了一定的成果。例如,文献[21]基于C3D神经网络和支持向量机提出了一种新颖的自适应特征提取方法,有别于常规算法,该方法根据RGB直方图检测场景的变化来测量样本中特征信息的冗余程度,同时基于光流分析提取视频中的关键帧作为C3D网络的输入,进而提取视频中的特征信息。虽然光流特征是基于人体行为识别任务最有效的特征之一,但特征提取过程计算量大,难以在现实中应用。为此,文献[22]基于I3D-Shufflenet的特征提取方法结合了I3D网络和轻量级模型ShuffleNet的低复杂度的优点,该方法有效地促进了特征的利用,降低了在特征提取过程中时间的消耗。

2　深度学习算法在行为识别领域的应用

深度学习作为行为识别领域主流技术,具有特征表示效果好、数据拟合能力强等的特点。常用的深度学习技术包括基于3D卷积神经网络模型、基于双流卷积神经网络模型以及混合网络模型等。

2.1　传统CNN网络

近年来CNN网络在图像分类和目标检测中发展迅速,如表1对常用CNN架构进行了汇总[23-31]。

受CNN在图像领域取得诸多成果的启发,学者们开始将CNN运用到人体行为识别领域。例如文献[32]通过改进轨迹计算的局部特征描述子捕获运动信息,同时使用8层卷积网络中后3层全连接层作为输出,并对输出向量取平均值后输入SVM中进行分类。文献[33]针对人体行为的运动特征和外观特征,分别利用改进稠密轨迹Fisher矢量表示和CNN激活。

2.2　3D卷积神经网络

诸多研究证明,卷积神经网络在从静态图像中提取空间信息具有较高的性能,然而人体行为识别是基于视频的3D时空信号,因此将CNN扩展并应用到3D视频的核心是引入对时间信息的利用。3D卷积神经网络是2D卷积神经网络的延伸,文献[34]首次在空间维度引入时间维度特征,以此获得视频中相邻帧间的上下文关系,并将其命名为3D CNN。3D卷积网络[35-40]在人工智能领域表现优异,国内外学者逐渐将其应用于行为识别,并取得了较好的效果,起初人们并没有考虑庞大的数据量对设备算力的依赖,产生了较大的计算成本;同时在实际应用过程中,直接通过3D卷积神经网络提取行为特征,也易引起梯度消失、梯度爆炸和过拟合问题。针对这些问题国内外学者展开了深入的研究。

表1　不同CNN架构对比分析

Table 1　Comparative analysis of different CNN architectures

架构	年份	相关工作	特点
AlexNet[23]	2012	1.增加Relu激活函数 2.Dropout层与数据增强	1.增强非线性表达 2.防止过拟合
ZFNet[24]	2013	引入可视化	内部操作可视化,有利于网络训练
VGG[25]	2014	用连续3×3卷积代替7×7等较大卷积	通过加深网络结构提高网络性能
GoogleNet[26]	2014	1引入1×1卷积 2.平均池化代替全连接	1.压缩降维,减少参数量 2.防止过拟合
ResnNet[27]	2015	1.核心单元模块化 2.采用Shortcut结构 3.引入BN层	1.简化网络结构,可以简单堆叠 2.解决梯度消失问题,提高网络性能 3.加速网络训练,使收敛更稳定
ShuffleNet[28]	2015	提出逐点群卷积和channel shuffle	减少冗余连接,提高模吧·B 型表达能力和训练速度
MobileNet V1[29]	2017	1.提出深度可分离卷积 2.引入超参数宽度因子与分辨率因子	1.减少了参数量和计算量 2.数据大小和维度、网络计算量和参数量可调
MobileNet V2[30]	2018	引入了反向残差结构与线性瓶颈结构(linear bottlenecks)	减少了参数和计算量,解决了梯度消失问题
MobileNet V3[31]	2019	1.使用5×5卷积代替3×3卷积 2.改进SE模块 3.提出H-Swish激活函数	1.加快网络收敛速度 2.放大有效权重,减小无效权重,提高网络性能 3.替代ReLU激活,提高模型精度

梯度消失问题即在反向传播过程中,梯度信息以指数形势减少,最后趋近于0,进而导致模型权重难以更新,使训练不能收敛到较好结果,模型丧失学习能力;相反,梯度爆炸问题是在反向传播过程中,梯度信息呈现指数形势增长趋势,导致误差梯度不断累加,使得模型权重更新过量,造成模型无法有效学习。本质上讲,梯度消失和梯度爆炸都是由于网络层数太深导致的反向传播过程中梯度信息的连乘效应。一般而言,适当降低模型复杂度,更换RuLE、ELU等激活函数,引入残差结构等方法可以使网络模型具有较好的收敛结果,达到提高模型泛化能力和识别效果的目的。如文献[41]通过跳过连接层和设置学习率的方法,解决分离操作引起的梯度消失问题。该方法通过在不影响识别率的同时,降低了模型复杂度,更加有利于部署。但模型复杂度的降低不利于网络有效提取特征信息,一定程度上导致了模型识别准确率的降低。

在3D卷积神经网络中,较多使用伪-3D残差网络(pseudo-3D residual network,P3D ResNet)来对网络模型进行优化。简而言之,P3D ResNet是将3×3×3的3D卷积滤波器分离为1个1×3×3空间卷积滤波器和1个3×1×1时间卷积滤波器。文献[42]在3D卷积神经网络的基础上,引入了P3D ResNet,该方法提高了3D卷积神经网络在人体行为识别领域的泛化能力。相似地,文献[43]用P3D残差网络替换3×3×3卷积运算,同时使用openpose预处理工具提取样本特征,进一步提高P3D Resnet的准确性。实现了自助银行场景下人体行为的有效识别,保证了人们的财产安全。虽然P3D残差网络可以使模型减少运算量,一定程度上解决了梯度消失和梯度爆炸问题,但在实际应用过程中,由于其使用的1×1卷积滤波器会造成特征信息部分丢失,同时也无法对多尺度特征信息有效融合,进而导致模型识别精度下降。例如文献[43]中自助银行场景下的行为识别模型在UCF101数据集其识别精度小于90%,在实际应用中更是不足70%。为此,文献[44]在其基础上,通过在残差网络中嵌套残差网络的方式,不仅解决了梯度消失问题,同时进一步提高了残差网络的性能,该方法对于人员跌倒的识别率高达97.4%,但是其未考虑多尺度视频特征对算法查准率的影响。文献[45]引入多尺度特征融合思想,通过组合特征映射来增强不同网络层的特征提取能力,以此提高行为的查准率。

过拟合是由于模型拟合参数时,由于样本数据的采样误差拟合进模型参数中导致的。通常产生过拟合问题的原因包括,训练数据不足,数据有较多噪声,模型过于复杂等。在实际应用过程中表现为模型的泛化能力不足,在训练集上表现良好,但在测试集上表现较差。通常可以通过对数据集扩容、随机抑制池化单元等方法来解决。文献[46]通过将Dropout引入到3D CNN网络池化层中,随机抑制池化单元,使池化层生成更多的子模型,同时引入双概率加权思想,使被选中神经元的激活值与其权重相结合,得到所有子模型的近似模型。虽然该方法解决了过拟合问题,增强了模型的泛化能力和识别精度,但随机池化过程是不可控的,其在抑制无效特征信息的同时,也会造成部分有效特征信息失活。

随着国内外学者对人体行为识别领域的深入研究,基于姿态信息和视频帧信息的3D卷积方法在各公开数据集上实现了较高的识别精度,但人体行为识别不仅需要考虑行为的类别,还需全面考虑环境信息对模型的影响。文献[47]提出基于P3D卷积网络和LSTM的行为识别方法,在一定程度上提高了模型的鲁棒性,经验证,其在识别准确率和速率方面均优于普通的3D卷积网络。P3D卷积方法使用预训练的P3D网络以及LSTM分别提取环境特征和行为特征,结合了两部分特征信息完成对行为的分类,此外,通过添加回归网络提高算法的识别速率,但其过于追求对于空间信息的提取,牺牲了大量的时间信息。针对该问题,研究者提出了两种解决方法,一种单纯提高算法对时间信息的利用率,如文献[48]中将提出的的轻量级C3D网络用于完成对雨天驾驶行为的预测。该方法使用五个3D卷积核和两个FC层中的最后一层FC层来学习最终帧的特征信息,通过利用更多有效的时间信息,提高模型的识别精度。另一种通过结合时间域和空间域特征信息,进而更好地研究人体行为的全局特征,如文献[49]在时间域上,考虑行为本身的相关性,在空间域上,以运动映射序列作为时空卷积网络的输入。该方法从不同角度充分识别和利用行为时空特征,同时一定程度上解决了3D卷积网络中存在的时间失配和噪声影响问题。

3D卷积神经网络注重空间运动信息,计算速度快,在人体行为异常识别领域具有较为广泛的应用,但不能否认的是,其具有的良好空间特征提取能力,是以牺牲时间特征为代价换来的,而且虽然计算速度和识别精度表现良好,但巨大的计算开销,一直是难以解决的问题。

2.3 混合深度学习网络

混合深度学习网络可根据不同的应用场景特点组合相应的模型,由于其集成多种网络模型的优点,在众多应用场景中均具有强大的特征提取能力和良好的识别精度[50-52]。目前主流的混合深度学习网络以3D卷积神经网络为基础,融合其他深度学习方法,以提高模型的泛化能力和识别精度。例如,文献[53]使用3DCNN模型和LSTM算法分别对光流信息和辅助信息进行特征提取并加以巩固,最后采用支持向量机对行为进行分类,实验证明其在多个数据集上识别精度均在90%以上。又如,文献[54]通过融合挤压激励网络(squeeze-and-excitation network,SE Net)、I3D、GRU网络的优点提出了SE-I3D-GRU模型,通过将三维SE Net分别引入至

I3D和GRU网络中,得到SE-I3D和SE-GRU网络,最后将二者进行合并,其在UCF101数据集上获得了93.2%的识别精度。此外,为充分挖掘CNN网络的潜力,文献[55]充分发掘LSTM网络与CNN网络的互补性,通过模型特征关系融合表示替代卷积融合策略,同时利用视频语义背景提高模型的行为预测性能,其在UCF101数据集上测试精度高达93.1%,在CCV数据集上其精度也达到了70%。同样的,文献[56]使用Shearlet变换提取多尺度特征信息,并通过循环神经网络对特征信息进行分类,实验证明使用Shearlet变换单独对特征提取过程建模,提高了行为分类信息的质量。RNN在行为识别应用中具有不同类型隐藏单元,为了降低RNN和隐藏单元的参数量,同时提高模型的分辨率,文献[57]结合高斯混合模型(GMM)和卡尔曼滤波(KF)提取人体行为特征信息,并使用门控循环网络(GRNN)对行为特征进行分类,该方法UCF101数据集上平均识别精度高达96.3%。

众多研究证明,混合算法在相应数据集上识别效果优异,但其组合困难,参数过多,资源消耗大,难以在现实中部署。因此,为了满足实际应用的需求,如何在不增加网络复杂度的同时,更高效地结合不同种类网络模型优点,以减少模型在实际应用中的资源消耗,仍需要进一步深入研究。

2.4　双流卷积神经网络

双流卷积神经网络通过提取人体静态表观特征和动态运动特征,从空间和时间两方面对特征进行深度分析,在人体识别领域具有良好的识别效果[58]。其网络架构如图2所示。

双流架构[59-64]是目前人体行为识别领域基准之一,国内外学者在其基础上对双流架构进行了进一步探索。最早的双流卷积模型以VGG-16卷积网络为基础,并在网络中分别针对时间通道和空间通道增加残差块,用以增强网络对于时间特征和空间特征的提取能力,最后将两通道特征信息加以融合利用。例如,文献[65]通过在VGG-16卷积网络中增加VMHI(VGG-16 and motion history image)和FRGB(faster R-CNN and RGB frames)两个模块组成的双流卷积神经网络,其中VMHI模块提取时间运动信息并将其输入至VGG-16卷积网络中,FRGB模块将RGB图像输入至R-CNN网络中训练,然后将两模块的输出进行融合,实验结果表明该方法不仅可以识别单人行为,亦可识别两人交互行为。虽然基于双流卷积神经网络对于人体行为时空特征提取效果较好,但仍存在对提取的时空特征难以有效利用的问题。针对该问题,国内外研究学者们在基于双流卷积网络的基础上提出了诸多的改进策略。优化网络结构是常见的改进方式之一,一方面针对VGG16网络无法充分提取行为特征信息的问题,将VGG16替换成更深层结构的网络,使模型具有提取更高维特征信息的能力。例如文献[66]验证了用ResNet101网络替换VGG16网络的可行性,文献[67]验证了用ResNet50替换VGG16网络的可行性。深层网络结构固然可以提取更高维的特征信息,但也易产生过拟合问题,使得模型的泛化能力下降。针对该问题,可以通过引入Dropout方法抑制部分深度神经网络提取的特征参数,以避免巨大参数量使模型过度拟合训练。文献[68]在特征映射中引入Dropout方法以降低网络中间层中相关的Rademacher复杂度。虽然Dropout方法可以有效减少模型参数量,解决过拟合训练问题,但是由于Dropout方法对于特征信息的抑制是随机的,其在抑制噪声信息的同时,也会导致部分有效特征失活。为此,文献[69]充分考虑每个神经元中信息,针对性的增加Dropout层。这样不仅可以摒弃高粘度神经元带来的噪声问题,保证损失最小化,还可以充分发挥Dropout方法的作用,有效地解决训练过程中出现的过拟合问题。

另一方面通过在残差网络中增加注意力模块对深层网络模型进行优化,目前大多数注意力模型以Encoder-Decoder框架为基础。Encoder-Decoder框架可以理解为将给定序列 X 首先通过编码方式转化为固定长度的向量,然后再将其解码为目标输出序列 Y 的过程,Encoder-Decoder框架如图3所示。

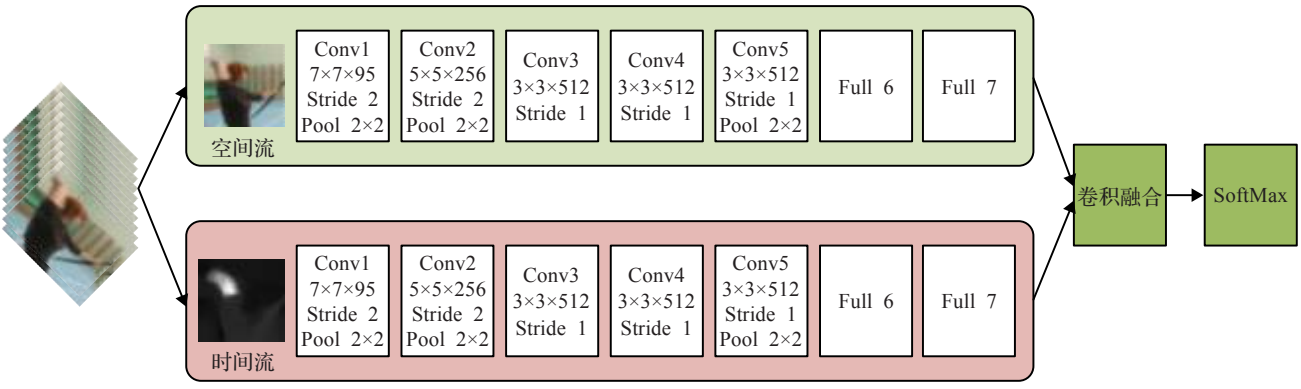

图2　双流卷积神经网络结构

Fig.2　Two-stream convolutional neural network frame

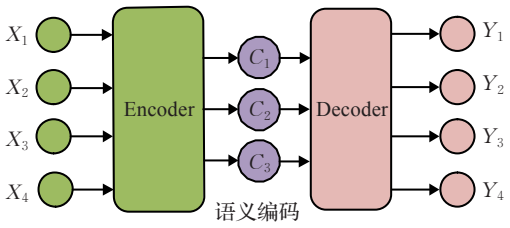

图3 Encoder-Decoder框架

Fig.3 Encoder-Decoder frame

Encoder-Decoder框架的提出,为构建可选择性提取特征信息的网络模型奠定了基础。文献[70]利用视觉注意力机制,构建了基于端到端的双流注意力LSTM模型,该模型可以根据需求,选择性提取光流图像中的有效特征,并根据深度特征相关层信息调整网络参数,用以模型的优化。为了更好地融合人体行为时间域特征信息和空间域特征信息,实现对视频图像中多通道的信息有效的利用,人们提出了可用于不同通道特征提取的注意力机制。例如,文献[71]中提出的跨模态的注意力模块(cross modality attention,CMA),通过有效融合视频中多模态信息,提高模型性能。文献[72]从空间注意力和时间注意力的互补性和共存关系入手,提出了双流协作学习的时空注意力模型(TCLSTA),通过静态空间特征和动态运动特征相互促进,增强模型特征学习能力。

经过众多学者的努力,基于双流卷积神经网络模型研究愈发成熟,在人体行为识别领域也取得了诸多成果。虽然双流网络能够很好的结合人体行为静态和动态特征信息,具有稳定性强、识别精度高的特点,但不可否认的是其具有的高性能是建立在大量数据样本训练的基础之上的,在实际应用中,很多场景均由于无法采集到足够的样本信息进行训练,这会使双流卷积神经网络在训练过程中出现过拟合等的问题,这也导致了实际中其无法达到理论上的识别精度。

2.5 Few-shot learning(FSL)

基于深度学习的异常行为识别模型具有高性能、高精度的特点,但训练深度学习模型需要大量的样本数据支撑,在许多实际场景中,收集大量样本并进行标注是非常困难的,这就导致很难将深度血虚模型应用于解决现实场景中的问题。如何使用少量样本、甚至无标签样本训练深度学习模型并使其达到具有大量标签样本支撑的模型的识别精度,更好地适应现实场景,不仅是人体行为识别领域的研究难题,更是整个深度学习应用领域亟需解决的问题。少样本学习方法(FSL)[73-76]旨在使用少量样本训练,使模型在有限的样本中快速适应新的类别。通过集成类别分类器和二域判别器,利用有限的标签信息进行模型训练,然后加以利用多任务生成对抗训练,可以实现网络的优化。文献[77]通过少样本对抗域适应方法(few-shot adversarial domain adaptation, FS-ADA)识别人体行为,使得网络可以在新场景下提取人体区域不变特征和类别判别特征。目前,大多数行为识别算法都遵循度量学习范式,即学习比较视频间相似性,但不同视频下相同行为在时间分布上存在类间差异,这会导致测试集和训练集在匹配时出现严重偏差。文献[78]从行为持续时间错位(action duration misalignment, ADM)与行为演化错位(action evolution misalignment,AEM)两个方面展开研究,首先通过时间转换模块(temporal transform module,TTM)学习时间仿射变换定位动作,忽略无关特征的同时,将每个视频特征时间偏差进行量化,其次将测试特征坐标化,通过时间重排和空间偏移预测匹配支持的时空行为演化。虽然传统FSL在许多场景中均具有良好的表现,但将其用于人体互动行为场景的识别时,由于人体行为的多样性和交互性,自适应分类器难以捕获表现模糊的特征信息,使得模型识别精度下降。针对该问题,可以通过重建视觉特征之间的关系来学习其类别间的潜在表示。文献[79]构建的动态图形网络(dynamic graph in graph network,Dgig-Net),将视觉子图嵌入到面向多任务的跨模型图中,提高了人体互动行为场景下的小样本识别模型的精度。

最近的研究表明,基于元学习和transformer的方法,可以很好地实现FSL。如果把机器学习看作寻找(X, Y)之间映射关系f的过程(如图4(a)所示),那么元学习可以理解为通过寻找(X, Y)的若干子集$((X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n))$的映射关系$(f_1, f_2, \cdots, f_n)$拟合为$F$,并将其推广至新的训练$(X^*, Y^*)$中,其过程可用图4(b)表示。

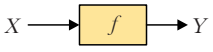

图4(a) 机器学习过程

Fig.4(a) Machine learning process

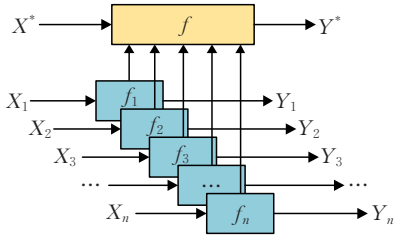

图4(b) 元学习过程

Fig.4(b) Meta learning process

基于元学习的少样本识别一般通过multi-head知识蒸馏方式以“端到端”的方式使网络具备自学习的能力。文献[80]通过multi-head知识方式,同时引入一般几何变换集的平移同变性和平移不变性思想,分别执行强制同变性和强制不变性操作,为了消除位置偏差,同时利用自监督学习方式,将仿射变换空间进行量化来优化模型,提高行为识别的精度。文献[81]通过自相关表

示(self correlational representation,SCR)和交叉相关注意力(cross correlational attention,CCA)模块寻找图像类间和类内的关系映射,通过在网络中嵌入这两个模块,以端到端的方式使网络具备“学会学习”的能力。目前大多数基于元学习的方法均是分别解析候选区域和新类之间的关系,而没有考虑二者之间的多重关系。文献[82]提出了基于异构卷积网络的FSL模型,该模型通过在候选区域节点和类节点之间进行高效消息传递,获得每个行为类的上下文感知特征,进而提高FSL模型的查准率。此外,通过将深度引导的自适应网络以元学习的方式训练同样可以达到少样本学习的目的。例如文献[83]将时间异步增强采样后的自适应元融合网络以元学习的方式训练,不仅缓解了模型样本数据短缺的问题,同时自适应地将特征熔断为两种不同的流,以此增强网络对特征信息的充分了利用。

Transformer[84]在自然语言处理领域中占主导地位,其具有独特的自注意力机制(self-attention)与位置编码(position encoding)机制,同时,具有可并行计算、无长距离依赖、输入参数共享等的特点。近几年,将transformer应用到计算机视觉领域同样取得了不俗的成功。文献[85]利用transformer架构具有的编码器-解码器结构优化共性学习和行为时空定位,实现了无类标签、间隔边界和框注释下的高精度行为识别。文献[86]利用迭代损失、自注意力机制以增强编码器-解码器结构中低层参数的渐变更新。文献[87]为降低模型计算量,在基于短时序的2D骨架模型中引入transformer,为模型的高准确、低延迟提出了一种新异常行为的解决方案。transformer的应用极大限度地解决了由于样本数据采集困难带来的诸如过拟合、欠拟合模型收敛速度慢、泛化能力低的问题。

2.6 其他算法

除3D卷积神经网络、双流卷积神经网络、少样本学习外,基于区域感兴趣点(regions of interest,ROI)的人体行为识别,通过降低数据规模,提高数据处理效率,文献[88]从人脸的区域感兴趣点提取Gabor特征,验证了基于区域感兴趣点的人体行为识别算法的可行性。人体行为识别的准确性与人体姿态具有较强的相关性,基于多流卷积神经网络可以通过融合人体姿态特征和其他特征,用以增强视频图像中人体行为的整体特征。融合姿态的多流卷积神经网络一般包括以下几个步骤:首先通过预训练的2D CNN网络将3D骨架序列转换成姿态进化图像(pose evolution images,PEI),其次,利用3D CNN网络从RGB视频中提取时间特征信息以及空间特征信息,然后,从RGB视频中提取人体ROI,最后将ROI输入至3D CNN网络中,并将四步的结果进行融合,得到最终行为的分类结果。文献[89]利用融合姿态的多流卷积神经网络实现了高精度的人体日常行为识别。文献[90]融合姿态数据和行为数据,并通过消融实验验证该方法与单纯的RGB图像、姿态数据相比更不易受外界环境干扰。在实际应用中,人体行为识别算法建立样本数据往往需要耗费极大的人力物力。基于图像聚类的方法可以自动生成数据标签,解决现实中行为数据标注难的问题。文献[91]提出了一种基于堆叠卷积auto encoder(SCAE)的聚类方法,该方法使用卷积神经网络(CNN)和长短期记忆网络(LSTM)构建深层网络,以研究多维框架中的人体行为识别视频的长期依赖性。除基于视觉的人体行为识别算法外,基于环境感知和可穿戴计算的人体行为识别方法也逐渐流行。例如,对于夜间人体行为的识别,由于受到光照条件的影响,往往使用红外相机获取人体行为信息。文献[92]通过捕获热量图像对夜晚行人不安全行为进行识别,其在传统CNN网络的基础上设计出了轻量级的CNN网络,同时将CNN网络与增强的随机森林分类器随机连接,以此来提高模型的识别精度。可穿戴计算的人体行为识别方法通过可穿戴传感器采集活动特征实现行为的识别,虽然相较于基于视觉的主流算法,不易受光线、遮挡等外界因素干扰,但由于缺乏相关技术规范,导致其推广速度缓慢。文献[93]基于CNN网络和滑动窗口构建了人体活动识别网络,该网络规范了应用中传感器的部署和数据的归一化方法,为可穿戴计算的人体行为识别技术规范的建立提供了参考。在行为识别过程中,如何在冗长、杂乱的视频中进行行为时空定位始终是难以解决的问题,为此,有学者提出了基于时间相似矩阵(temporal similarity matrices,TSM)的网络,该网络可在视频对之间进行微粒相似度建模,使它们之间生成时间类激活映射(temporal class activation maps,TCAM)。文献[94]使用TCAM作为时间注意力机制实现了在未修辑的视频中对异常行为进行定位。

2.7 模型迁移方法

虽然深度学习模型算法在行为识别领域取得了不小的成果,同时在现实中也得到了广泛的应用,但对于许多现实场景仍然存在一定的局限性,例如在许多现实场景中对于样本数据的获取成本高昂、耗费时间,甚至无法获取。少样本学习虽然一定程度上能减少模型对于样本数据量的需求,但无法从根本上解决该问题。因此,在对许多模型进行训练时,由于无法对样本数据很好的聚类,导致模型难以得到较好训练的现象普遍存在。然而,迁移学习将目光聚焦于知识转移,很好地解决了上述问题。迁移学习是通过将一个或多个源任务中的知识用于目标任务改进的一种学习方法。通常我们可以把迁移学习看作深度学习算法的扩展,它解决了深度学习模型算法由于训练数据不足导致模型泛化能力不足、识别率低的问题。本质上讲,根据迁移方法的不同,可将迁移学习方法分为样本迁移、特征

迁移、模型迁移(也叫作参数迁移)、关系迁移4种,其中模型迁移是当下最常用同时也是最有效的的迁移学习方法之一。

在人体行为识别领域,传统算法和深度学习算法难以解决在新的样本数据和测试数据具有不同特征分布条件下,模型泛化能力低的问题。模型迁移通过考虑源任务和目标任务之间的互通相关性和互联差异,增强目标任务的泛化能力。文献[95]将模型迁移思想用于自然视频图像预训练的CNN模型中,进行人体头部运动趋势识别,将其检测率提高到了98%以上。文献[96]证明了广泛的迁移信息和少量个人信息相结合,足以减少模型对于主观数据的依赖,提高模型的泛化能力。

模型迁移一般在预训练和微调模型模式下进行,在模型没有足够样本数据时,通过预训练和微调模型可以使模型在迁移时快速适应新的数据集,加快其收敛速度,同时有效解决由于数据样本缺失导致模型过拟合的问题。例如,文献[97]将模型迁移和集合学习分类器用于卷积神经网络模型,增强了模型提取抽象特征的能力并加速了模型收敛速度。又如,文献[98]通过深度模型迁移方法识别CNN模型和微调CNN模型以提取的高级特征的行为,通过在12个数据集上对二者进行测试,证明了微调CNN模型对于模型迁移的有效性。

虽然模型迁移极大地解决了深度学习算法对于样本数据不足导致的模型泛化能力低、收敛速度慢、识别精度不足等问题,但模型迁移也具有一定的局限性,其具有不错识别精度建立在源任务和目标任务共享一些超参数的先验分布的假设之下,当目标任务与源任务之间不存在相关性,或者相关性较少时,对于模型的迁移很大概率不会成功。另外,在模型迁移过程中也可能出现两者之间关系呈负相关的情况,那么就会导致目标任务受源任务误导,出现负面影响,这种现象也被称为负转移。

2.8 算法性能对比

人体异常行为识别领域常用数据集包括UCF101[99]和HMDB-51[100]。UCF101是一个源自YouTube并被广泛使用的基础测试集,它包括101个动作类,约13 320个视频,每个类至少包含100个视频片段,视频片段的分辨率为320×240,帧率为25 frame/s,其中约9 500个训练样本和3 700个测试样本。HMDB51数据集源自网络真实视频或电影,共有51个类别,约6 766个视频片段,其中约3 570个训练样本和1 530个测试样本。如表2为各算法在UCF101和HMDB51数据集上的表现及特点。

由表2可以看到在UCF101和HMDB51数据集上,混合网络的平均识别精度较好,根本原因在于混合网络集合了多种网络的优势,同时也与大量的样本数据有关,这导致其整体结构庞大、参数过多,因此在实际应用中资源消耗严重、对硬件设施要求极高,难以在现实中部署。少样本学习集成类别类器和二域判别器,利用有

表2 各算法性能对比

Table 2 Performance comparison of each algorithm

算法		ACC/%			特点
		UCF101	HMDB51	Others	
3D-CNN	I3D[35]	95.6	74.8	—	改善通道特征组成,促进有用信息利用
	C3D[36]	82.3	51.6	—	推理速度快,具有更紧凑的特征表示
	T-C3D[37]	91.8	62.8	—	降低网络计算成本,实现实时处理
	P3D[38]	88.6	—	80.8	深入提高结构多样性,以提高网络性能
	R(2+1)D[39]	97.3	78.7	—	简化3D卷积结构,提高模型泛化能力
	3D-MRCNN[40]	85.7	—	—	具有较高的准确性和操作效率
Hybrid network	Two-steam CNNs+PI-LSTM[50]	93.2	71.8	—	充分利用长期动态视觉线索,提高模型分辨率
	2D CNN+P3D[51]	94.3	68.6	—	使行为外观与运动特征高效融合,提高模型识别精度
	iDT(FV)+Objects+VideoLSTM[52]	92.2	64.9	—	增添时间注意力提高动作定位准确性
Two-Stream	TSN[60]	94.9	71.0	89.6	基于分段采样和聚合方案模拟长时间结构
	DenSenet+BNIception[61]	89.3	—	92.1	充分利用不同视频阶段的时间相关性
	VGG16[62]	90.6	58.2	—	融合外观和运动特征,以便更好利用这种时空特征信息
	BILSTM[63]	—	—	78.6	充分利用长距离光流信息,解决了行为识别过程中的闭塞问题
	ResNet50+TSM[64]	92.2	66.1	—	强调细节特征,提高时序关系的辨识能力
Few-Shot	HF-AR[73]	89.0	63.1	—	通过跨模态操作提高视频分类性能
	BD-CSPN[74]	—	—	90.2	利用标签传播减小类内偏置,提高了模型的性能
	CBT[75]	79.5	44.6	—	无需矢量化或预训练
	CPDE[76]	—	—	89.3	减少了语义偏差对模型性能的影响
Others	iDT+FV[101]	85.9	57.2	—	通过探索不同BoVW框架,改进密集轨迹的互补性
	STGR-GCN[102]	—	—	92.3	提高模型对于骨骼关节点间高阶相关性的分析能力
	IPL-JPDA[103]	—	—	93.2	为模型训练提供多模态信息,提高了模型泛化能力

限的标签信息进行模型训练,大大减少了模型对于样本数据量的需求,但由于人体行为的多样性,模型中分类器难以捕获表现模糊的特征信息,这是少样本学习平均识别精度低的根本原因之一。3D卷积神经网络注重运动信息,具有良好的空间特征提取能力,运算速度快,然而三维卷积会产生大量的模型参数,导致较高的运算成本和资源消耗。此外,使用3D卷积网络提取特征信息时容易造成梯度消失、梯度爆炸、收敛效果差等问题,虽然通过降低模型复杂度以及引入残差块的方法可以有效提高模型效果,但是模型复杂度的降低,也会一定程度上导致模型识别率的下降,同时残差块中使用的卷积滤波器会造成部分特征信息的丢失,也无法对多尺度特征信息达到很好的融合效果。基于双流架构算法注重时空信息、准确率较高,例如TSN的稀疏采样方法可以覆盖整个视频的各个时间段,使网络更好地利用时间通道特征,但其采样过程是随机的,不能保证采样结果均为有效的特征信息。总而言之,3D-CNN、混合网络、双流卷积网要想实现较高识别精度,需要对样本中每个类进行大量标记,而FSL只需要对少量样本进行标记就可以实现较高的识别精度,这也是FSL具有较高研究价值的意义所在。

3 总结与展望

3.1 总结

人体行为识别技术多被应用于智慧医疗、智慧城市、人机交互、安全生产等领域,很好地保障了人们的生产和生活安全,另外,其相关技术对入侵检测、视频语义检索等方面研究也具有极高的参考价值。虽然目前深度学习算法使得行为识别技术稳固发展,但不能否认的是大部分算法仍存在空间特征表示与时间特征表示难、计算复杂度高等问题,同时目前的算法只针对特定场景,不具有一定的泛化能力,需要大量的样本数据进行训练,其较高的计算复杂度对硬件算力要求高,难以应用于现实问题。随着人工智能技术的发展以及人们生产生活迫切需求,未来的算法逐渐偏向于泛化能力强、计算复杂度低、少样本甚至无样本数据训练。基于元学习与transformer的FSL模型构建是未来的研究热点之一。

3.2 展望

综合基于深度学习的人体行为识别研究和应用现状及存在问题,提出如下几点展望。

(1)跨场景行为识别模型。现有的模型通常是在特定场景下进行训练,仅能适应特定场景,无法做到跨场景识别。在实际应用中,模型迁移至其他场景时,需要对模型重新训练,这无疑增加了巨大的工作量。因此,如何在模型训练过程中融入环境信息,使模型具备跨场景识别的能力是未来的研究方向之一。

(2)现实中部署难。目前大多数高识别精度模型往往具有较高的复杂度,这导致模型在实际应用过程中需要部署在高算力的硬件设备之上,同时高复杂度也会使模型难以进行实时检测,为模型在现实中部署带来难题。如何在保证模型识别精度的同时降低其复杂度,具有较高的研究价值。

(3)行为预测。旨在从实时视频流中准确判别存在安全隐患的行为,遏制存在安全隐患的行为。这不仅对模型的识别速度、识别精度有一定的要求,同时要求模型对于行为的判断要有一定的前瞻性,目前大多数算法只做到了事后识别,而未做到事前预防,因此对于该领域的研究还需要更为深入。

(4)少样本学习模型。现实中许多场景对于样本数据的采集始终是难以解决的问题,由于数据样本的缺少,无法进行有效的训练。目前以元学习与transformer为主的FSL模型虽然极大地减少了模型对于数据样本的依赖,但相较于其他模型其识别精度也有所下降。因此,如何在减少样本数据的同时,提高模型的性能,保证其具有较高的识别精度,仍需要进一步探索。

4 结束语

概述了现有人体行为识别领域常用特征提取方法,以及3D卷积神经网络、双流卷积神经网络、FSL等算法在行为识别领域的发展与应用,对比了各类算法在UCF101和HMDB51数据集上的性能表现,对各类算法优缺点进行了总结,同时对基于深度学习的人体行为识别模型发展提出了几点展望,以期在理论和实践上对人体行为识别模型的发展提供参考。

参考文献:

[1] HATIRNZA E,SAH M,DIREKOGLU C.A novel framework and concept-based semantic search Interface for abnormal crowd behaviour analysis in surveillance videos[J].Multimedia Tools & Applications,2020,79:17579-17617.

[2] 黄凯奇,陈晓棠,康运锋,等.智能视频监控技术综述[J].计算机学报,2015,38(6):1093-1118.
HUANG K Q,CHEN X T,KANG Y F,et al.Intelligent visual surveillance:A review[J].Chinese Journal of Computers,2015,38(6):1093-1118.

[3] QIAN Y J,GUO M L,JIN W Y,et al.A model based method of pedestrian abnormal behavior detection in traffic scene[C]/Proceedings of IEEE First International Smart Cities Conference(ISC2),2015:1-6.

[4] LENTZAS A,VRAKAS D.Non-intrusive human activity recognition and abnormal behavior detection on elderly people:A review[J].Artificial Intelligence Review,2019,

53:1-47.

[5] 于乃功,柏德国.基于姿态估计的实时跌倒检测算法[J].控制与决策,2020,35(11):2761-2766.
YU N G,BAI D G.Real-time fall detection algorithm based on pose estimation[J].Control and Decision,2020,35(11):2761-2766.

[6] 王佳铖,鲍劲松,刘天元,等.基于工件注意力的车间作业行为在线识别方法[J].计算机集成制造系统,2021,27(4):1099-1107.
WANG J C,BAO J S,LIU T Y,et al.Online method for worker operation recognition based on attention of workpiece[J].Computer Integrated Manufacturing Systems,2021,27(4):1099-1107.

[7] LU J,YAN W Q,NGUYEN M.Human behaviour recognition using deep learning[C]//Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance(AVSS),2018:1-6.

[8] VISHWAKARMA D K,KAPOOR R.Hybrid classifier based human activity recognition using the silhouette and cells[J].Expert Systems with Applications,2015,42(20):6957-6965.

[9] MAITY S,BHATTACHARJEE D,CHAKRABARTI A.A novel approach for human action recognition from silhouette images[J].IETE Journal of Research,2017,63(2):160-171.

[10] TANG Z,GU R,HWANG J N.Joint multi-view people tracking and pose estimation for 3D scene reconstruction[C]//Proceedings of the IEEE International Conference on Multimedia and Expo(ICME),2018:1-6.

[11] CHAI Y,SAPP B,BANSAI M,et al.Multipath:Multiple probabilistic anchor trajectory hypotheses for behavior prediction[J].arXiv:1910.05449,2019.

[12] AN L,TSOU M H,CROOK S E S,et al.Space-time analysis:Concepts,quantitative methods,and future directions[J].Annals of the Association of American Geographers,2015,105(5):891-914.

[13] XIA L,LI Z.A new method of abnormal behavior detection using LSTM network with temporal attention mechanism[J].The Journal of Supercomputing,2021,77(4):3223-3241.

[14] WANG Y,WANG S,ZHOU M,et al.TS-I3D based hand gesture recognition method with radar sensor[J].IEEE Access,2019,7:22902-22913.

[15] SUN L,JIA K,YEUNG D Y,et al.Human action recognition using factorized spatio-temporal convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision,2015:4597-4605.

[16] DU Y,FU Y,WANG L.Skeleton based action recognition with convolutional neural network[C]//Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition(ACPR),2015:579-583.

[17] KAMAL S,JALAL A,KIM D.Depth images-based human detection,tracking and activity recognition using spatiotemporal features and modified HMM[J].Journal of Electrical Engineering and Technology,2016,11(6):1857-1862.

[18] SI C,CHEN W,WANG W,et al.An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:1227-1236.

[19] 苏江毅,宋晓宁,吴小俊,等.多模态轻量级图卷积人体骨架行为识别方法[J].计算机科学与探索,2021,15(4):733-742.
SU J Y,SONG X N,WU X J,et al.Skeleton based action recognition algorithm on multi-modal lightweight graph convolutional network[J].Journal of Frontiers of Computer Science and Technology,2021,15(4):733-742.

[20] ZHANG K,LING W.Joint motion information extraction and human behavior recognition in video based on deep learning[J].IEEE Sensors Journal,2019,20:11919-11926.

[21] ZHAO Z,ZOU W,WANG J J.Action recognition based on C3D network and adaptive keyframe extraction[C]//Proceedings of the IEEE 6th International Conference on Computer and Communications(ICCC),2020:2441-2447.

[22] LIU G,ZHANG C,XU Q,et al.I3D-shufflenet based human action recognition[J].Algorithms,2020,13(11):301.

[23] YAN L C,YOSHUA B,GEOFFREY H.Deep learning[J].Nature,2015,521:436-444.

[24] ZEILER M D,FERGUS R.Visualizing and understanding convolutional networks[C]//Proceedings of the European Conference on Computer Vision,2014:818-833.

[25] SIMONYAN K,VEDALDI A,ZISSERMAN A.Deep inside convolutional networks:Visualising image classification models and saliency maps[J].arXiv:1312.6034,2013.

[26] BALLESTER P,ARAUJO R M.On the performance of GoogLeNet and AlexNet applied to sketches[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence,2016.

[27] TARG S,ALMEIDA D,LYMAN K.Resnet in resnet:Generalizing residual architectures[J].arXiv:1603.08029,2016.

[28] NGUYEN N C,VUDINH T.A new architecture for tele-radiology networks[C]//Proceedings of the International Conference on Advanced Technologies for Communications(ATC),2015:394-399.

[29] KIM W,CHOI H K,JANG B T,et al.Driver distraction detection using single convolutional neural network[C]//

Proceedings of the International Conference on Information and Communication Technology Convergence (ICTC),2017:1203-1205.

[30] SANDLER M, HOWARD A, ZHU M, et al.Mobilenetv2:Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:4510-4520.

[31] HOWARD A, SANDLER M, CHU G, et al.Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1314-1324.

[32] JAIN M, GEMERT J V, SNOEK C G M.University of Amsterdam at THUMOS challenge 2014[M]//THUMOS challenge 2014:notebook papers.Orlando, FL:Center for Research in Computer Vision, University of Central Florida,2014.

[33] WANG L, QIAO Y, TANG X.Action recognition and detection by combining motion and appearance features[J]. THUMOS14 Action Recognition Challenge,2014,1(2):2.

[34] JI S, XU W, YANG M, et al.3D convolutional neural networks for human action recognition[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012,35(1):221-231.

[35] CARREIRA J, ZISSERMAN A.Quo Vadis, action recognition? a new model and the kinetics dataset[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017:6299-6308.

[36] TRAN D, RAY J, SHOU Z, et al.Convnet architecture search for spatiotemporal feature learning[J].arXiv: 1708.05038,2017.

[37] LIU K, LIU W, GAN C, et al.T-C3D:Temporal convolutional 3D network for real-time action recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence,2018:7138-7145.

[38] LU B, LV Z, ZHU S.Pseudo-3D residual networks based anomaly detection in surveillance videos[C]//Proceedings of the Chinese Automation Congress (CAC), 2019: 3769-3773.

[39] TRAN D, WANG H, TORRESANI L, et al.A closer look at spatiotemporal convolutions for action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:6450-6459.

[40] PENG B, YAO Z, WU Q, et al.3D Convolutional neural network for human behavior analysis in intelligent sensor network[J].Mobile Networks and Applications, 2022,80:1-10.

[41] HU Z, HU Y, LIU J, et al.3D separable convolutional neural network for dynamic hand gesture recognition[J]. Neurocomputing,2018,318:151-161.

[42] QIU Z, YAO T, MEI T.Learning spatio-temporal representation with pseudo-3D residual networks[C]//Proceedings of the IEEE International Conference on Computer Vision,2017:5533-5541.

[43] WANG M, ZHU Y, SUN Z, et al.Abnormal behavior detection of ATM surveillance videos based on pseudo-3D residual network[C]//Proceedings of the IEEE 4th International Conference on Cloud Computing and Big Data Analysis(ICCCBDA),2019:412-417.

[44] 王新文,谢林柏,彭力.跌倒异常行为的双重残差网络识别方法[J].计算机科学与探索,2020,14(9):1580-1589.
WANG X W, XIE L B, PENG L.Double residual network recognition method for falling abnormal behavior[J]. Journal of Frontiers of Computer Science and Technology, 2020,14(9):1580-1589.

[45] JIANG H, PAN Y, ZHANG J, et al.Battlefield target aggregation behavior recognition model based on multi-scale feature fusion[J].Symmetry,2019,11(6):761.

[46] SHEN H X, LI Y, CHAEN H, et al.Research on human action recognition based on improved pooling Algorithm[C]//Proceedings of the Chinese Control And Decision Conference(CCDC),2020:3306-3310.

[47] YU G, LIU J, ZHANG C.An abnormal behavior recognition method based on fusion features[C]//Proceedings of the International Conference on Intelligent Robotics and Applications,2021:222-232.

[48] TSENG C K, LIAO C C, SHEN P C, et al.Using C3D to detect rear overtaking behavior[C]//Proceedings of the IEEE International Conference on Image Processing (ICIP),2019:151-154.

[49] ZHAO C, CHEN M, ZHAO J, et al.3D behavior recognition based on multi-modal deep space-time learning[J]. Applied Sciences,2019,9(4):716.

[50] YU S, CHENG Y, XIE L, et al.A novel recurrent hybrid network for feature fusion in action recognition[J].Journal of Visual Communication and Image Representation, 2017,49:192-203.

[51] LI S, ZHAO Z, SU F.A spatio-temporal hybrid network for action recognition[C]//Proceedings of the IEEE Visual Communications and Image Processing(VC) IP,2019:1-4.

[52] LI Z, GAVRILYUK K, GAVVES E, et al.VideoLSTM convolves, attends and flows for action recognition[J]. Computer Vision and Image Understanding,2018,166: 41-50.

[53] TANBERK S, KILIMCI Z H, TUKEL D B, et al.A hybrid deep model using deep learning and dense optical flow approaches for human activity recognition[J].IEEE Access,2020,8:19799-19809.

[54] JIN W, XUE Y, MENG X, et al.Research on behavior recognition algorithm based on SE-I3D-GRU network[J]. High Technology Letters,2021,27(2):163-172.

[55] JIANG Y G,WU Z,TANG J,et al.Modeling multimodal clues in a hybrid deep learning framework for video classification[J].IEEE Transactions on Multimedia,2018,20(11):3137-3147.

[56] ALAZZAWI N A.Human action recognition based on hybrid deep learning model and shearlet transform[C]// Proceedings of the 12th International Conference on Information Technology and Electrical Engineering (ICITEE),2020:152-155.

[57] JAOUEDI N,BOUJNAH N,BOUHLEL M S.A new hybrid deep learning model for human action recognition[J].Journal of King Saud University-Computer and Information Sciences,2020,32(4):447-453.

[58] 钱慧芳,易剑平,付云虎.基于深度学习的人体动作识别综述[J].计算机科学与探索,2021,15(3):438-455.
QIAN H F,YI J P,FU Y H.Review of human action recognition based on deep learning[J].Journal of Frontiers of Computer Science and Technology,2021,15(3):438-455.

[59] ULLAH H,KHAN S D,ULLAH M,et al.Two stream model for crowd video classification[C]//Proceedings of the 8th European Workshop on Visual Information Processing(EUVIP),2019:93-98.

[60] WANG L,XIONG Y,WANG Z,et al.Temporal segment networks for action recognition in videos[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,41(11):2740-2755.

[61] YE Q,LIANG Z,ZHONG H,et al.Human behavior recognition based on time correlation sampling two-stream heterogeneous grafting network[J].Optik,2022,251:168402.

[62] FEICHTENHOFER C,PINZ A,ZISSERMAN A.Convolutional two-stream network fusion for video action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:1933-1941.

[63] WANG C H,WEI Y Q,GUO D,et al.Human behavior recognition under occlusion based on two-stream network combined with BiLSTM[C]//Proceedings of the Chinese Control and Decision Conference,2020:3311-3316.

[64] 张红颖,安征.基于改进双流时空网络的人体行为识别[J].光学精密工程,2021,29(2):420-429.
ZHANG H Y,AN Z.Human action recognition based on improved two-stream spatiotemporal network[J].Optics and Precision Engineering,2021,29(2):420-429.

[65] LIU C,YING J,YANG H,et al.Improved human action recognition approach based on two-stream convolutional neural network model[J].The Visual Computer,2021,37(6):1327-1341.

[66] SARABU A,SANTRA A K.Distinct two-stream convolutional networks for human action recognition in videos using segment-based temporal modeling[J].Data,2020,5(4):104.

[67] YANG X,LIU L,WANG N,et al.A two-stream dynamic pyramid representation model for video-based person re-identification[J].IEEE Transactions on Image Processing,2021,30:6266-6276.

[68] TANG Y,WANG Y,XU Y,et al.Beyond dropout:Feature map distortion to regularize deep neural networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence,2020:5964-5971.

[69] ZUNINO A,BARGAL S A,MORERIO P,et al.Excitation dropout:Encouraging plasticity in deep neural networks[J].International Journal of Computer Vision,2021,129(4):1139-1152.

[70] DAI C,LIU X,LAI J.Human action recognition using two-stream attention based LSTM networks[J].Applied Soft Computing,2020,86:105820.

[71] CHI L,TIAN G,MU Y,et al.Two-stream video classification with cross-modality attention[C]//Proceedings of the IEEE International Conference on Computer Vision Workshop(ICCVW),2019:4511-4520.

[72] PENG Y,ZHAO Y,ZHANG J.Two-stream collaborative learning with spatial-temporal attention for video classification[J].IEEE Transactions on Circuits and Systems for Video Technology,2018,29(3):773-786.

[73] KUMAR N,NARANG S.Few shot activity recognition using variational inference[J].arXiv:2108.08990,2021.

[74] LIU J,SONG L,QIN Y.Prototype rectification for few-shot learning[C]//Proceedings of the European Conference on Computer Vision,2020:741-756.

[75] SUN C,BARADEL F,MURPHY K,et al.Learning video representations using contrastive bidirectional transformer[J].arXiv:1906.05743,2019.

[76] ZOU Y,ZHANG S,CHEN K,et al.Compositional few-shot recognition with primitive discovery and enhancing[C]// Proceedings of the 28th ACM International Conference on Multimedia,2020:156-164.

[77] LI X,HE Y,ZHANG J A,et al.Supervised domain adaptation for few-shot radar-based human activity recognition[J].IEEE Sensors Journal,2021,21(22):25880-25890.

[78] LI S,LIU H,QIAN R,et al.TA2N:Two-stage action alignment network for few-shot action recognition[J].arXiv:2107.04782,2021.

[79] LIU X,JI Z,PANG Y,et al.DGIG-Net:Dynamic graph-in-graph networks for few-shot human-object interaction[J].IEEE Transactions on Cybernetics,2021(1):1-13.

[80] RIZVE M N,KHAN S,KHAN F S,et al.Exploring complementary strengths of invariant and equivariant

representations for few-shot learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021:10836-10846.

[81] KANG D,KWON H,MIN J,et al.Relational embedding for few-shot classification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2021:8822-8833.

[82] HAN G,HE Y,HUANG S,et al.Query adaptive few-shot object detection with heterogeneous graph convolutional networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2021:3263-3272.

[83] FU Y,ZHANG L,WANG J,et al.Depth guided adaptive meta-fusion network for few-shot video recognition[C]//Proceedings of the 28th ACM International Conference on Multimedia,2020:1142-1151.

[84] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[C]//Advances in Neural Information Processing Systems,2017:5998-6008.

[85] YANG P,METTES P,SNOEK C G M.Few-shot transformation of common actions into time and Space[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021:16031-16040.

[86] FAN R,CHU W,CHANG P,et al.An improved single step non-autoregressive transformer for automatic speech recognition[J].arXiv:2106.09885,2021.

[87] MAZZIA V,ANGARANO S,SALVETTI F,et al.Action transformer:A self-attention model for short-time human action recognition[J].arXiv:2107.00606,2021.

[88] BOUGHIDA A,KOUAHLA M N,LAFIFI Y.A novel approach for facial expression recognition based on gabor filters and genetic algorithm[J].Evolving Systems,2022,13:331-345.

[89] BYEON Y H,KIM D,LEE J,et al.Body and hand-object ROI-based behavior recognition using deep learning[J].Sensors,2021,21(5):1838.

[90] 周波,李俊峰.基于多流卷积神经网络的行为识别[J].计算机系统应用,2021,30(8):118-125.
ZHOU B,LI J F.Behavior recognition based on multi-stream convolutional neural network[J].Computer Systems & Applications,2021,30(8):118-125.

[91] HSUEH Y L,LIE W N,GUO G Y.Human behavior recognition from multiview videos[J].Information Sciences,2020,517:275-296.

[92] LEE E J,KO B C,NAM J Y.Recognizing pedestrian's unsafe behaviors in far-infrared imagery at night[J].Infrared Physics & Technology,2016,76:261-270.

[93] 何坚,郭泽龙,刘乐园,等.基于滑动窗口和卷积神经网络的可穿戴人体活动识别技术[J].电子与信息学报,2022,44(1):168-177.
HE J,GUO Z L,LIU L Y,et al.Human activity recognition technology based on sliding window and convolutional neural network[J].Journal of Electronics & Information Technology,2022,44(1):168-177.

[94] XIE T T,TZELEPIS C,FU F,et al.Few-shot action localization without knowing boundaries[J].arXiv:2106.04150,2021.

[95] KUJANI T,KUMAR V D.Head movements for behavior recognition from real time video based on deep learning ConvNet transfer learning[J].Journal of Ambient Intelligence and Humanized Computing,2021(1):1-15.

[96] RESCIGNO M,SPEZIALETTI M,ROSSI S.Personalized models for facial emotion recognition through transfer learning[J].Multimedia Tools and Applications,2020,79:35811-35828.

[97] YANG X,ZHANG Y,LV W,et al.Image recognition of wind turbine blade damage based on a deep learning model with transfer learning and an ensemble learning classifier[J].Renewable Energy,2021,163:386-397.

[98] BU Q,YANG G,MING X,et al.Deep transfer learning for gesture recognition with WiFi signals[J].Personal and Ubiquitous Computing,2020,939:1-12.

[99] SOOMRO K,ZAMIR A R,SHAH M.UCF101:A dataset of 101 human actions classes from videos in the wild[J].arXiv:1212.0402,2012.

[100] KUEHNE H,JHUANG H,Garrote E,et al.HMDB:A large video database for human motion recognition[C]//Proceedings of the International Conference on Computer Vision,2011:2556-2563.

[101] PENG X,WANG L,WANG X,et al.Bag of visual words and fusion methods for action recognition:Comprehensive study and good practice[J].Computer Vision and Image Understanding,2016,150:109-125.

[102] LI B,LI X,ZHANG Z,et al.Spatio-temporal graph routing for skeleton-based action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2019:8561-8568.

[103] FU Z,HE X,WANG E,et al.Personalized human activity recognition based on integrated wearable sensor and transfer learning[J].Sensors,2021,21(3):885.