

基于深度学习的视频人体行为识别研究进展

孙德刚^{1,2}, 白入文^{1,2}, 李 敏^{1,2}, 孟 博³, 李凌涵^{1,2}, 杨 阳^{1,2},
姜 森^{1,2}, 任俊星^{1,2}, 李风发⁴, 黄子豪⁵

¹中国科学院信息工程研究所 北京 中国 100093

²中国科学院大学 北京 中国 100049

³北京理工大学 北京 中国 100081

⁴西安电子科技大学 西安 中国 710126

⁵华北电力大学 北京 中国 102206

摘要 “大智移云物”的成熟发展不断推动着视频监控的自动化、智能化, 智能视频监控已经逐渐替代传统视频监控, 成为各行业安防的重要组成部分。在智能视频监控系统中, 识别人体行为对有效发现潜在的危险因素、场所动态监管、异常事件预警等都有重要作用。然而在真实视频监控场景中识别人体行为仍面临很大的挑战。本文旨在为视频监控中人体行为识别技术的研究提供必要的参考, 从 RGB 视频、人体骨骼与 RGB+D 视频三种模态数据全面概述了近六年深度行为识别模型的研究进展。本文主要依据模型的识别精度并结合模型的大小、计算效率和推理速度, 比较基于不同数据模态的深度模型架构, 并分析了各种方法应用不同数据来识别人体行为的优点和局限性。最后重点讨论了智能视频监控系统中的人体行为识别面临的挑战及未来研究的潜在方向。

关键词 智能视频监控; 人体行为识别; 深度学习; 多模态数据

中图分类号 TP306.1、TP309.2 DOI号 10.19363/J.cnki.cn10-1380/tn.2026.05.19

Research Progress of Human Activity Recognition in Videos Based on Deep Learning

SUN Degang^{1,2}, BAI Ruwen^{1,2}, LI Min^{1,2}, MENG Bo³, LI Linghan^{1,2}, YANG Yang^{1,2},
JIANG Miao^{1,2}, REN Junxing^{1,2}, LI Fengfa⁴, HUANG Zihao⁵

¹ Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

² University of Chinese Academy of Sciences, Beijing 100049, China

³ Beijing Institute of Technology, Beijing 100081, China

⁴ Xidian University, Xi'an 710126, China

⁵ North China Electric Power University, Beijing 102206, China

Abstract The mature development of big data, artificial intelligence, mobile Internet, cloud computing and the Internet of Things accelerate automation and intelligence of video surveillance, intelligent video surveillance has gradually replaced traditional surveillance, as an important part of the security in various industries. In intelligent video surveillance systems, recognizing human activities plays an important role in effective discovering potential risk factors, dynamic supervision of scenes, and early warning of abnormal events. However, recognizing human activity in real video surveillance scenarios faces significant challenges. This paper aims to provide a necessary reference for the research of human activity recognition, and provides a comprehensive overview of the research progress of deep activity recognition models in the past six years from three modal data, RGB video, human skeleton and RGB+D video. This paper compares different model architectures based on various data modalities based mainly on the recognition accuracy of the models and taking into account the size, computational efficiency and inference speed of the models, and analyzes the advantages and limitations of various approaches which apply different data to recognize human activities. Finally, it focuses on the challenges of human activity recognition in intelligent video surveillance systems and potential directions for future research.

Key words intelligent video surveillance; human activity recognition; deep learning; multimodal data

通讯作者: 孙德刚, 博士, 研究员, Email: sundegang@iie.ac.cn。

本课题得到国家重点研发计划项目(No. 2019YFB1005205)资助。

收稿日期: 2021-03-16; 修改日期: 2021-07-20; 定稿日期: 2023-08-29

1 引言

在“大智移云物”成熟发展的时代背景下, 计算机视觉技术在图形处理器(Graphics Processing Unit, GPU)等强大计算设备的支持下, 以深度学习方法作为主要的处理和分析手段, 为视频内容的理解和分析提供新的可能。计算机视觉技术的目标是期待机器能够代替人眼来自动识别和理解视觉信息。深度学习方法以其强大的表示学习能力, 使机器视觉在图像识别任务上能比人类视觉完成得更优秀, 即针对某些特定任务, 机器与人眼相比能从图像上获得更多的信息。近几年 ImageNet 大规模视觉识别挑战赛(ImageNet large scale visual recognition competition, ILSVRC)的报告结果显示计算机能够实现的图像识别错误率已远远低于人类的 5.1%^[1]。计算机视觉技术的突飞猛进和深度学习的发展不仅拓宽了图像领域的应用, 同时也让视频领域的研究者看到了新的希望。传统视频监控依托计算机视觉技术逐步从“被动监控”转向“主动监控”, 进入智能化监控的时代。监控系统的功能日益丰富, 安防效能不断提升。人体行为分析作为智能视频监控系统中的一个重要技术模块, 对监控场所中存在的危险因素感知、异常事件预警以及安防管控都有着重要意义。

1.1 视频人体行为识别概述

视频人体行为识别是计算机通过对视频图像帧序列的处理, 来获取并分析视频数据中包含的信息, 从而能够自动识别视频中人的行为。该研究任务涉及目标检测、语义分割、运动分析以及姿态识别等多个计算机视觉领域内的研究课题, 因而挑战性很大, 但也有着广泛的应用前景。在互联网与多媒体技术飞速发展的时代, 前端视频硬件设备发展成熟, 道路监控摄像头、人脸识别门禁、银行安保摄像头等安防设施在日常生活中随处可见, 视频人体行为识别在安防领域大有可为。

数据总是深度学习方法发展的第一驱动力。目前性能优越的人体行为识别方法大多依赖强监督的学习方式和大规模带标签的训练数据。也有一些研究工作^[2-3]在识别人体行为时采用弱监督和无监督的学习方式。这些工作具有很好的参考价值, 但其有限的性能表现远无法满足实际场景的应用需求。因而本文综述的行为识别方法均采用强监督的学习方式, 其依赖各种模态的数据展现了优越的性能。从数据角度出发, 这些数据模态包括 RGB 视频、深度数据、人体骨骼数据, 如图 1 所示。图 1 概述了针对不同视频数据模态、采用不同深度模型架构的视频人体行为识别研究工作。

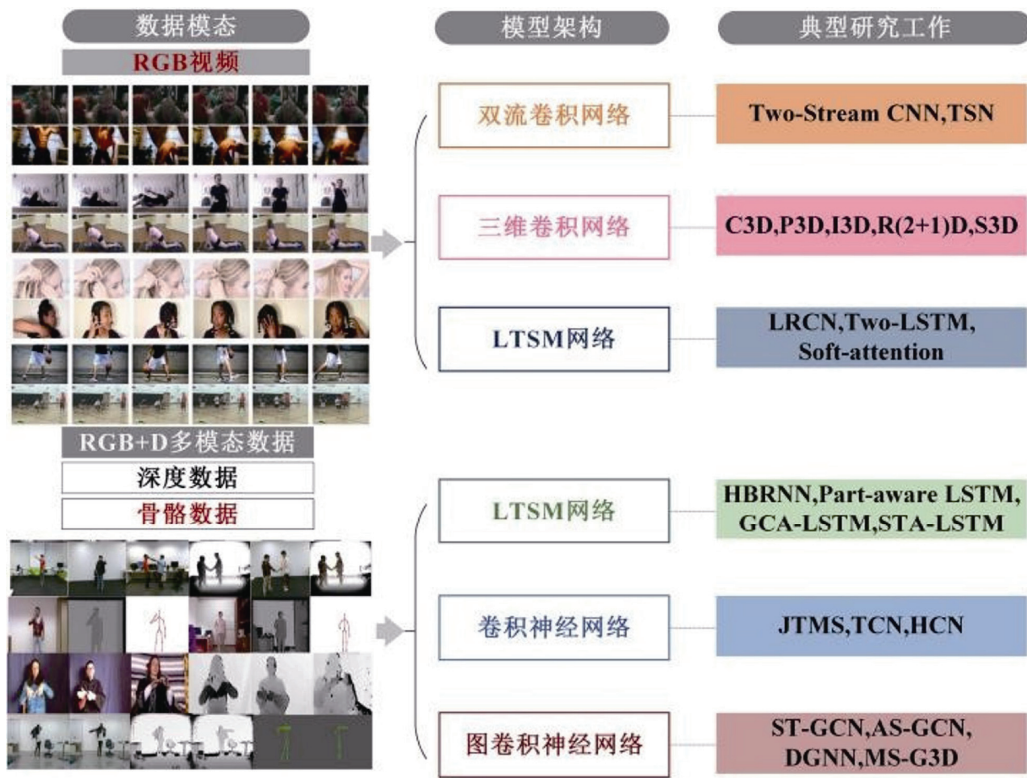


图 1 视频人体行为识别的深度学习研究方法研究概述

Figure 1 An overview of deep learning methods for video human activity recognition

早期的研究者尝试将在图像任务中表现良好的成熟模型扩展应用于 RGB 视频。由于视频数据是包含时间维度的三维空时体, 包含了更丰富且复杂的信息, 因而这些深度模型从二维图像数据到视频数据的扩展研究并没有实现研究者期望的效果。首先, 视频数据高昂的标注成本限制了视频数据集的规模, 也限制了深度模型对视频数据的表达能力。其次, 复杂的背景、视角变换、光照变化、人的尺度变化以及运动速度等因素也给视频中的人体行为识别带来额外的挑战。

一些研究方法尝试仅利用深度数据来识别人体行为, 但没有实现理想的识别效果。首先, 深度图像中缺少了颜色纹理和人体姿态信息, 因而包含的具有判别力的信息较少; 再者, Kinect 获取的深度图像数据由于 Kinect 软件和自身硬件原因, 包含了大量的噪声; 另外, 现有的小规模深度图像数据仍不足以驱动卷积神经网络等深度学习模型学习有效的表示, 因而使用深度图像数据训练得到的深度学习模型容易过拟合。近两年随着计算机视觉领域研究的不断深入, 人体骨骼数据能够很容易地通过两种方式获得, 一是使用包含人体骨骼关键点估计与追踪系统的微软 Kinect 设备, 从深度图像中估计得到; 另一种是依赖实时且可靠的人体姿态估计算法^[4]从 RGB 视频数据中估计得到。近两年有大量的研究工作^[5-10]表明人体骨架序列能很好地表示人体运动信息。

本文主要分析比较了基于 RGB 和骨骼两种模态数据的深度学习模型架构, 并简单阐述了当前基于多模态数据融合的研究工作及面临的挑战。

1.2 相关综述工作

近些年来国内外有一些综述文章对人体行为识别的方法进行总结。Guo 等人^[11]分析了静态图像的行为识别方法。Kang 等人^[12]从特征表达与编码和行为分类两个过程分析了行为分类与时序行为检测算法。Han 等人^[13]和 Presti 等人^[14]主要综述了基于骨骼数据进行行为识别的特征表达和方法。这些综述中更多地分析了基于手工特征的行为识别方法。中山大学胡建芳等人^[15]系统地介绍了深度图像数据和骨骼数据两种模态数据的行为识别方法, 包含了传统方法和深度学习方法。其中深度学习的方法主要介绍了目前应用最为广泛的卷积神经网络(Convolutional Neural Networks, CNN)和循环神经网络(Recurrent Neural Network, RNN)。类似地, Wang 等人^[16]分析和比较了十个最新的基于 Kinect 深度相机的人体行为识别算法, 发现基于骨骼数据的特征比深度特征更具鲁棒性, 基于大型数据集训练的深度学习方法要比手工

特征方法好得多。上述工作都主要综述了 RGB 数据、深度数据和骨骼数据中的一种或两种数据模态的行为识别方法。

文献[17-19]分别综述了 2016 年、2017 年、2018 年以前的研究工作, 证明了深度学习方法在学习行为特征表达上的优越性。Asadi-Aghbolaghi 等人^[17]主要根据时序维度的建模方式, 将当前行为和姿态识别的深度学习方法分为 2D CNN 模型, 利用运动建模(如光流)的模型, 3D CNN 模型和 2/3D 与其他模型的混合模型。Wang 等人^[18]与 Hu 等人^[15]的研究工作相比, 详细地综述了行为识别与姿态识别中的运动识别方法, 并且分别针对 RGB、深度信息、骨骼和 RGB+D 多模态数据, 分析比较了深度学习中 CNN 和 RNN 获得空时信息的方法。除此以外, Wang 等人^[18]也详细介绍了多模态的行为识别数据集, 并列出了各个算法在不同数据集上的性能。Zhang 等人^[19]从手工提取的特征表达和深度学习的特征表达两方面全面综述了行为识别的方法, 并总结概括了行为识别中时序行为检测方法与人物交互方法, 指出了当前行为识别在时序行为检测和人物交互上存在的挑战。本文主要综述了近六年深度学习方法在视频人体行为识别中的应用。

Presti 等人^[14]和 Zhang 等人^[19]均将行为识别分为行为分类与时序行为检测两个子任务。行为分类一般是对一段仅包含一个行为实例的视频进行分析, 并将其归类为一个动作类别。然而, 现实生活中大部分视频都是包含多种行为实例与冗余时间段的长视频, 是未经过事先剪切的。时序行为检测可从这样的长视频中检测每种行为的开始、结束时间并确定行为类别。时序行为检测是比行为分类更复杂且更有挑战性的研究任务。行为分类任务中经典的深度学习模型(如 TSNet^[20], C3D^[21], I3D^[22]等)常常扩展应用于时序行为检测任务中。

实际上, 目前视频人体行为识别的研究工作大多致力于提高行为分类模型的性能, 并且研究最广泛的是对单人行为的识别, 即行为分类。换句话说, 近些年大多数文献中对视频人体行为识别的研究关注于人体行为的准确分类。因此本文不区分行为识别与行为分类的概念, 认为视频人体行为识别任务是将输入视频中的人体行为序列, 映射为对应的行为标签, 其大致流程如图 2 所示。

本文将基于 RGB 输入数据模态的行为识别模型根据空时建模方式的不同进行分类介绍, 将基于骨骼数据输入模态的行为识别模型根据骨骼关键点的表达方式不同进行分类介绍, 同时也概述了融合 RGB+D 多模态数据进行人体行为识别的深度学习方法。

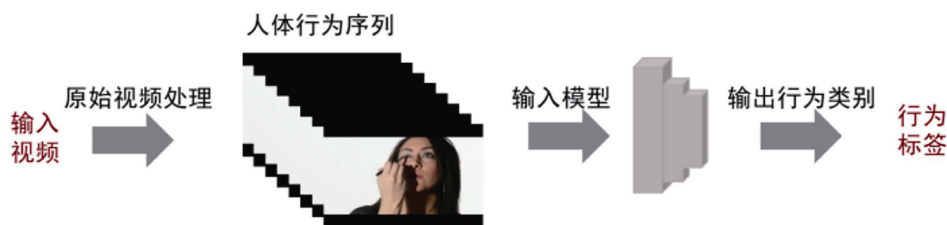


图2 人体行为识别流程示意图

Figure 2 Procedure diagram of human activity recognition

总的来说, 本文的主要贡献有以下几个方面:

(1) 针对视频人体行为识别任务, 本文详细介绍了近六年来关于 RGB 视频和骨骼数据两种数据模态的深度学习模型的研究进展, 行为识别方法的年历表概览如图 3 所示。

(2) 本文在几个目前主流的行为识别基准数据集上比较最新的基于 RGB 视频、骨骼数据以及多模态数据的行为识别模型性能, 包括 RGB 视频数据集 UCF101^[23]、Kinetics^[24], 骨骼数据集 Kinetics-skeleton^[5], 还有多模态数据集 NTU RGB+D^[25]、SYSU 三维人物交互数据集^[26]。

(3) 本文对比并分析了目前性能优越的行为识别模型的大小和推理速度, 通过比较各种行为识别模型的优势与不足, 把握当前行为识别方法应用于实际生活中的挑战以及未来的研究方向。

2 数据集介绍

视频内容理解是极具挑战性的研究课题。为了使深度模型更好地理解复杂的视频内容, 近些年来视频领域的研究者不断收集和标注视频数据来构建大规模的视频数据集。2014 年发表的 Sports 1M^[27] 视频数据集包含了 487 个体育类别, 主要用于识别不同的体育运动。2015 年发表的 FCVID 数据集^[28] 中的视频涵盖不同的场景、动物和事件, 常用于视频分类和视频内容理解。同为 2015 年发表的 ActivityNet 数据集^[29] 的视频总时长达到 849 小时, 其中每段视频平均包含 1.5 个行为实例, 训练标签包含行为实例标签和行为的开始与结束时间。大量未修剪的长视频训练数据常用于时空行为检测任务和复杂的视频行为描述任务中。

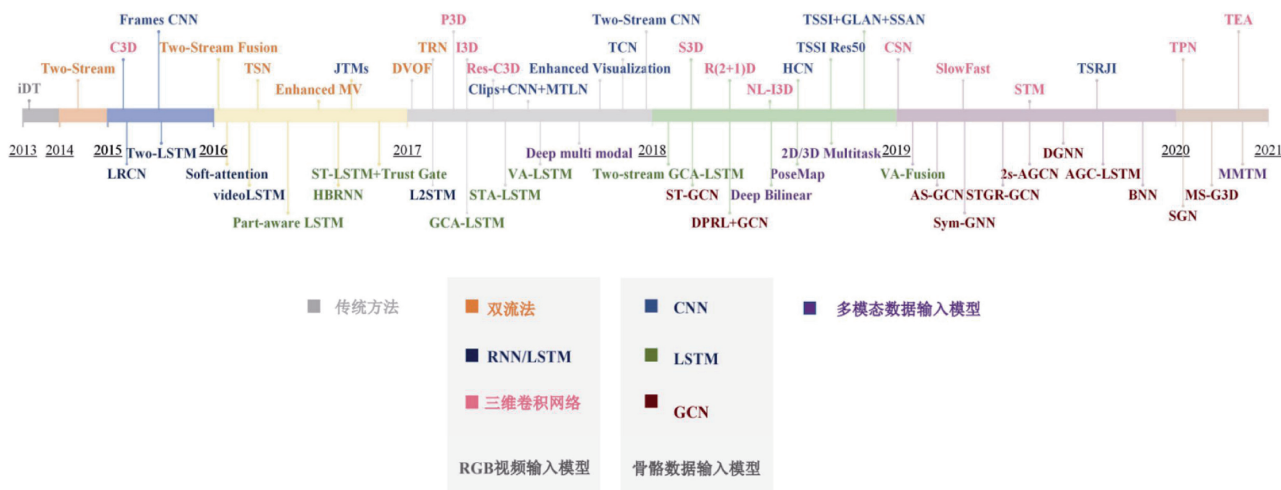


图3 视频行为识别方法年历表概览

Figure 3 Chronological overview of video activity recognition method

本节介绍了行为识别任务中主流的 RGB 视频数据集和 RGB+D 多模态数据集。大量先进的行为识别方法采用这些数据集作为评估的基准数据集。

2.1 RGB 视频行为识别数据集

深度学习方法的一个典型特征是受数据驱动

的。深度学习模型应用于图像领域展现出的优越性能在很大程度上归功于大规模的图像数据集 ImageNet^[30]。该图像数据集包含了近 1500 多万张图像, 共有 2 万多个类别。2012 年公开的 UCF101 视频数据集^[23]是从 YouTube 上收集得到的, 包含了多种

多样的人们日常行为。之后多年,研究者大多以 UCF101 视频数据集作为评估行为识别模型的基准数据集。为了进一步提升深度学习模型对视频中人体行为的理解能力,DeepMind 发布了 Kinetics 视频数据集^[24],它是高质量的 YouTube 视频链接集合,涵盖了多种多样的人类行为。Carreira 等人^[22]实验证明了使用 Kinetics 预训练的模型和使用 ImageNet 数据集预训练后再使用 Kinetics 数据集微调的模型在分类精度上几乎一样,说明 Kinetics 的视频数据量应用于视频领域已相当于图片任务中的 ImageNet 数据集。

2.1.1 UCF101 行为数据集

UCF101 数据集^[23]中包含了从 YouTube 上收集的 13320 个真实行为视频片段,涉及单人动作、乐器演奏、人物交互、人人互动、体育运动 5 大类人类日常行为,一共有 101 个行为类别,每类动作都有 25 个分组,每组中包含来自同一个长视频的 4 到 7 个视频片段,它们有着相同的背景或演员。不同动作的复杂程度不一样,视频片段的长度也不一样。剪辑片段中 5~10s 的视频片段占大多数。大多数 YouTube 视频是由用户自主录制并上传的,来自无约束的现实环境,在摄像机运动、视点变化、杂乱背景、物体外观和姿态、分辨率不一致和光照条件等方面均存在较大的差异,因而 UCF101 数据集挑战性很大,是经典的用于评估行为识别模型的基准数据集。UCF101 数据集划分为 3 个训练/测试数据子集,每个子集中训练和测试的视频片段不能由同一个长视频剪切得到。来自 7 个不同长视频的视频片段构成了每个子集的测试集,对应子集的训练集来自剩下的 18 个长视频。

2.1.2 Kinetics 行为数据集

2016 年最早公开的 Kinetics 数据集^[24]包含 400 类动作,每个动作都有 400~1150 个视频片段,每段视频的时长都在 10 秒左右,一共有 30 多万个视频。到 2017 年, Kinetics 扩充到了 600 个行为类别,大约 50 万个视频片段。2019 年最新的 Kinetics 数据集已从 600 类进一步增加到 700 类,每一类行为至少包含 600 个视频片段,视频片段总量也增加到 65 万以上。Kinetics 数据集的行为类别主要包含三大类:单人行为,人物交互和人人交互。Kinetics 数据集和 UCF101 数据集一样均来自无约束的真实场景,细粒度划分行为类别也给 Kinetics 数据集上的行为识别带来额外的挑战。

2.1.3 RGB 数据集对比及评估方法

上述两个 RGB 视频数据集的对比如表 1 所示,示例图如图 4 所示。在视频人体行为识别任务中,一

般采用识别精度来评估深度学习模型的性能。在 UCF101 和 Kinetics 数据集上计算识别精度的方法是计算平均精度均值(mAP)。

表 1 RGB 视频数据集对比

Table 1 Comparison of RGB video datasets

数据集	年份	行为类别数	每个类别的视频片段数	RGB 视频片段的平均时长	总视频片段数
UCF101	2012	101	最小 101	6s	13320
Kinetics400	2017	400	最小 303	10s	306245
Kinetics600	2018	600	最小 519	10s	482622
Kinetics700	2019	700	最小 532	10s	650317

具体来说,假设数据集一共有 C 个行为类别,在测试时会预测验证集中每一个视频片段属于某个类别的概率,预测结果为概率值从高到低排列的列表,列表长度为 C ,其中每个元素 c 对应一个行为类别。在对比行为类别的预测结果与真值时,一般取每个视频片段的预测得分列表中前 k 个预测概率值 $P(k)$, $\text{rel}(k)$ 是指示函数,如果第 k 个标签与真值对应, $\text{rel}(k)$ 为 1, 否则为 0。因此,如果验证集中共有 n 个视频片段,某个行为类别的平均精度(AP)的计算方式如公式(1)所示:

$$\text{AP}(c) = \frac{\sum_{k=1}^n (P(k) \times \text{rel}(k))}{\sum_{k=1}^n (\text{rel}(k))} \quad (1)$$

平均精度均值(mAP)是所有类别的平均精度求和后再对所有行为类别数取均值。

$$\text{mAP} = \frac{1}{C} \sum_{c=1}^C \text{AP}(c) \quad (2)$$

使用 UCF101 数据集评估深度学习模型时,一般只取得分最高的类别标签作为预测标签($k=1, \text{Top-1}$),若该预测标签与真值对应,则认为该预测结果是真阳性(True Positive)。而在 Kinetics 数据集中,至少 400 类的行为类别划分与 UCF101 数据集的 101 类相比是更细粒度的,因而一个视频片段可能包含多种动作。例如,“打电话”时“喝水”,“梳头发”时“打哈欠”等等。所以在 Kinetics 数据集上进行评估时,计算 Top-1 和 Top-5 两种分类精度,即选择得分最高和预测概率列表中的前 5 个标签作为预测的行为类别标签,只要预测标签中包含真值则认为预测结果为真阳性。

本文在比较模型的推理速度时,选用了两个评价指标。一个是每秒帧率(Frames Per Second, FPS),即每秒模型可处理的视频帧的数量。另一个是每秒浮点运算次数(Giga Floating-point operations per second, GFLOPS)。



图 4 RGB 视频行为识别数据集示例图

Figure 4 Example diagrams of RGB video activity recognition datasets

2.2 RGB+D 多模态数据集

近些年随着微软的 Kinect、华硕的 XtionTM 等深度相机开发成本的降低, 利用多模态的 RGB+D 视频数据进行人体行为识别也成为备受关注的研究问题。与 RGB 视频数据相比, 多模态的 RGB+D 视频数据包含 RGB 视频, 深度和人体骨骼三种模态的数据, 对环境变化、光照变化、人体外貌变化等与行为无关的

外界因素具有较强的鲁棒性, 并且深度相机中的深度信息可提供丰富的场景信息。在 1.1 节中提到, 人体骨骼数据在视频人体行为识别中已经有了很好的应用范例, 因此本文主要介绍在骨骼行为识别研究中使用的多模态(骨骼)数据集: NTU RGB+D 数据集, kinetics-skeleton 数据集以及 SYSU 三维人物交互数据集。RGB+D 多模态的示例图如图 5 所示。



图 5 RGB+D 视频行为识别数据集示例图

Figure 5 Example diagrams of RGB+D video activity recognition datasets

2.2.1 NTU RGB+D 数据集

NTU RGB+D^[25]是由南洋理工大学使用 Microsoft Kinect v2 收集的人体行为数据集, 它包含深度图、3D 关键点信息、RGB 视频帧以及 IR 序列 4 种数据模式。其中用于人体骨骼行为识别任务的 3D 骨骼关键点数据由 25 个身体关键点的 3D 坐标组成。该数据集收集了 40 个对象的 60 个行为类别, 包含日常行为、交互动作和健康相关的动作, 一共 56880 个视频样本和 400 万帧图像。该数据集有两种评估协议: 交叉受试者(CrossSubject, CS)和交叉视角(CrossView, CV)。两种协议下的训练集和测试集的构成不同, 前者的训练集与测试集中分别包含一半受试者(20 个)演示的行为, 后者的训练集中则包含了两个摄像机拍摄的视频样本, 其余摄像机拍摄的视频样本作为测试集。在 2019 年, 该团队将 NTU RGB+D 数据集增加到 114480 个视频样本, 包含 120 类人类日常行为, 是目前规模最大的可用于人体骨骼行为识别的数据集。

2.2.2 Kinetics-skeleton 数据集

Kinetics-skeleton 数据集是使用 OpenPose^[4]工具从 Kinetics-400 数据集中提取人体骨骼关键点数据得到的。在提取骨骼关键点之前, 首先调整所有视频的分辨率到 340×256 的分辨率, 帧率为每秒 30 帧。OpenPose 工具估计每个人的 18 个身体关键点的二维坐标及其对应的置信得分。如果超过两个人, 置信度较低的将被忽略。由此构成了 Kinetics-skeleton 骨骼数据集。

2.2.3 SYSU 三维人物交互数据集

SYSU 三维人物交互数据集^[24]是由中山大学收集的, 一共包含 480 个视频片段, 包含 12 种不同的人物交互行为。其中 40 个受试者进行 12 种不同的动作: 喝水、倒水、打电话、玩手机、背背包, 打包包裹, 坐椅子, 移动椅子, 拿出钱包, 从钱包中取出, 拖地, 扫地。对每个行为, 每个受试者仅使用电话、椅子、包、钱包、拖把和扫帚这六种物品中的一个。每个受试者的骨骼数据包含 20 个关键点。与其他数据集相比, 该数据集的主要挑战性在于某些行为运动姿态以及涉及的交互物品外观相似。该数据集有两种评估协议。在设置-1(setting-1)中, 一半的样本用于训练, 剩下的用于测试。在设置-2(setting-2)中, 一半的受试者用于训练, 其余的用于测试。

上述三个数据集的对比如表 2 所示。

3 基于 RGB 视频数据的行为识别

为获得视频数据的空间和时间信息, 早期的研究者利用成功应用于图像任务的 2D CNN 模型来构

成具有空间和时间两个分支的双流框架, 分别获取视频图像帧的外观特征和帧间的运动特征。由于不同的人体行为持续时间不同, 视频的时间维度也包含了丰富的判别信息。一些研究作为学习更全局的视频时序信息, 采用基于长短期记忆网络(Long Short-Term Memory, LSTM)^[31]的 RNN 沿着时间维度对视频帧序列建模。另外, 有一些研究工作则将 2D CNN 扩展为 3D CNN, 3D 网络使用 3D 卷积核可同时获得视频数据的空时特征。这些深度学习模型获取视频时间维度信息的方式不同, 但是均关注于如何获得视频图像帧中更具判别力的外观特征以及视频帧在时序上的外观变化, 从而识别 RGB 视频数据中的人体行为。

表 2 RGB+D 多模态数据集对比
Table 2 Comparison of RGB+D multi-modal datasets

RGB+D 视频数据集	年份	行为类别数	骨骼关键点个数	总视频样本量
NTU RGB+D 60	2016	60	25	56880
NTU RGB+D 120	2019	120	25	114480
kinetics-skeleton	2018	101	18	306245
SYSU 3DHOI	2017	12	20	480

经典的手工提取特征的方法 iDT(improved Dense Trajectories)^[32]是在深度学习方法应用于视频人体行为识别任务之前性能最好的方法。iDT 使用密集采样的光流获得像素点在时间上的运动轨迹, 并沿着该轨迹提取四种手工特征。最后采用 Fisher Vector 编码提取到的手工特征, 构成用于分类的特征向量。然而编码得到的超高维特征向量极大地限制了算法的计算效率。早期提出的深度学习模型在视频人体行为识别任务中的应用远远没有实现理想的效果, 在集成 iDT 方法之后仍能获得一定的性能提升。本文仍使用 iDT 作为对比算法之一。

3.1 双流卷积网络

Simonyan 等人^[33]设计了包含空间和时间两个子网络的双流卷积网络, 单个视频图像帧输入到 ImageNet 预训练的空间子网络获得视频帧的空间特征, 时间子网络基于视频帧间计算的密集光流捕获帧间时序运动信息, 最后采用后融合的方式融合两个分支的类别得分得到最终的识别结果。该方法在 UCF101 数据集上的识别精度为 88%, 也是首次 CNN 在视频行为识别任务中的表现超过了手工特征提取的方法 iDT。结合光流输入的双流架构是深度学习模型应用到视频行为识别研究任务的一次成功尝试, 其良好的表现激发了该领域的研究者对双流架

构的关注。随着图像任务中深度学习模型的不断改进,在双流架构下改变骨干网络是提高视频行为识别分类精度的一种方式。

双流融合法(Two-stream Fusion)^[34]使用 VGG-Net^[35]深度模型作为骨干网络。除此以外,与双流卷积网络在全连接层后简单地融合分类得分不同,双流融合法在最后一个卷积层后融合特征用于行为识别,实验证明后者能进一步提高模型性能。然而,这两种双流方法有一个共同的缺陷是没有考虑较长时间上视频帧间的动态变化,空间流仅处理单个视频图像帧,时间流一次仅能处理连续 10 帧的堆叠光流场。为获得视频中长期的时序特征,时序分割网络(Temporal segment networks, TSN)^[20]将整个视频分为 3 段(segment),输入到双流网络中得到每段视频的行为类别概率,并通过段共识函数融合类别得分来得到最终视频级的预测结果。TSN 在 UCF101 数据集上的识别精度达到 94.2%,优越的性能引起了研究者的关注。为更好地学习整个视频的特征表达,Lan 等人^[36]在 TSN 模型的基础上提出深度局部特征(DVOF),将深度网络提取的局部特征聚合为全局特征后,再利用该全局特征学习视频级的行为标签。2017 年 Zhou 等人^[37]提出时序关系推理(TRN)可高效地获得多个时间尺度上的行为时序关系,该方法在时序相关性较强的 Something-Something 视频数据集^[38]上与 TSN 方法相比有了很大的性能提升。

双流架构借助光流计算来表达视频帧序列的时间信息,有着优越的性能表现。但是光流的计算需要消耗大量的时间和计算资源。也有方法尝试能代替光流表示运动信息的方法,Zhang 等人^[39]提出使用计算成本很低的运动矢量(motion vector)来替代光流,实现了能实时进行行为识别的双流框架,推理速度达到每秒 390.7 帧。但是运动矢量中仅包含粗糙的运动模式,严重降低了卷积神经网络的识别性能。

本文在 UCF101 数据集上从精度和推理速度两方面比较上述双流方法的性能,如表 3,“—”表示论文中均未显示或无法衡量。

表 3 双流卷积网络在 UCF101 数据集上的性能对比
Table 3 Performance comparison of two-stream convolutional networks on UCF101 dataset

方法	输入数据	UCF101	推理速度
iDT+FV ^[32]	RGB+Flow	85.9	—
Two-stream ^[33]	RGB+Flow	88.0	—
TSN ^[20]	RGB+Flow	94.0	14fps
DVOF ^[36]	RGB+Flow	94.9	—
Enhanced MV ^[39]	RGB	86.4	390fps

3.2 基于 LSTM 的循环神经网络

CNN 模型在时间维度上仅能处理有限长的视频帧序列。上述双流卷积网络在时间上一次最多仅能处理 10 个视频帧,TSN^[20]和 DVOF^[36]则是间接采用时序聚合方式对整个视频建模,三维卷积网络 C3D^[21]和 I3D^[22]分别能够一次处理 16 帧和 64 帧图片。然而实际生活中人类行为通常持续数百个视频帧,有些行为可通过静止的视频帧或短期的帧间关系准确识别,如握手、打电话等,而有些行为则必须基于长期的时序关系才能区分,如走与跑,跳远与跳高等。总之,学习视频帧间长期的时序关系能够提供更完整的行为信息,有利于识别持续时间长或静态外观易混淆的行为。与 CNN 这样的前馈神经网络不同,RNN 的循环结构允许信息在网络中持续存在,这样可使用先前的视频帧来帮助理解当前帧。但是严格随时间整合状态信息的简单 RNN 模型有一个重要限制是梯度消失效应^[40],即当链接的信息相隔较长时间时,RNN 将无法学习链接信息。特殊的 RNN 网络 LSTM^[31]使学习长期的时序依赖成为可能。LSTM 单元中的隐藏状态增加了非线性机制与简单的门控函数,使状态在传播过程中不会被更新、更改或重置。

Donahue 等人^[41]提出的长期循环卷积网络(Long-term recurrent Convolutional Networks, LRCN)验证了 LSTM 在视觉问题中对序列数据建模的优势,它能够处理可变长度的由 CNN 提取的视频帧级特征。Ng 等人^[42]的研究工作采用了两种视频帧级的特征聚合方式,一种是与视频帧时序无关的特征池化,一种是与视频帧时序相关的 LSTM,两种方式一致证明聚合更多的帧级特征能显著提高行为的分类精度,并且使用 LSTM 聚合 30 个连续视频帧级特征的结果相当于卷积池化聚合 120 个视频帧级特征的结果,这也说明视频帧序列中的时序关系对识别行为的重要性。论文同样在双流框架下分别使用 RGB 和光流图像作为输入。

Sharma 等人^[43]在 LSTM 中引入了软注意力机制(soft-attention),使模型能够有选择地关注每一个视频帧中与行为类别相关的空间区域来推断视频中的人体行为。注意力机制的引入在一定程度上弥补了基于 LSTM 的 RNN 学习序列数据空间结构信息上的不足,然而 LSTM 对图像的矢量化使其在学习视觉信息方面仍不如卷积结构有效。VideoLSTM^[44]为更好地学习视频中人体行为的空时相关关系,在 LSTM 单元中引入卷积运算来编码视频帧中的空间相关信息,同时结合光流图像学习注意力机制。上述大多数模型倾向于在 CNN 的基础上使用 LSTM。因

为 CNN 在表达视频帧的空间外观信息方面有着不可替代的优势, LSTM 则能很好地描述全局的视频时序特征。但通过这样的方式学习到的视频帧间运动信息隐含地假设视频中的运动在不同的空间位置上是静止的。Sun 等人^[45]提出了 Lattice LSTM(L2STM), 采用多模态数据训练的方式同时训练两个 LSTM 分支的记忆单元, 有效增强了时间上动态建模能力。上述基于 LSTM/RNN 的视频人体行为识别方法在 UCF101 数据集上的精度如表 4 所示。

表 4 LSTM/RNN 深度行为识别模型在 UCF101 数据集上的性能对比

方法	输入数据	预训练	UCF101
LRCN ^[41]	RGB+Flow	ImageNet	82.3
Two-LSTM ^[42]	RGB+Flow	ImageNet+ 1M Sports	88.6
Soft-attention ^[43]	RGB	ImageNet	75.8
videoLSTM ^[44]	RGB+Flow	ImageNet	88.9
L2STM ^[45]	RGB+Flow	ImageNet	93.6

3.3 三维卷积网络

视频数据本身是包含空间和时间两种成分的三维体, 学习鲁棒的视频特征表达是具有挑战性的, 主要难点是学习人体行为在空间和时间两个维度上的相关关系。前文中提到的在双流框架下扩展 2D CNN 的深度模型都是在 2D CNN 获取视频帧中空间信息的基础上, 不断探索时序上运动特征的表达方式, 这里运动特征指的是视频帧外观的改变。与之相比, 3D CNN 对视频数据的处理是更直观的, 可同时学习视频帧序列中的空间和时间特征。3D 卷积与 2D 卷积的运算过程如图 6 所示, 图 6(a)表示对单通道图像进行 2D 卷积运算后得到 2D 特征图, 图 6(b)表示对 3D 视频空时体进行 3D 卷积运算后得到的特征图也同样是保留了时间维度的 3D 体。

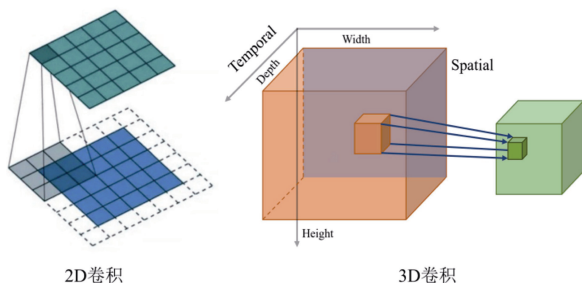


图 6 2D 与 3D 卷积过程示例图

Figure 6 Example diagram of 2D and 3D convolution process

Ji 等人^[46]首先提出将 3D 卷积核应用于视频行为识别任务中, 然而构成的浅层 3D CNN 的模型参数量与 2D CNN 相比呈指数级增长。Karpathy 等人^[27]的研究工作发现这样的 3D CNN 的性能表现与在单个视频帧上使用 2D CNN 相比差异很小, 并没有体现出使用 3D 网络表达时间维度特征的优势。2015 年 Facebook 实验室的 Du 等人的研究工作 C3D^[21]与之前的 3D CNN 相比首次实现了不错的行为分类精度。它是仅有 11 层的浅层网络, 在推理速度上有很大的优势, 尤其是与需要光流计算的深度模型相比。Du 等人^[21]通过 C3D 模型的特征可视化, 证明了 3D CNN 有很好的学习视频空时特征的能力。然而 C3D 网络仍然没有解决巨大的模型参数量的问题。152 层 ResNet^[47]的深度模型大小为 235MB, 浅层的 C3D 网络的模型参数量却达到了 321MB。这样大的模型参数量训练起来是困难的, 并且无法在像 ImageNet 这样大规模的图片数据集上预训练, 浅层的网络也限制了模型的性能。Du 等人^[48]在残差网络的框架下实现深度三维残差卷积网络, 该模型的推理速度是 C3D 的两倍并且模型参数也少了两倍, 其良好的性能表现验证了三维卷积比二维卷积更适合进行空时特征学习。

有限的视频数据量限制了深度模型学习视频特征的能力, 在 ImageNet 图片数据集上预训练的二维卷积网络可在一定程度上缓解这个问题。为进一步提高 3D CNN 模型的泛化能力, Pseudo-3D^[49]在残差学习的框架下改进 C3D 网络, 其包含三种不同的残差块, 实验证明混合三种形式的残差空时卷积的性能最好。类似地, I3D 基于 Inception-V1^[50]深度网络实现了 3D 空时卷积。它们采用的方式都是在性能优越的深度模型基础上将 2D 卷积运算扩展为 3D 卷积运算, 而后按照空间和时间两个维度分解为 2D 空间卷积和 1D 时序卷积(如图 7)。这样扩展得到的 3D CNN 可使用在 ImageNet 数据集上预训练的模型参数, 因而具有更强的特征表达能力。P3D 和 I3D 与 C3D 相比, 模型的深度更深且参数量有明显的减

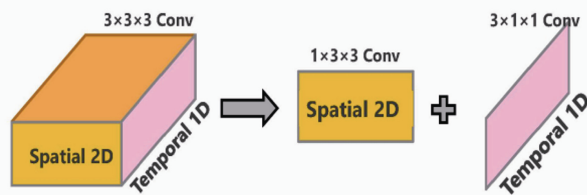


图 7 3D 卷积分解成(2+1) D 卷积(图片来自[49])

Figure 7 3D convolution is decomposed into (2+1) D convolution

少, 它们的识别性能均展现了很好的学习空时表达的能力。

2018 年 Facebook 的研究工作 $R(2+1)D^{[51]}$ 在 P3D 研究工作的基础上只采用了一种不包含瓶颈的空时残差块。谷歌 Deepmind 团队则是在 I3D 网络的基础上提出了深度模型 S3D^[52], 它在准确率、模型尺寸和计算效率上都实现了比 I3D 更好的性能。这两个研究工作进一步地探索了 3D 卷积获取视频人体行为序列中信息的能力, 实现了很好的行为识别精度与推理速度的平衡, 证明了 3D 卷积核在空间和时间上的分解可获得更强的空时特征表达能力。在视频的行为分类任务中应用 2D 可分离卷积大大提升了精度与计算能力, 3D 卷积核分解的性能也通过 $R(2+1)D$ 和 S3D 良好的性能表现得到了验证。受此启发, Facebook 在 2019 年的一个研究工作信道分离卷积网络 (Channel-Separated Convolutional Network, CSN)^[53], 考虑了卷积运算中通道交互的因素, 将一个 3D 卷积核分为 $1 \times 1 \times 1$ 的传统卷积, 用于通道交互; $k \times k \times k$ 的深度卷积用于局部空时交互, 采用信道分离卷积的方式会产生更高的训练误差但模型的泛化性能更好, 相当于对模型的正则化。CSN 简单且高效, 在进一步提高模型识别精度的同时, 计算速度提高了 2 到 3 倍。

上述研究作为更好地将 3D 卷积应用于视频人体行为识别任务, 均采用了分解 3D 卷积核的方式。最初的动因在于使用 ImageNet 预训练后能够大大减少已标注的视频数据需求。随着研究的不断深入, 研究者发现应用 3D 卷积核的分解能够提高精度并减少计算开销。Kinetics 大规模视频数据集的出现使可利用的视频行为识别数据量超过了 30 万条。Hara 等人^[54]使用 Kinetics 数据集训练 152 层深的残差三维网络, 获得了和 I3D 模型相当的识别精度, 说明 Kinetics 视频数据集的规模足够支持从头开始训练深度三维卷积模型, 这极大地促进了 3D 卷积网络在视频人体行为识别中的应用。

3D/(2D+1D)卷积在建模长期时序关系时, 直接采用局部的 3D/(2+1)D 卷积运算处理局部时间窗 (C3D, I3D), 并在深度网络中反复堆叠局部卷积来间接地建模长期时序关系。然而, 在进行大量的局部卷积运算后, 远距离帧的有用特征已被弱化并且无法很好地获得。Wang 等人^[55]则是在 I3D 三维网络的基础上, 使用 ResNet-101 骨干网络, 通过加入非局部模块来获得视频中更长距离的空时依赖关系。在非局部 (Non-local) 运算模块中, 一个位置的特征响应是通过计算所有位置的特征加权和得到的 (如图 8 所示)。He 等人^[56]提出 StNet, 通过对堆叠的视频帧应

用 2D 卷积获得局部空时关系, 并在局部空时特征图上分别进行信道和时间的卷积来获得全局时序依赖关系。这里将视频采样为 T 段, 并在每段中采样 N 个连续的视频帧堆叠成超图像, 构成 $[T \times 3N \times H \times W]$ 的张量输入到网络中。

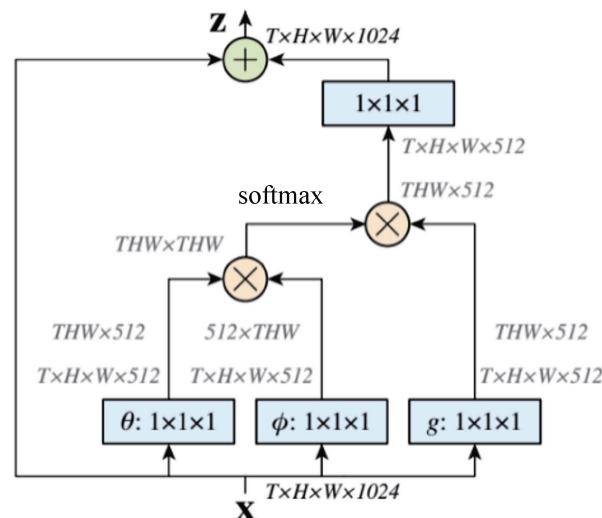


图 8 非局部模块 (图来自文献[55]) “ $\square$ ”表示矩阵乘法, “ $\oplus$ ”表示元素级求和。图中使用张量变换表示特征图的维度变化过程

Figure 8 Non-local module (figure from Ref. [55]) " $\square$ " denotes matrix multiplication, the " $\oplus$ " denotes that the element level sum. In the figure, tensor transformation is used to represent the dimension change process of the feature graph

文献[53-54]都认为动作执行的快慢有利于区分一些视觉外观相似度很高的动作。Feichtenhofer 等人^[57]采用不同的帧率分开处理视频片段, 慢路径可获得少数稀疏帧中包含的语义信息, 快路径则能够获得快速变化的运动, 最终性能基于非局部网络有了进一步的提升。本文在对比该模型与其他模型的性能时, 选择了 3D ResNet-50 作为骨干网络, 当原始的视频帧输入为 64 帧时, 慢路径的输入帧以 8 帧的步长采样为 8 帧, 同时快路径保持 64 帧的输入时, 模型的识别性能最好。类似地, Yang 等人^[58]为处理不同速率的视频帧, 构建了特征级的时序金字塔网络 (Temporal Pyramid Network, TPN), 与 SlowFast 网络^[57]不同的是, 该网络的输入是单一速率采样的视频帧, 之后对获取的特征进行空间语义调制。

为找到更好的、轻量级的运动表达方式, Jiang 等人^[59]提出了一个新的空时运动编码模块 (STM), 并将其整合到 2D ResNet 中。STM 模块包含信道上的时空模块 (CSTM) 和信道上的运动模块 (CMM), 同

时编码空时和运动特征。与 STM 直接将时空特征和运动特征一起编码的方式不同, Li^[60]首先采用运动激励(ME)模块, 利用相邻帧间的运动表达来获得特征中对运动敏感的信息, 之后用多时序聚合(MTA)模块进行长期的时序建模。两个简单且轻量的模块集成到 2D ResNet 中构成时序激励和聚合模块

(TEA)。两个模块对比如图 9 所示。

本文在 UCF101 和 Kinetics 两个基准数据集上对上述基于 3D 网络及其改进版本的深度模型从参数量、计算效率以及分类精度三方面进行比较, 如表 5 所示。在表 5 中, 如果没有特意指出模型采用的骨干网络为 2D ResNet, 则默认 ResNet 骨干网络为 3D ResNet。

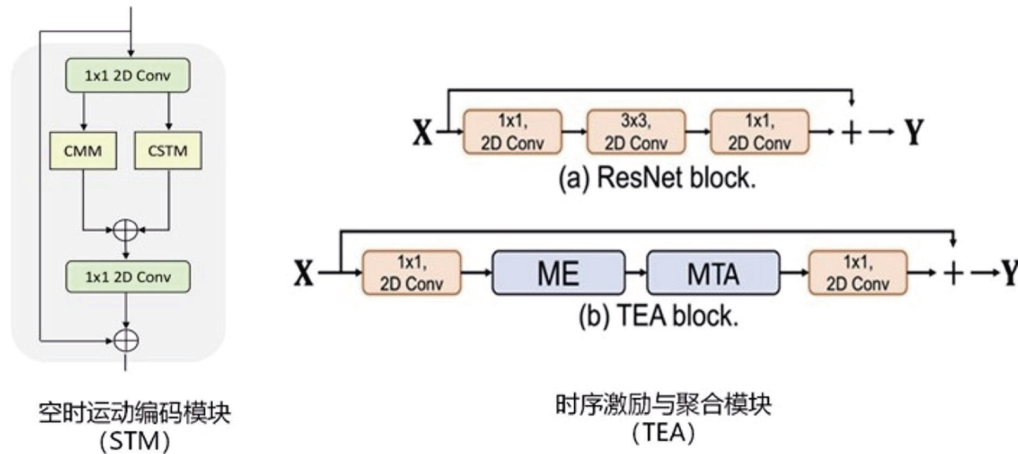


图 9 空时运动编码模块(STM)和时序激励与聚合模块(TEA)对比。图片来自[55-56]

Figure 9 Space-temporal and motion encoding module (STM) and temporal excitation and aggregation module (TEA) are compared. The images are from [55-56]

表 5 基于 3D 网络及其改进版本的深度模型在 UCF101 和 Kinetics 数据集上的性能比较

Table 5 Performance comparison of 3D network-based and its improved version of the depth model on UCF101 and kinetics datasets

方法	骨干网络	训练数据	输入数据	模型大小(M)	GFLOPS	UCF 101	Top1	Top5
			(frames×size×crops×clips)					
C3D ^[21]	INet	Sports-1M	16×112×112×1×all	321	N/A	82.3	N/A	N/A
Res-C3D ^[48]	ResNet-34	Sports-1M	16×112×112×1×10	64.9	38.5	85.8	67.4	87.2
P3D ^[49]	ResNet-152	ImageNet+Sports-1M	16×160×160×1×20	261	N/A	88.5	N/A	N/A
I3D-RGB ^[22]	Inception	ImageNet+Kinetics	All×224×224×1×1	12.7	544	95.6	71.7	89.3
R(2+1)D-RGB ^[51]	ResNet-34	Sports-1M+Kinetics	32×224×224×1×10	63.6	152.4	96.8	74.3	91.4
S3D ^[52]	Inception	ImageNet+Kinetics	64×224×224×1×1	8.77	66.38	N/A	72.2	90.06
S3D-G-RGB ^[52]	Inception	ImageNet+Kinetics	64×224×224×1×1	11.56	71.38	96.8	74.7	93.4
ip-CSN-152 ^[53]	ResNet-152	Kinetics	32×224×224×3×10	32.8	108.8	N/A	78.8	93.5
Non-localI3D ^[55]	ResNet-50	ImageNet+Kinetics	128×224×224×3×10	35.33	282	N/A	76.5	92.6
StNet ^[56]	2D ResNet-50	ImageNet+Kinetics	(25×15)×256×256×1×1	33.16	189.29	N/A	69.85	N/A
SlowFast ^[57]	ResNet-50	Kinetics	(8+64)×224×224×3×10	N/A	65.7	N/A	77	92.6
TPN ^[58]	ResNet-50	Kinetics	32×256×256×3×10	N/A	213	N/A	78.9	93.5
STM ^[59]	2D ResNet-50	ImageNet+Kinetics	16×244×244×3×10	24	66.5	96.2	73.7	91.6
TEA ^[60]	2D ResNet-50	ImageNet+Kinetics	16×244×244×3×10	N/A	70	96.9	76.1	92.5

模型的计算效率和参数量, 不仅与模型本身的性能有关, 也与输入视频数据的处理有关。在进行视频人体行为识别任务时, 一般要对输入的原始视频进行两个操作: (1) 将原始视频采样成包含 T 个连续帧的视频片段(clips); (2) 利用随机缩放和角裁剪(crop)进行数据增强, 并固定每帧的裁剪区域大小。最终模型输入视频数据的形式为: 帧数 $T \times$ 每帧裁剪区域大小 $\text{Size} \times$ 每帧的裁剪数目 $\text{crops} \times$ 视频采样的片段数 clips 。本章节按照文献[50][56], 评估各个模型时采用两种评估协议: (1) 效率协议。在评估模型计

算效率时, 基于 1 个连续 T 帧的视频片段(clips)并且每帧仅进行一次裁剪(crop)。(2) 精度协议。从视频中随机采样 N 个不同的片段(clips), 平均所有剪辑的得分来得到最终的预测结果。

3.4 小结

本文在前面三个小节中分别介绍了基于双流卷积网络, 循环神经网络(LSTM/RNN)以及三维卷积网络的深度学习模型, 这些模型都成功应用于从 RGB 视频数据中识别人体行为。三种模型架构示意图如图 10 所示。

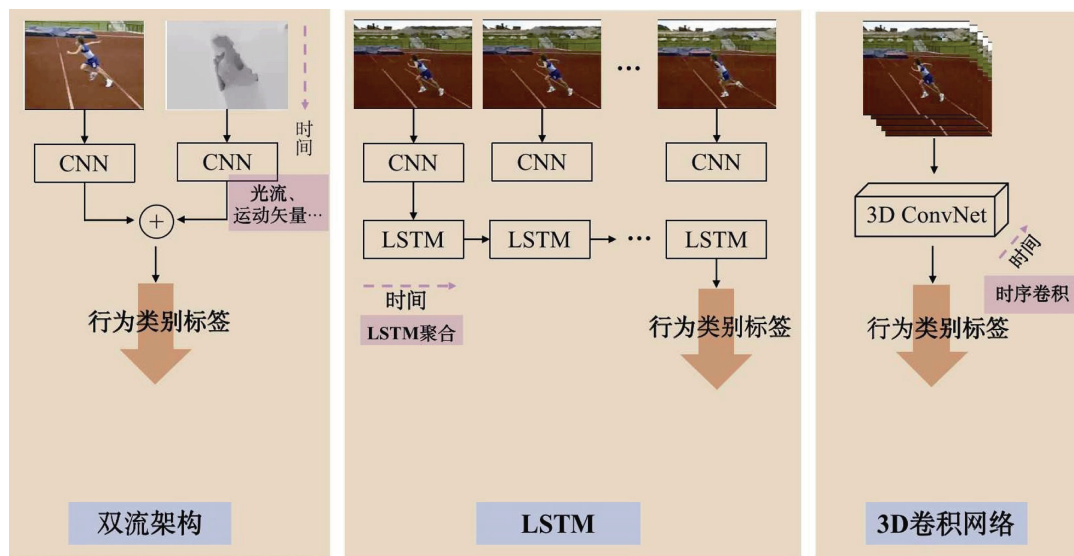


图 10 双流法、LSTM 以及 3D 卷积网络的模型架构示意图

Figure 10 The diagram of model architectures of two-stream method, LSTM and 3D convolutional network

针对这三类深度模型的研究, 研究者不断地克服视频分析和理解中面临的两个巨大挑战。第一个挑战在于如何更有效地获得视频帧上具有判别力的外观信息。上述深度行为识别模型主要依赖 CNN 强大的提取图像外观特征的能力, 并采取选择关键视频帧, 增加注意力机制等方式从空间上获取视频帧中的行为信息。另一个更大的挑战是获得时间维度上视频行为序列的信息。LSTM 能够对更长时间的视频序列建模, 并很好地学习其时序动态变化, 3D CNN 则利用时间维度上的卷积运算来获得视频时序特征表达。

结合表 3, 表 4 和表 5 中的深度模型性能对比可以得出, 由于表达与聚合空间特征的能力十分有限, 基于 LSTM/RNN 的深度模型即使结合计算开销巨大的光流, 其识别精度也远远不如其余两类方法。因此虽然 LSTM 允许堆叠更多的视频帧序列来建模长期的时序关系, 但这类深度模型并不是近几年 RGB 视频行为识别研究的关注重点。除此以外, 从这类模型

的复杂度来看, 其在计算效率上没有任何优势。因此本文在表 4 中仅从行为识别精度这一方面比较了基于 LSTM/RNN 的深度模型的性能, 没有具体体现这类模型的推理速度。

2017 年之前, 以 TSN 为代表的双流卷积网络在 UCF101 行为识别数据集上实现了很高的分类精度, 激发了研究者对双流架构的广泛关注, 与之相比三维卷积网络因模型参数量大、识别精度低、可供训练的视频数据少等因素的限制而发展十分缓慢。得益于 ImageNet 图片数据集上的预训练参数在三维卷积模型上的应用以及大规模视频行为数据集 Kinetics 的提出, 深度三维卷积网络学习视频数据表达的优势得以显现。除此以外, 2017 年之前光流普遍被认为是一种很好的表征人体运动的方式, Facebook 实验室 2017 年的研究工作^[61]提出了对光流不同的见解, 发现光流计算总能提高视频行为识别的精度是因为光流图像包含了视频图像帧的外观不变信息。光流的局限性在于它并不是一种有效的表征运动的方式

且计算成本很高,这也促使研究者的关注度逐渐转向三维卷积网络。

随着图像任务中应用的深度学习模型的不断发展,在视频中应用的深度学习模型沿着一致的发展方向不断提高模型性能。一方面是模型深度不断加深,从最初仅有 11 层的浅层网络 C3D^[19],发展为在残差框架下实现的 199 层的 P3D^[45]和 152 层的

CSN^[53],模型深度的不断加深在很大程度上依赖于可用于训练的公开视频数据量的增加。一方面是深度模型的轻量化并寻求速度和精度的平衡,从模型参数量不到 9M 的 S3D^[52]。视频人体行为识别的研究总是选择在图像任务中应用很成功的深度模型作为骨干网络(如图 11),最常用的则是残差网络 ResNet^[47]和 Inception^[50]。

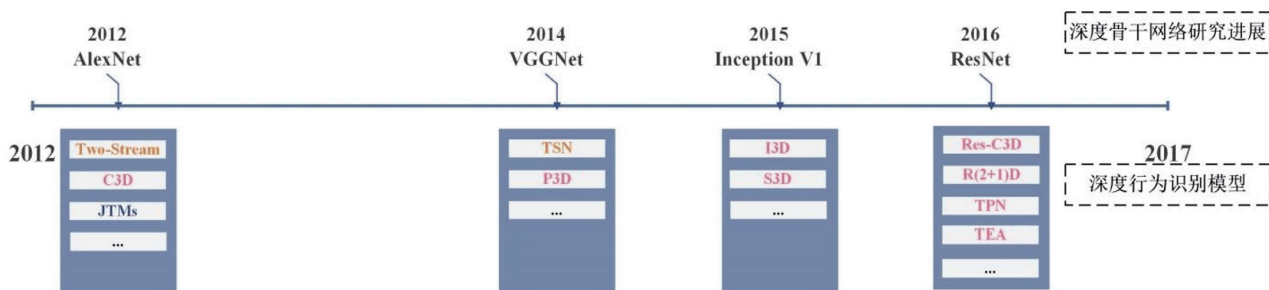


图 11 双流法与 3D 卷积网络的时序演进图

Figure 11 Timelines of two-stream methods, 3D convolutional networks

对包含空间和时间成分的三维视频数据而言,时间维度的信息对理解视频中的人体行为来说至关重要。三维卷积相比二维卷积在学习视频时空特征的表达上也更具优势。然而,虽然 3D CNN 可直接对 RGB 视频输入的时空信息建模,但很多方法,如 I3D^[22],R(2+1)D^[51]仍然在集成光流之后,能够进一步提高性能。为此,研究者也在不断探索除了光流之外,能够更好地表达运动信息的方式^[59-60],并将空时特征和时间运动特征整合到 2D CNN 框架中,这种方式简单且有效。

4 基于人体骨骼数据的行为识别

人体骨骼数据是由人体关键点的位置组成的,人体关键点有 2D 和 3D 两种位置坐标表示方式,是更抽象的人体表达。依赖于实时且可靠的人体关键点检测算法,人体骨骼数据可通过 Kinect 这样的深度传感器获得,也可从 RGB 视频中获得。骨骼数据能很好地克服 RGB 视频数据中存在的复杂背景、光照变化、人体外貌变化以及运动速度等与行为无关的不确定因素的影响,与深度数据相比具有更好的抗噪能力。近两年来,利用人体骨骼数据进行视频人体行为识别的研究工作越来越多并受到广泛关注。这类研究工作与基于 RGB 视频数据的研究类似,同样可分为三类,基于 LSTM 的循环神经网络,基于卷积神经网络的,和基于图卷积神经网络(Graph Convolutional Network, GCN)的。

4.1 基于 LSTM 网络的骨骼行为识别

早期基于骨骼数据的行为识别模型通常是将一

系列骨骼数据看作时间序列进行处理,试图提取人体姿态的静态特征及其连续的时间动态来表示人体动作。Du 等人^[62]提出一个端到端的循环神经网络 HBRNN,将特征学习和序列分类集成到一个框架中。该模型将人体骨骼分为五个部分,分别送入五个双向的递归神经网络(BRNNs)进行局部特征提取,后续分层次地将局部特征融合为全局特征用于行为分类,这样逐层特征融合的方式显式地将空时结构信息编码为高层次的表达。Shahroudy 等人^[25]将整个身体的时序变化划分为身体部位的变化,提出了部位感知的 LSTM 模型(Part-aware LSTM),该模型独立地学习各个身体部位间共同的时序变化模式,整合成全局表达用于行为识别。Liu 等人^[63]将传统的 LSTM 扩展到空时域,每个身体关键点从邻近的关键点和先前的视频帧中接收上下文信息来编码空时上下文,并添加新的门控机制分析输入数据的可靠性。Li 等人^[64]在 2017 年提出全局上下文感知注意力网络(GCA-LSTM),借助全局上下文信息选择性地关注包含更多判别信息的身体关键点。Li 等人^[65]2018 年在 GCA-LSTM 网络的基础上增加反复注意力迭代机制,对全局上下文记忆单元进行逐步优化来提升模型的注意力性能。除此以外模型进一步扩展为双流架构,分别整合细粒度的身体关键级别级的注意力和粗粒度的身体部位级别级的注意力。

微软亚洲研究院的兰翠玲等人对如何利用 LSTM 学习人体骨骼序列中的行为信息做了深入的研究^[66-69]。Zhu 等人^[66]提出的深度 LSTM 网络模型有 3 个 LSTM 层和两个前馈层,LSTM 层和前馈层交

替使用, 模型的全连接层和 LSTM 神经元的部分运用了两种不同类型的强正则化方式, 减轻过拟合并使网络鲁棒地学习不同骨骼关键点间复杂的共现关系和运动动态。Song 等人^[67]提出一种具有时空注意力机制的多层 LSTM 网络用于行为识别。模型中的空间注意力模块自动选择每一帧的主要关键点, 时序注意力模块给不同的帧分配不同程度的重要性。Zhang 等人^[68]没有采用人类定义的标准处理三维骨骼数据, 而是提出一个新的视角自适应方案来解决骨骼视角变化, 通过自动调整每一帧的观察点, 以获得新视角下的骨骼表示。而后 Zhang 等人^[69]在此工作的基础上, 融合

RNN 视角自适应网络与 CNN 视角自适应网络, 进一步提高了模型的性能。这验证了视角自适应方案在循环神经网络与卷积神经网络上的有效性。

4.2 基于 CNN 的骨骼行为识别

LSTM 因其强大的时间序列建模能力, 在骨骼行为识别任务上取得了不错的效果, 但这种结构无法有效地学习骨骼关键点间的空间关系。CNN 具有很好的学习结构信息的能力, 因而有很多研究者使用 CNN 挖掘骨骼关键点间复杂的空间关系。这种方法大多采用将骨骼序列编码为伪图像, 即二维矩阵的方式, 如图 12。

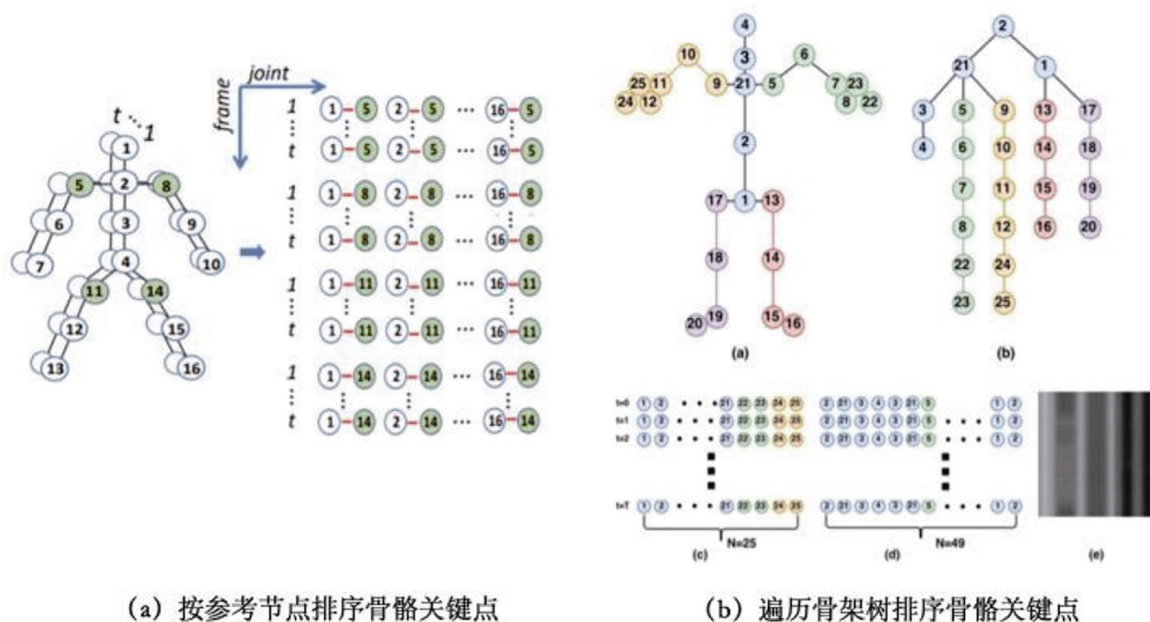


图 12 骨骼序列表示成伪图像(图来自[74][75])

Figure 12 Skeletal sequences are represented as pseudo-images

Du 等人^[70]将每个骨骼序列表示为二维矩阵的形式, 连接每个骨骼关键点坐标构成矩阵的行, 沿着时间顺序堆叠每一帧的向量构成矩阵的列, 这样的时空编码图量化成一个图像后送入 CNN 网络进行行为预测。Wang 等人^[71-72]采用颜色编码将关节轨迹的空间结构和动力学特征表示为三种纹理图像, 即时空编码图量化成一个图像后送入 CNN 网络进行行为预测。Wang 等人^[71-72]采用颜色编码将关节轨迹的空间结构和动力学特征表示为三种纹理图像, 即关节轨迹图(Joint Trajectory Maps, JTM), 然后送入三个微调后的 AlexNet CNNs^[73]进行行为分类。Ke 等人^[74]将骨骼序列表示为 3 个片段, 每个片段包含 4 种不同的骨骼关键点的空间结构信息。为获得丰富的骨骼关键点的空间关系, Ke 分别选择左肩、右肩、左髋关节、右髋关节这 4 个在身体运动中保持相对稳定的

身体关键点作为参考关键点, 通过计算身体关键点到参考关键点的相对位置, 将骨骼序列的每一帧转换为四种不同的关键点空间关系, 从而获得更为紧凑的骨骼关键点表达。作者通过多任务学习的方式对所有时间步长的特征向量进行并行处理, 并用于人体的行为识别。该工作在一定程度上缓解了人体骨骼序列产生的稀疏数据的问题。Liu 等人^[75]为消除视角变换对骨骼关键点坐标位置的影响, 采用基于序列的视角不变变换处理骨骼数据。转换后的骨骼图像可视化为彩色图像后进行视觉增强和运动增强, 并将其输入到多分支的 CNN 框架下进行人体行为识别。Li^[76-77]为整合骨骼运动特征, 将骨骼序列的二维图像表达方式以及连续两帧间的关键点坐标差输入到双流网络中。Li^[77]利用分层特征学习的方式, 将所有身体关键点的点集特征逐步聚合为全局特征, 从

而获得身体不同关键点间的相关性。

将 LSTM 用于骨骼行为识别的最大障碍之一是所学习的模型参数与其隐藏表示间缺乏清晰的联系。Kim 等人^[78]在三维人类行为分析的背景下, 通过使用残差连接重新阐述 TCN 模型, 提供易于解释的隐藏表示和模型参数。Yang 等人^[79]提出一种树结构骨架图(Tree Structure Skeleton Image, TSSI)表示, 可保留更丰富的骨骼结构信息, 避免了将所有关键点按固定顺序连接起来导致的结构信息丢失。具体方法用深度一阶算法遍历一棵骨架树, 边越少, 连接的两个身体关键点的相关性越强。将生成的骨骼关键点表示量化成图像后, 输入 ResNet-50 深度模型。Caetano^[80]结合了参考关键点^[74]和树结构骨架^[79]的优点, 获得了良好的性能。

4.3 基于 GCN 的骨骼行为识别

CNN 的卷积核具有参数共享机制和加权平均机制, 卷积核参数在模型训练时迭代更新的过程, 也是卷积核提取图像像素矩阵中不同的形状和空间特征的过程。CNN 的卷积运算仅处理具有规则网格结构数据。GCN 将卷积扩展到人体骨骼等任意结构的数据。近些年来, GCN 越来越受关注并应用到图像分类文件分类, 半监督学习以及行为识别等任务中。现有的关于图卷积的研究可分为两类^[81]。第一种是空域图卷积^[5,7-10,82,83]。对每个节点, 根据其与其他节点的连接构造子图。然后聚合子图中每个节点的值来进行卷积。这种方法的挑战性在于正确选择节点的排序和处理子图间的基数差异。第二种是频域图卷积^[6,84-87]。该方法的关键是计算图的傅里叶变换。

香港中文大学 Yan 等人^[5]根据人体骨骼关键点序列构建空时图, 如图 13 所示, 然后送入到 ST-GCN 模型中通过空时图卷积运算进行行为识别。这是首次将 GCN 用于人体骨骼行为识别的研究工作。由于人体骨骼本来是以图的形式存在, ST-GCN 提供了一个使用 GCN 直接从人体骨骼序列图中获得空间特征以及时序动态的新方法, 验证了图卷积建模人体运动模式的有效性。Tang 等人^[6]采用深度渐进强化学习方法从输入人体骨骼序列中提取一定数量的关键帧, 并使用频域图卷积获得骨骼关键点间的依赖关系。考虑到 ST-GCN 仅能从骨骼图中获得直接相连的身体关键点的局部特征, Li 等人^[7]构建行为链式推理模块(AIM)来获得身体关键点间更丰富的依赖关系, 实验证明在行为识别任务前端并行增加姿态预测模块有助于保留更细节的动作信息, 实现更好的行为识别性能。Li 等人^[8]2019 年进一步提出了共生模型 Sym-GNN, 在多任务学习的框架下, 同时进行

行为识别与行为预测任务, 并采用多分支多尺度的图卷积神经网络提取空间和时间特征, 该模型在两个任务上都获得了较好的性能。Li 等人^[9]则是考虑利用骨骼关键点的位置和运动信息学习动态的图拓扑结构, 从而能够自适应地学习骨骼关键点间的高阶连通性。类似地, Shi 等人^[10]为构造自适应的图卷积网络, 对可表示数据共同模式的全局图以及表示单个数据模式的单个图进行参数化, 这两类图结构随模型的卷积参数一起训练并更新, 同时采用将骨骼关键点间的一阶信息和二阶信息融合的双流框架。同年, Shi 等人^[82]也尝试充分利用关键点和骨骼间的关系, 将人体骨骼表示为有向无环图, 并且在训练过程中通过学习图的拓扑结构来构造自适应图。Ga 等人^[84]提出了基于图回归的图卷积网络(GR-GCN), 通过学习数据的底层图结构增强图的表达能力。GR-GCN 考虑每个身体关键点与同一坐标系中的相邻关键点、前后坐标系中的相关关键点的相关性。

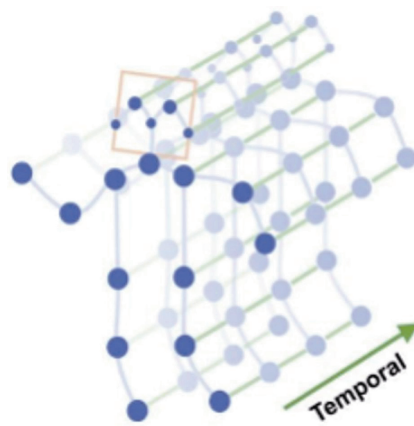


图 13 人体骨骼空时图(图来自[5])

Figure 13 Human skeletal spatial-temporal graph

由于人体运动模式的复杂性, 从骨骼数据中识别人体行为具有很大的挑战。为进一步增强模型的泛化能力, 获得更具判别的空时特征, 近些年来研究者不断地将图卷积进行扩展。Si 等人^[83]提出注意力增强的图卷积 LSTM 网络模型 AGC-LSTM, 该模型可从人体骨骼序列中学习具有判别力的空时特征以及骨骼关键点间在空间和时间上的共现关系, 在预测行为标签时考虑了骨骼关键点的全局和局部特征。实际上, 不同的人做同一个动作, 姿态也会很不相同。这样显著的类内差异使类间差异变得模糊。Zhao 等人^[85]利用图卷积获得身体关键点间的相互依赖关系, 利用长短期记忆网络(LSTM)获得身体姿态的时序变化, 最后将该神经网络遵循贝叶斯框架进一步扩展为概率模型, 将模型参数视为随机变量。这

种扩展使模型更好地处理数据变化。Zhang 等人^[86]首次尝试将骨骼关键点的高级语义(如帧索引和关键点类型)用于骨骼行为识别;提出了轻量级的模型SGN。该模型将图卷积层和卷积层结合,可分层次地利用关键点级和帧级的时空相关性。

最近,Li 等人^[88]和Liu 等人^[87]的研究工作证明了考虑多尺度的骨骼关节间的依赖关系可以有效学习鲁棒的人体运动模式。Li 等人^[88]将人体骨骼表示为不同级别的图拓扑结构并以金字塔的形式输入到网络中,从而构建不同尺度的骨骼关节间的运动关联性。Liu 等人^[87]提出了统一的多尺度空时卷积网络MS-G3D,采用多路由的方式同时建模复杂的空时域间骨骼关键点的依赖关系并进行长期、多尺度的特征聚合,从而从人体骨骼序列中获得鲁棒的运动特征。

4.4 小结

近些年来将人体骨骼数据用于行为识别的研究工作备受关注,原因有以下几点:(1) 包含人体骨骼关键点 3D 坐标的骨骼数据是高级抽象表达。从生物学上讲,人类可通过观察身体关节运动来识别动作类别,不需要任何外观信息;(2) 低成本的深度摄像机和骨骼关键点估计技术的发展使获得人体骨骼数据更为容易;(3) 与 RGB 视频相比,骨架表示不包含环境噪声(例如背景杂乱、照明条件、服装等),对视角和外观的变化具有较好的鲁棒性;(4) 低维的特征表达在计算上是更高效的。目前,基于骨骼数据的行为识别模型按照骨骼关键点的表达方式不同大

致可分为三类,基于 LSTM 方法将人体骨骼关键点表示成时间序列的形式,基于 CNN 的方法直接根据人体骨骼关键点的自然连接将其表示成伪图像,基于图卷积的方法则将人体骨骼表示成图的形式。这三类研究方法的关注重点都在于如何从人体骨骼序列中学习到有判别力的空间和时间特征,如图 14 所示。

本文分别在 Kinetics-skeleton, NTU-RGBD 60, NTU-RGBD 120, 以及 SYSU 四个行为识别的基准数据集上评估并对比基于人体骨骼数据的行为识别算法,如表 6。

基于 LSTM 的循环神经网络最早被应用于人体骨骼行为识别任务中。直观理解是,利用序列的时间信息可很好地帮助理解人体姿态的运动。采用这种遍历规则,将人体骨骼数据表示为时间上的骨骼关键点序列。注意力机制常被应用于该网络来选择关注那些包含更多判别力信息的骨骼关键点、身体部位或更重要的某一帧人体骨骼。然而,虽然 LSTM 能够建模长期的骨骼关键点间的时序依赖,但 CNN 在学习人体骨骼关键点的高级特征、提取更丰富的骨骼关键点间空间结构信息方面更具优势。基于 CNN 的骨骼行为识别算法将原始的人体骨骼序列编码为伪图像后用于行为识别。研究者探索了不同人体骨骼关键点的特征编码方式,如骨骼关键点轨迹图、相邻帧间的骨骼关键点坐标差、选择参考关键点、树结构骨架图等,期望获得丰富且可靠的骨骼关键点间的空间关系。

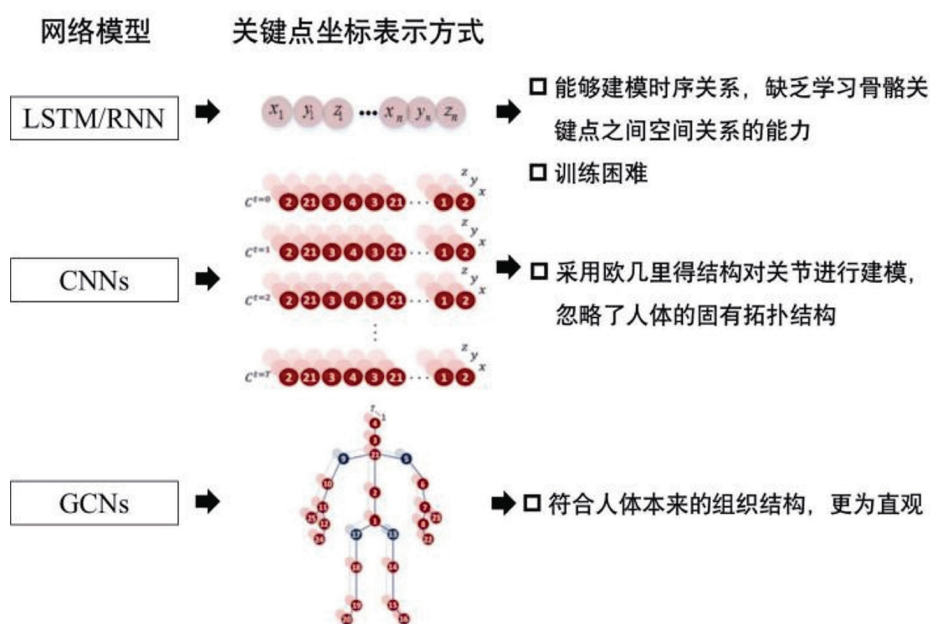


图 14 不同模型对人体骨骼关键点坐标的不同表示方式示意图

Figure 14 Diagram of the different representations of joint coordinates of human skeletons in different models

表 6 基于人体骨骼的行为识别模型的性能对比

Table 6 Performance comparison of human skeleton-based activity recognition models

年份	模型	方法	SYSU		Kinetics-skeleton		NTU-RGBD 60		NTU-RGBD 120	
			setting-1	setting-2	Top-1	Top-5	CS	CV	CS	CV
2016	HBRNN[62]	RNN	—	—	—	—	59.1	64	—	—
2016	Part-aware LSTM[25]	RNN	—	—	16.4	35.3	62.9	70.3	25.5	26.3
2016	ST-LSTM+Trust Gate[63]	RNN	—	76.5	—	—	69.2	77.7	55.7	57.9
2017	GCA-LSTM[64]	RNN	—	77.8	—	—	74.4	82.8	58.3	59.2
2017	STA-LSTM[66]	RNN	—	—	—	—	73.4	81.2	—	—
2017	VA-LSTM[67]	RNN	76.9	77.5	—	—	79.8	88.9	—	—
2018	Two-stream GCA-LSTM[65]	RNN	—	79.1	—	—	76.1	84	61.2	63.3
2019	VA-Fusion[68]	RNN	86.7	86.2	—	—	89.4	95	—	—
2015	Frames+CNN[70]	CNN	—	—	—	—	68.7	73	—	—
2016	JTMS[71]	CNN	—	—	—	—	73.4	75.2	—	—
2017	Clips+CNN+MTLN[74]	CNN	—	—	—	—	79.6	84.8	58.4	57.9
2017	Enhanced Visualization[75]	CNN	—	—	—	—	80	87.2	60.3	63.2
2017	Two-stream CNN[76]	CNN	—	—	—	—	83.2	89.3	—	—
2017	TCN[78]	CNN	—	—	20.3	40	74.3	83.1	—	—
2018	HCN[77]	CNN	—	—	—	—	86.5	91.1	—	—
2018	TSSI+Resnet50[79]	CNN	—	—	—	—	73.1	76.5	—	—
2018	TSSI + GLAN + SSAN[79]	CNN	—	—	—	—	82.4	89.1	—	—
2019	TSRJI [80]	CNN	—	—	—	—	73.3	80.3	65.5	59.7
2018	ST-GCN[5]	GCN	—	—	30.7	52.8	81.5	88.3	72.4	71.3
2018	DPRL+GCNN[6]	GCN	—	76.9	—	—	83.5	89.8	—	—
2019	AS-GCN[7]	GCN	—	—	34.8	56.3	86.8	94.2	77.7	78.9
2019	Sym-GNN[8]	GCN	—	—	37.2	58.1	90.1	96.4	—	—
2019	STGR-GCN[9]	GCN	80.7	81.9	33.6	56.1	86.9	92.3	—	—
2019	2s-AGCN[10]	GCN	—	—	36.1	58.7	88.5	95.1	—	—
2019	DGNN[82]	GCN	—	—	36.9	56.9	89.9	96.1	—	—
2019	AGC-LSTM[83]	GCN+LSTM	—	—	—	—	89.2	95	—	—
2019	GR-GCN[84]	GCN+CNN	—	77.9	—	—	87.5	94.3	—	—
2019	BNN[85]	GCN+LSTM	—	82	—	—	81.8	89.0	—	—
2020	SGN[86]	GCN+CNN	89.3	90.6	—	—	89	94.5	79.2	81.5
2020	MS-G3D[88]	GCN	—	—	38.0	60.9	91.5	96.2	86.9	88.4

LSTM 与 CNN 在处理人体骨骼数据时首先要对其进行网格化处理, 无法避免三维空时信息的丢失, 并且利用深度模型直接从这样的网格数据中获得的骨骼关键点间长期的时序依赖与空间关系显然是不准确的。鉴于深度学习从欧氏空间数据中提取潜在特征的有效性, GCN 逐渐成为新的研究热点, 它能够泛化卷积神经网络到非欧氏结构的数据上。人体骨骼本来是以图的形式存在的, 骨骼关键点是顶点, 骨骼是边, 利用图卷积网络进行人体骨骼行为识别时能够充分利用人体骨骼数据的拓扑结构, 不再需要依赖任何手动的部分或者事先设定某种规则来分析数据的空间模式。

近两年, 有大量的研究工作使用 GCN 从骨骼空时图中学习人体运动模式。同样, 为获得鲁棒的人体运动特征, 理想的情况是可从骨骼空时图中获得局部连接和空间分离的骨骼关键点间局部与全局相关性, 以及不同时间尺度上的时序依赖关系。仅从原始的人体骨骼图中学习到的高级特征表达可能包含的判别信息有限。为充分利用图的拓扑信息, 获得更丰富的骨骼关键点间的相关性, 现有的大多基于图卷积的方法主要采用两种方式: (1) 添加额外的信息,

如骨骼信息、骨骼关键点运动信息、骨骼关键点的语义信息、增加帧间的骨骼关键点连接等; (2) 改变图卷积的邻接矩阵, 如自适应学习邻接矩阵参数, 利用高阶多项式的邻接矩阵等。由于人体行为的持续时间不同, 获得短期与长期的时序依赖关系是另一个研究重点。目前大多数图卷积采用的时序建模方式是采用 1D 时序卷积模块建模时间动态, 并采用空域时域交错的模块。这样的方式没有办法建立骨骼关键点的跨时空联系。在最近的方法中^[9,88], 研究者开始关注利用跨空时的关键点间的相关性。

5 基于 RGB+D 多模态数据的行为识别

基于 RGB 视频的行为识别已获得广泛的研究。其优异的性能表现在视频人体行为识别领域占据着主导地位。但这种方法只是视频到行为类别的映射, 无法区分人体和视频背景, 缺乏对人体运动的明确建模。骨骼数据来自 RGB 或深度视频中, 提供了简洁且信息丰富的人体动作表示。基于骨骼的行为识别展现了良好的性能, 成为研究者关注的模式。然而, 由于现实生活中人体运动动力学的复杂性, 视频的人体行为识别仍然具有一定的挑战性。

人体运动的复杂性主要体现在三方面,一是人体运动往往涉及不同身体部位的协调,即不同身体部位间存在时空依赖性;二是运动模式中存在长期的时序依赖性;三是不同个体由于年龄、习惯、体重等主体因素的不同,使得不同个体的动作模式存在显著变化。因而不同个体即使进行相同的动作,在运动的程度和速度上也会有很大的不同。总之,空间、时间和主体因素都增加了视频人体行为识别的挑战,也限制了实际场景中从单个数据模态中提取特征来进行人体行为识别的方法应用。骨骼数据可鲁棒地抵抗杂乱背景和与动作无关的运动这两方面的干扰,但也缺失了 RGB 视频帧中包含的细节外观纹理特征,无法利用视频中出现的大部分上下文线索和全局线索。除此以外,尽管二维姿态估计有了显著的进步,但在深度视频中其性能仍然不如三维姿态估计。由于实际应用中复杂的背景和人体部位的自遮挡,估计得到的人体骨骼数据不完全可靠,可能会误解人体形态。

近些年来有不少研究者关注如何从不同的数据模态中提取并融合相关模态的信息,以实现更好的性能。Shahroudy 等人^[89]使用级联因数分解层来寻找各个模态间的共享表达,并孤立特定模态的信息。人体行为可从不同的数据模式和不同的动作执行阶段中识别。Hu 等人^[90]采用双线性池化模块整合不同维度的输入模态-时间张量,从而能够从不同的数据模态和动作历史序列的长度中学习特征。动作历史序列的长度即表示不同的动作执行阶段。Liu 等人^[91]将人体姿态估计图聚合为热图和人体骨骼,并在双流框架中使用卷积网络提取特征,融合后的特征用于最终预测,实验证明热图包含了能够反映人体形态的丰富信息。Luvizon 等人^[92]在多任务框架下,同时进行二维/三维的人体姿态估计与行为识别,能够利用人体骨骼关键点相对时间变化的语义信息和单帧及身体部位概率图上的视觉信息(如图 15 所示)。最近,Joze 等人^[93]探索了不同空间维度的多模态卷积层特征融合,实现了 RGB 和骨骼两种数据模态的缓慢融合。

本文在 NTU RGB+D 60 数据集上展现了上述基于 RGB+D 多模态数据的行为识别方法性能,如表 7。基于 RGB+D 多模态数据的方法在 NTU RGB+D 行为数据集上获得了不错的结果。然而与骨骼行为识别的方法相比,其识别精度仍有一定差距。实际上对基于多模态数据的方法来说,如何找到一种合适的融合不同模态数据的方法是具有挑战性的问题。多模态数据融合在输入级(即早期融合)、决策

级(即晚期融合)或中间均可实现。晚期融合仍是多模态学习的主要方法,其原因为可使用一个简单的池化运算或一个注意力机制来融合每个流的一维预测分数。然而,不同模态不同层次的特征包含有关输入数据的不同信息,找到合适的在中间层融合不同模态数据的方法,是研究者关注并仍需不断探索的问题^[89,93]。

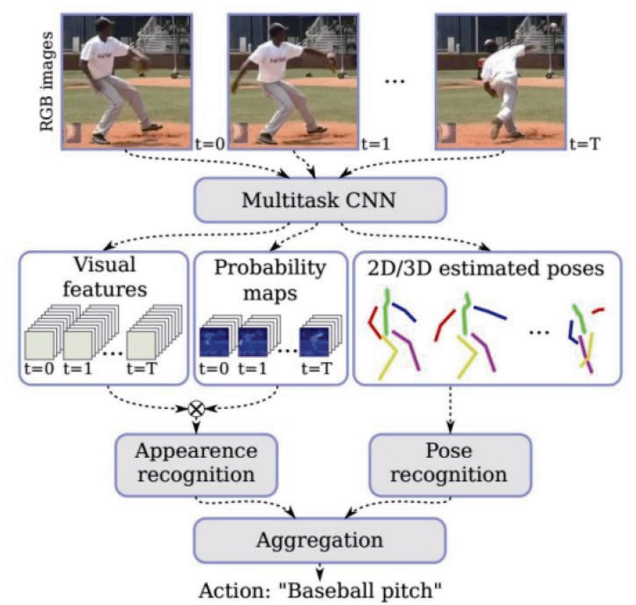


图 15 姿态估计与行为识别的多任务学习框架图(图来自^[88])

Figure 15 Multitask learning framework diagram for pose estimation and activity recognition (Figure from ^[88])

表 7 基于 RGB+D 多模态数据的行为识别模型精度
Table 7 Accuracy of activity recognition model based on RGB+D multimodal data

年份	模型	输入数据模态	NTU RGB+D 60	
			CS	CV
2017	Deep multi modal ^[89]	RGB+Depth+Skeleton	74.90	—
2018	Deep Bilinear ^[90]	RGB+Depth+Skeleton	85.40	90.70
2018	Posemap ^[91]	RGB+Skeleton	91.71	95.25
2018	2D/3D Multitask ^[92]	RGB+Skeleton	85.55	—
2020	MMTM ^[93]	RGB+Skeleton	91.99	—

6 挑战及下一步研究方向

从表 5, 表 6, 表 7 的结果来看, 目前先进的深度行为识别模型在大型的行为识别基准数据集上均实现了不错的精度。然而现实情况是这些深度学习模型如果应用到真实的视频监控场景中, 还远远无法

实现人们期望的效果。基于多模态视频数据的行为识别方法的不足可概括如下:

(1) RGB 视频图像帧包含的颜色、纹理与形状信息可提供丰富的与人体行为相关的外观特征。但从 RGB 视频中难以显式地建模人体运动变化。因此, 从 RGB 视频数据中识别人体行为的深度模型应用于多人的复杂场景时, 难以区分视频中不同的个体和不同的行为。

(2) 基于人体骨骼的行为识别方法由于缺乏视频帧的细节纹理特征, 无法利用视频帧的全局与局部信息。而仅从人体骨骼关键点中获得的信息有限。

(3) 基于多模态数据的行为识别方法仍需要不断探索合适的数据融合方式来提高模型视觉感知与运动表征的能力。

人体行为识别方法在智能视频监控系统中的应用仍面临很大的挑战。

(1) 在实际的监控视频场景中, 基于 RGB 视频的深度行为识别模型易受杂乱背景、光照变化等各种环境因素的影响。环境中其他对象引起的相互遮挡和人本身的自遮挡也是真实视频中容易出现的问题, 这样的遮挡会使人体骨骼关键点的估计失效。

(2) 实际生活中人体骨骼的动作姿态存在显著的类内差异, 导致类间差异模糊, 同样限制了骨骼行为识别方法的性能。

(3) 真实视频流包含了大量长时间的非动作冗余视频段。目前人体行为识别的研究更多是对分割好的视频行为片段进行分类。如何在空间和时间维度上进行明确的人体行为定位, 确定人体动作开始与结束发生的位置, 则是视频时序行为检测任务的研究重点。有研究者将当前表现良好的深度行为识别模型扩展应用于时空维度的行为检测, 但效果和速度都无法达到目前的应用要求。

(4) 目前性能优越的行为识别方法大多依赖强监督的学习方式和大规模的带标签的训练数据。视频数据的标注耗时耗力, 成本极高。如果要识别训练集中没有的行为类别, 模型将需要重新训练。因此强监督的行为识别方法在实际场景中的应用成本和效率都受到了限制。

结合深度学习方法在视频人体行为识别任务中的研究现状, 考虑金融、工业、地产、教育、零售等多个领域中视频监控智能化的应用需求, 下一步的研究重点应该关注于: 结合计算机视觉领域内的知识包括目标检测、人体跟踪等辅助行为识别加强弱监督或无监督的行为识别方法研究, 寻找一个能有效应用于实际场景中的通用行为识别框架。

在多任务学习框架下, 探索不同任务间的相互促进关系, 比如行为预测与行为识别, 姿态估计与行为识别等越来越低成本的视频拍摄设备, 使多模态视频图像的获取越来越便捷化。多模态数据的融合方法研究也是人体行为识别任务的一个重要研究方向。

如何从真实的视频流中实现快速的行为检测并准确识别行为仍是当前极具挑战且亟待解决的一个问题。

参考文献

- [1] Xin T. 信通院发布《人工智能发展白皮书-技术架构篇(2018年)》[J]. *Robot Technique and Application*, 2018(5): 3. (信通院. 信通院发布《人工智能发展白皮书-技术架构篇(2018年)》[J]. *机器人技术与应用*, 2018(5): 3.)
- [2] Wang L M, Xiong Y J, Lin D H, et al. UntrimmedNets for Weakly Supervised Action Recognition and Detection[EB/OL]. 2017: arXiv: 1703.03329. <https://arxiv.org/abs/1703.03329>.
- [3] Su K, Liu X L, Shlizerman E. PREDICT & CLUSTER: Unsupervised Skeleton Based Action Recognition[EB/OL]. 2019: arXiv: 1911.12409. <https://arxiv.org/abs/1911.12409>.
- [4] Cao Z, Simon T, Wei S H, et al. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 1302-1310.
- [5] Yan S J, Xiong Y J, Lin D H. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, 32(1).
- [6] Tang Y S, Tian Y, Lu J W, et al. Deep Progressive Reinforcement Learning for Skeleton-Based Action Recognition[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 5323-5332.
- [7] Li M S, Chen S H, Chen X, et al. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 3590-3598.
- [8] Li M S, Chen S H, Chen X, et al. Symbiotic Graph Neural Networks for 3D Skeleton-Based Human Action Recognition and Motion Prediction[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(6): 3316-3333.
- [9] Li B, Li X, Zhang Z F, et al. Spatio-Temporal Graph Routing for Skeleton-Based Action Recognition[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 8561-8568.
- [10] Shi L, Zhang Y F, Cheng J, et al. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 12018-12027.
- [11] Guo G D, Lai A. A Survey on still Image Based Human Action Recognition[J]. *Pattern Recognition*, 2014, 47(10): 3343-3361.
- [12] Kang S M, Wildes R P. Review of Action Recognition and Detection Methods[EB/OL]. 2016: arXiv: 1610.06906. <https://arxiv.org/abs/1610.06906>.

- [13] Han F, Reily B, Hoff W, et al. Space-Time Representation of People Based on 3D Skeletal Data: A Review[J]. *Computer Vision and Image Understanding*, 2017, 158: 85-105.
- [14] Lo Presti L, La Cascia M. 3D Skeleton-Based Human Action Classification: A Survey[J]. *Pattern Recognition*, 2016, 53: 130-147.
- [15] Hu J F, Wang X H, Zheng W S, et al. RGB-D Action Recognition: Recent Advances and Future Perspectives[J]. *Acta Automatica Sinica*, 2019, 45(5): 829-840.
(胡建芳, 王熊辉, 郑伟诗, 等. RGB-D 行为识别研究进展及展望[J]. *自动化学报*, 2019, 45(5): 829-840.)
- [16] Wang L, Huynh D Q, Koniusz P. A Comparative Review of Recent Kinect-Based Action Recognition Algorithms[J]. *IEEE Transactions on Image Processing*, 2020, 29: 15-28.
- [17] Asadi-Aghbolaghi M, Clapés A, Bellantonio M, et al. A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences[C]. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition*, 2017: 476-483.
- [18] Wang P C, Li W Q, Ogunbona P, et al. RGB-D-Based Human Motion Recognition with Deep Learning: A Survey[J]. *Computer Vision and Image Understanding*, 2018, 171: 118-139.
- [19] Zhang H B, Zhang Y X, Zhong B N, et al. A Comprehensive Survey of Vision-Based Human Action Recognition Methods[J]. *Sensors*, 2019, 19(5): 1005.
- [20] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]. *European conference on computer vision*, 2016: 20-36.
- [21] Tran D, Bourdev L, Fergus R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]. *2015 IEEE International Conference on Computer Vision*, 2016: 4489-4497.
- [22] Carreira J, Zisserman A. Quo Vadis, Action Recognition a New Model and the Kinetics Dataset[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 4724-4733.
- [23] Soomro K, Zamir A R, Shah M. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild[EB/OL]. 2012: arXiv: 1212.0402. <https://arxiv.org/abs/1212.0402>.
- [24] Kay W, Carreira J, Simonyan K, et al. The Kinetics Human Action Video Dataset[EB/OL]. 2017: arXiv: 1705.06950. <https://arxiv.org/abs/1705.06950>.
- [25] Shahroudy A, Liu J, Ng T T, et al. NTU RGB D: A Large Scale Dataset for 3D Human Activity Analysis[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 1010-1019.
- [26] Hu J F, Zheng W S, Lai J H, et al. Jointly Learning Heterogeneous Features for RGB-D Activity Recognition[C]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016: 2186-2200.
- [27] Karpathy A, Toderici G, Shetty S, et al. Large-Scale Video Classification with Convolutional Neural Networks[C]. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 1725-1732.
- [28] Wu Z X, Jiang Y G, Davis L, et al. LSVC2017: Large-Scale Video Classification Challenge[C]. *The 25th ACM international conference on Multimedia*, 2017: 1978-1979.
- [29] Heilbron F C, Escorcia V, Ghanem B, et al. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding[C]. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 961-970.
- [30] ImageNet[EB/OL] [2021-03-11]. <http://www.image-net.org/>.
- [31] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. *Neural Computation*, 1997, 9(8): 1735-1780.
- [32] Wang H, Schmid C. Action Recognition with Improved Trajectories[C]. *2013 IEEE International Conference on Computer Vision*, 2014: 3551-3558.
- [33] Simonyan K, Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos[EB/OL]. 2014: arXiv: 1406.2199. <https://arxiv.org/abs/1406.2199>.
- [34] Feichtenhofer C, Pinz A, Zisserman A. Convolutional Two-Stream Network Fusion for Video Action Recognition[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 1933-1941.
- [35] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[EB/OL]. 2014: arXiv: 1409.1556. <https://arxiv.org/abs/1409.1556>.
- [36] Lan Z Z, Zhu Y, Hauptmann A G, et al. Deep Local Video Feature for Action Recognition[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017: 1219-1225.
- [37] Zhou B L, Andonian A, Oliva A, et al. Temporal Relational Reasoning in Videos[C]. *European Conference on Computer Vision*, 2018: 831-846.
- [38] Goyal R, Kahou S E, Michalski V, et al. The “something something” Video Database for Learning and Evaluating Visual Common Sense[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 5843-5851.
- [39] Zhang B W, Wang L M, Wang Z, et al. Real-Time Action Recognition with Enhanced Motion Vector CNNs[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2718-2726.
- [40] Bengio Y, Simard P, Frasconi P. Learning Long-Term Dependencies with Gradient Descent is Difficult[J]. *IEEE Transactions on Neural Networks*, 1994, 5(2): 157-166.
- [41] Donahue J, Hendricks L A, Guadarrama S, et al. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description[C]. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 2625-2634.
- [42] Ng J Y H, Hausknecht M, Vijayanarasimhan S, et al. Beyond Short Snippets: Deep Networks for Video Classification[C]. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 4694-4702.
- [43] Sharma S, Kiros R, Salakhutdinov R. Action Recognition Using Visual Attention[EB/OL]. 2015: arXiv: 1511.04119. <https://arxiv.org/abs/1511.04119>.
- [44] Li Z Y, Gavriluk K, Gavves E, et al. VideoLSTM Convolves, Attends and Flows for Action Recognition[J]. *Computer Vision and Image Understanding*, 2018, 166: 41-50.
- [45] Sun L, Jia K, Chen K, et al. Lattice Long Short-Term Memory for Human Action Recognition[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 2166-2175.
- [46] Ji S W, Xu W, Yang M, et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(1): 221-231.
- [47] 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*, 2016: 770-778.
- [48] Tran D, Ray J, Shou Z, et al. ConvNet Architecture Search for Spatiotemporal Feature Learning[EB/OL]. 2017: arXiv: 1708.05038. <https://arxiv.org/abs/1708.05038>.
- [49] Qiu Z F, Yao T, Mei T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 5534-5542.
- [50] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[C]. *The 32nd International Conference on International Conference on Machine Learning - Volume 37*, 2015: 448-456.
- [51] Tran D, Wang H, Torresani L, et al. A Closer Look at Spatiotemporal Convolutions for Action Recognition[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 6450-6459.
- [52] Xie S N, Sun C, Huang J, et al. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-Offs in Video Classification[C]. *European Conference on Computer Vision*, 2018: 318-335.
- [53] Tran D, Wang H, Feiszli M, et al. Video Classification with Channel-Separated Convolutional Networks[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2020: 5551-5560.
- [54] Hara K, Kataoka H, Satoh Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 6546-6555.
- [55] Wang X L, Girshick R, Gupta A, et al. Non-Local Neural Networks[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7794-7803.
- [56] He D L, Zhou Z C, Gan C A, et al. StNet: Local and Global Spatial-Temporal Modeling for Action Recognition[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 8401-8408.
- [57] Feichtenhofer C, Fan H Q, Malik J, et al. SlowFast Networks for Video Recognition[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2020: 6201-6210.
- [58] Yang C Y, Xu Y H, Shi J P, et al. Temporal Pyramid Network for Action Recognition[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 588-597.
- [59] Jiang B Y, Wang M M, Gan W H, et al. STM: SpatioTemporal and Motion Encoding for Action Recognition[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2020: 2000-2009.
- [60] Li Y, Ji B, Shi X T, et al. TEA: Temporal Excitation and Aggregation for Action Recognition[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 906-915.
- [61] Sevilla-Lara L, Liao Y Y, Güney F, et al. On the Integration of Optical Flow and Action Recognition[C]. *German Conference on Pattern Recognition*, 2019: 281-297.
- [62] Du Y, Fu Y, Wang L. Representation Learning of Temporal Dynamics for Skeleton-Based Action Recognition[J]. *IEEE Transactions on Image Processing*, 2016, 25(7): 3010-3022.
- [63] Liu J, Shahroudy A, Xu D, et al. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition[C]. *European Conference on Computer Vision*, 2016: 816-833.
- [64] Liu J, Wang G, Hu P, et al. Global Context-Aware Attention LSTM Networks for 3D Action Recognition[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 3671-3680.
- [65] Liu J, Wang G, Duan L Y, et al. Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM Networks[J]. *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, 2018, 27(4): 1586-1599.
- [66] Zhu W T, Lan C L, Xing J L, et al. Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, 30(1).
- [67] Song S J, Lan C L, Xing J L, et al. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data[C]. *The Thirty-First AAAI Conference on Artificial Intelligence*, 2017: 4263-4270.
- [68] Zhang P F, Lan C L, Xing J L, et al. View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 2136-2145.
- [69] Zhang P F, Lan C L, Xing J L, et al. View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(8): 1963-1978.
- [70] Du Y, Fu Y, Wang L. Skeleton Based Action Recognition with Convolutional Neural Network[C]. *2015 3rd IAPR Asian Conference on Pattern Recognition*, 2016: 579-583.
- [71] Wang P C, Li W Q, Li C K, et al. Action Recognition Based on Joint Trajectory Maps with Convolutional Neural Networks[J]. *Knowledge-Based Systems*, 2018, 158: 43-53.
- [72] Wang P C, Li Z Y, Hou Y H, et al. Action Recognition Based on Joint Trajectory Maps Using Convolutional Neural Networks[C]. *The 24th ACM international conference on Multimedia*, 2016: 102-106.
- [73] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. *Communications of the ACM*, 2017, 60(6): 84-90.
- [74] Ke Q H, Bennamoun M, An S J, et al. A New Representation of Skeleton Sequences for 3D Action Recognition[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 4570-4579.
- [75] Liu M Y, Liu H, Chen C. Enhanced Skeleton Visualization for View Invariant Human Action Recognition[J]. *Pattern Recognition*, 2017, 68: 346-362.
- [76] Li C, Zhong Q Y, Xie D, et al. Skeleton-Based Action Recognition with Convolutional Neural Networks[C]. *2017 IEEE International Conference on Multimedia & Expo Workshops*, 2017: 597-600.
- [77] Li C, Zhong Q Y, Xie D, et al. Co-Occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation[EB/OL]. 2018: arXiv: 1804.06055. <https://arxiv.org/abs/1804.06055>.
- [78] Kim T S, Reiter A. Interpretable 3D Human Action Analysis with Temporal Convolutional Networks[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017: 1623-1631.
- [79] Yang Z Y, Li Y C, Yang J C, et al. Action Recognition with Spa-

- tio-Temporal Visual Attention on Skeleton Image Sequences[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, 29(8): 2405-2415.
- [80] Caetano C, Br  mond F, Schwartz W R. Skeleton Image Representation for 3D Action Recognition Based on Tree Structure and Reference Joints[C]. *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images*, 2019: 16-23.
- [81] Bruna J, Zaremba W, Szlam A, et al. Spectral Networks and Locally Connected Networks on Graphs[EB/OL]. 2013: arXiv: 1312.6203. <https://arxiv.org/abs/1312.6203>.
- [82] Shi L, Zhang Y F, Cheng J, et al. Skeleton-Based Action Recognition with Directed Graph Neural Networks[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 7904-7913.
- [83] Si C Y, Chen W T, Wang W, et al. An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 1227-1236.
- [84] Gao X, Hu W, Tang J X, et al. Optimized Skeleton-Based Action Recognition via Sparsified Graph Regression[C]. *The 27th ACM International Conference on Multimedia*, 2019: 601-610.
- [85] Zhao R, Wang K, Su H, et al. Bayesian Graph Convolution LSTM for Skeleton Based Action Recognition[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2020: 6881-6891.
- [86] Zhang P F, Lan C L, Zeng W J, et al. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 1109-1118.
- [87] Li M, Chen S, Zhao Y, et al. Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction[C/OL]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE*, 2020: 211-220[2021-05-31]. <https://ieeexplore.ieee.org/document/9157584/>. DOI:10.1109/CVPR42600.2020.00029.
- [88] Liu Z Y, Zhang H W, Chen Z H, et al. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 140-149.
- [89] Shahroudy A, Ng T T, Gong Y H, et al. Deep Multimodal Feature Analysis for Action Recognition in RGB-D Videos[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(5): 1045-1058.
- [90] Hu J F, Zheng W S, Pan J H, et al. Deep Bilinear Learning for RGB-D Action Recognition[C]. *European Conference on Computer Vision*, 2018: 346-362.
- [91] Liu M Y, Yuan J S. Recognizing Human Actions as the Evolution of Pose Estimation Maps[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 1159-1168.
- [92] Luvizon D C, Picard D, Tabia H. 2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 5137-5146.
- [93] Vaezi Joze H R, Shaban A, Iuzzolino M L, et al. MMTM: Multimodal Transfer Module for CNN Fusion[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 13286-13296.



孙德刚 现任中国科学院信息工程研究所正研级研究员、客座博士生导师, 中国科学院重大科技任务局副局长、中国科学院大学网络安全学院教授。主要从事信息安全保密技术研究。Email: sundegang@iie.ac.cn



白入文 于 2018 年在河北大学电子信息工程学院获得学士学位, 现在中国科学院信息工程研究所攻读博士学位。研究领域为计算机视觉、视频内容分析和人体行为识别。Email: bairuwen@iie.ac.cn



李敏 现任中国科学院信息工程研究所正研级高级工程师、中国科学院大学网络安全学院教授。研究领域为计算机视觉与智能分析、室内目标定位和物联网等。Email: limin@iie.ac.cn



孟博 于 2007 年在北京工业大学自动化工程专业获得硕士学位。目前在北京理工大学光学与光子学院攻读博士研究生。研究领域为计算机断层扫描、光谱 CT 成像和深度学习。Email: bit.meng@live.com



李凌涵 于 2018 年在西安电子科技大学通信工程专业获得学士学位。现在中国科学院信息工程研究所攻读网络空间安全专业硕士研究生。研究领域为基于深度学习的智能视频分析。Email: lilinghan@iie.ac.cn



杨阳 于 2017 年在北京交通大学计算机科学与技术专业获得学士学位。目前在中国科学院信息工程研究所攻读基于深度学习的图像处理技术博士学位。研究领域为物体检测、动作识别和姿势估计。Email: yangyang1995@iie.ac.cn



姜淼 于 2019 年在南京理工大学计算机科学与技术专业获得学士学位。现在中国科学院信息工程研究所攻读基于深度学习的图像处理技术硕士学位。研究领域物体检测和动作识别。Email: jiangmiao@iie.ac.cn



任俊星 于 2014 年在北京航空航天大学通信工程专业获得硕士学位。现任中国科学院信息工程研究所中级工程师。研究领域为信息安全、电磁安全和深度学习。Email: renjunxing@iie.ac.cn



李风发 于 2021 年在西安电子科技大学信息工程专业获得学士学位。现在中国科学院大学计算机应用技术专业攻读硕士学位。研究领域为行为识别。研究兴趣包括目标检测、网络空间安全。Email: 523825936@qq.com



黄子豪 于 2021 年在华北电力大学自动化专业获得学士学位。现在中国科学院大学计算机应用技术专业攻读硕士学位。研究领域为目标检测。研究兴趣包括图像超分辨率。Email: 120171070514@ncepu.edu.cn