信息与控制　2025 年　第 54 卷　第 1 期：1 ~ 27

DOI：10.13976/j.cnki.xk.2024.3091　　CSTR：32166.14.xk.2024.3091

基于人体骨架的动作识别：综述与展望

孟祥璞[1,2]，李　硕[3]，苑明哲[2,3]，王文洪[1,2]，张志佳[1]，宋纯贺[3]，曹飞道[2]

1. 沈阳工业大学人工智能学院，辽宁 沈阳　110870；
2. 广州工业智能研究院，广东 广州　511458；
3. 中国科学院沈阳自动化研究所，辽宁 沈阳　110016

基金项目：国家自然科学基金面上项目(62273337)；中国科学院科技服务网络计划(STS) - 东莞专项(20211600200072)
通信作者：苑明哲，mzyuan@sia.cn　收稿/录用/修回：2024 - 09 - 18/2024 - 11 - 05/2025 - 01 - 03

摘要

人体动作识别在多场景、多任务下具有多样的研究价值，在智能安防、自动驾驶、人机交互等领域存在广泛的应用前景。基于人体骨架的动作识别已进行了广泛研究，但还没有文献系统地整理其发展历程，并剖析更深层次的内在逻辑。本文整理了基于人体骨架的动作识别的主要发展历程，按照技术方法将其整理归纳为循环神经网络、卷积神经网络、图卷积神经网络、Transformer 四大技术路线，并梳理了其不同的发展脉络，分析了两大关键技术点：空间建模与时间建模，指出了构建丰富表征输入信息的方法论；同时讨论了人体骨架模态在多模态融合中对动作识别的重要意义；最后，对人体骨架动作识别技术方法和实际应用进行了展望。

关键词

动作识别
人体骨架
深度学习
图卷积神经网络
Transformer

中图法分类号：TP391

文献标志码：A

Action Recognition Based on Human Skeleton: Review and Prospect

MENG Xiangpu[1,2], LI Shuo[3], YUAN Mingzhe[2,3], WANG Wenhong[1,2], ZHANG Zhijia[1], SONG Chunhe[3], CAO Feidao[2]

1. *College of Artificial Intelligence*, *Shenyang University of Technology*, *Shenyang* 110870, *China*;
2. *Guangzhou Institute of Industrial Intelligence*, *Guangzhou* 511458, *China*;
3. *Shenyang Institute of Automation*, *Chinese Academy of Sciences*, *Shenyang* 110016, *China*

Abstract

Human action recognition holds diverse research value across various scenarios and tasks, with promising applications in intelligent security, autonomous driving, and human-computer interaction. Although extensive research has been conducted on action recognition using human skeletal data, a systematic review of its development trajectory and underlying logic remain lacking. We review the major milestones in human skeletal action recognition, categorizing them into four key technological approaches: recurrent neural networks, convolutional neural networks, graph convolutional networks, and transformers. The developmental contexts of these methods are outlined, with an analysis of two key technological aspects: spatial modeling and temporal modeling. Strategies for constructing rich input representations are also highlighted. Additionally, the significance of skeletal modalities in multimodal integration for action recognition is discussed.

Keywords

action recognition;
human skeleton;
deep learning;
graph convolutional neural network;
Transformer

Finally, we discusse future directions for techniques and applications in human skeletal action recognition.

0 引言

人体动作识别(Human Action Recognition, HAR)旨在理解人类的行为,根据人体行为轨迹进行动作分类。其具体流程是输入一段或整个视频,将视频中人体的动作特征进行表示,进而实现动作识别分类,常被用来进行人体行为识别,如摔倒、翻越栅栏等。HAR 任务与传统的计算机视觉任务如图像分类[1-4]、目标检测[5-8]相比,面临着多维度信息表达、动作分类不统一、数据标注困难、干扰因素多等挑战。在人工智能技术广泛应用的背景下,动作识别研究者们也在智能交通[9]、智慧课堂[10-11]、智慧安防[12-15]和人机交互[16-18]等诸多领域进行了人体动作识别的多种尝试。还有一些研究者引入其他领域如自然语言处理(Natural Language Processing, NLP)[19-20]、边缘计算[21]等技术与人体动作识别进行交叉研究,进而推动计算机视觉的理论和技术发展,为图像内容的理解提供新的思路和方法,因此人体动作识别已成为重要和热门的研究课题之一。

人体骨架数据模态只使用包含人体关键节点的信息,而不需要处理复杂的背景信息,比常见的 RGB(Red-Green-Blue)数据模态需要处理的数据量更小,因此计算效率更高[22-23]。也与 RGB 模态易受环境变化影响不同,骨架模态对光照变化和背景噪声鲁棒性更好,同时在现实世界应用中,使用骨架模态也可以更好地保护隐私。因此,基于人体骨架模态的动作识别受到研究者们的广泛关注。

早期的骨架数据使用动作捕捉传感器[24-26]获取,精度较高,但是捕捉系统组成复杂,对场地大小及环境布置要求高,往往局限于实验场景,不适合在户外的日常场景使用。近年来通过姿态估计技术[27-28]可以直接在图像、视频等 RGB 数据集上提取人体 2 维甚至 3 维骨架信息,使得人体骨架动作识别可以在更真实、更复杂的场景下研究与应用。

当前一些研究者已经对动作识别领域进行了阶段性研究总结。张会珍等[29]从人体动作特征提取方法角度进行分析,将人体动作识别分为 3 个阶段:1)在视频帧中检测运动信息并提取底层特征;2)对行为模式进行建模研究;3)建立动作行为类别与底层视觉特征高层语义信息间的对应关系。SUN 等[30]总结了可以表达人类行为的多种数据模态,对各种输入数据模态类型的 HAR 深度学习方法进行了全面的研究调查,总结了多种数据模态在人体动作识别的使用方法,并分析了在各种应用场景不同数据模态的应用优势。毕春艳等[31]回顾了 2 维图像与视频数据在应用深度学习技术建模的不同之处,详细介绍了视频数据时序建模及参数优化的方法,分析了常用的动作识别数据集和度量参数,总结了如何在视频数据中更好地对时间信息进行建模。刘宝龙等[32]将基于骨架的人体动作识别划分为监督、半监督、无监督三大类并做了分析与比较。WANG 等[33]对比了 RGB 数据和骨架数据在人体动作识别的方法,并简要介绍了提取骨架的姿态估计算法。上述文献对人体骨架动作识别方法路线、相关数据集做了阶段性总结,但是没有总结骨架模态在动作识别中的重要作用,也没有剖析骨架动作识别技术发展的内在逻辑。

本文尝试从基于骨架模态出发串联起 HAR 任务的发展过程,对既有研究进行总结归类。首先简要概述了人体动作识别的概念以及常用的方法,然后聚焦于骨架动作识别,整理了常用数据集并分析了其发展过程。而后综合整理了应用在骨架动作识别的循环神经网络(Recurrent Neural Network, RNN)[34]、卷积神经网络(Convolutional Neural Network, CNN)[35]、图卷积网络(Graph Convolutional Network, GCN)[36]及基于 Transformer[37-38]的四大技术路线,从不同技术路线梳理了骨架动作识别的发展脉络,分析了基于骨架技术发展的内在逻辑,阐述了各种技术的发展史及现有变体,并做了综合比较,深入探讨骨架模态的优势与如何发挥其优势,包括骨架模态的来源、建模以及多模态融合的问题,整理了基于骨架模态的最新研究进展,同时介绍了一些基于骨架的人体动作识别应用,最后对骨架动作识别未来发展进行了展望。

1 人体动作识别

1.1 任务概述

早期的动作识别是在 RGB 图像数据上进行识别研究的,动作特征的注释与提取表示主要靠人工观察和设计,通常先利用手工提取特征(如轮廓剪

影、时空兴趣点、人体关节点、运动轨迹等)[39-40]，然后将这些特征转化为特征向量，之后送入到机器学习分类器(如支持向量机、隐马尔可夫模型、条件随机场等)中进行动作分类，进而实现行为识别等下游任务。这类方法的优点是模型结构简单、易实现，缺点是手工设计特征严重依赖先验知识，所提取的特征表达能力有限，很难设计更深层次的抽象特征，因此难以适应复杂多变的行为场景，在一些大型数据集上识别效果不佳。

近年来，随着深度学习技术的发展，神经网络模型能够在各种复杂的场景下提取到更深层次的特征信息[41]，并进行高效的特征融合。基于深度学习的动作识别技术不仅提高了识别速度，还能够取得较好的识别准确率。根据输入数据模态的不同，HAR 最受关注的方法有基于 RGB 图像和基于骨架数据两大类。RGB 图像通常就是普通相机采集到的视频数据，范围来源广泛，包含日常生活、工作、体育运动等多种动作场景，涵盖居家、体育场、野外等多种背景。骨架数据可以通过动作捕捉传感器直接对人体动作进行捕获得到人体骨架信息，也可以使用姿态估计技术在其他数据模态(通常是 RGB 模态)提取出人体骨架信息，后者更具有实用性。

基于 RGB 模态的方法通常使用端到端的方案直接处理 RGB 数据集。首先获取视频数据中的动作特征，然后进行特征融合，最后进行分类。因为基于 RGB 模态的方法是在全 RGB 数据中提取特征，通常参数量比较大，计算成本比较高，且容易受到 RGB 数据中环境变化的影响。动作识别关注的核心在于人体的姿态变化，那么从人体姿态变化的角度出发，只需要对人体关节点的位置变化进行判断，自然就可以断定人体动作。骨架动作识别就是一种基于人体关节变化思路解决动作识别的方案，与 RGB 的方法不同，骨架动作识别不需要在空间维度上处理大量不相关的背景信息，只专注于人体动作，不易受到如光照、拍摄角度、服装背景改变等因素的影响，通常具有更好的鲁棒性[33]。在时间维度上，对各关节点的变化进行建模，可以更加精准确定肢体动作变化的幅度，对于人体姿态变化能够更好的把控建模，并且骨架数据结构紧凑、运算效率高，有着更快的识别速度。对比其他模态，骨架动作识别有其显著特点：1）直接关注于人体动作，不需要无关背景的冗余计算；2）关节点之间联系性强，有明确的几何关系与物理意义，能够更好地表征人体运动特征；3）骨架数据结构紧凑、信息密集、抽象程度高、不易受外界环境变化影响。4）可从其他多种模态中提取出来，泛用性更广。

骨架动作识别处理范式如图 1 所示，通过动捕设备或姿态估计技术在实际场景下或其他数据上获取骨架数据，然后送入到识别算法中进行空间与时间建模，然后对建模后或获取的特征进行融合，再送入到分类器计算得到分类概率，最终给出预测分类类别。

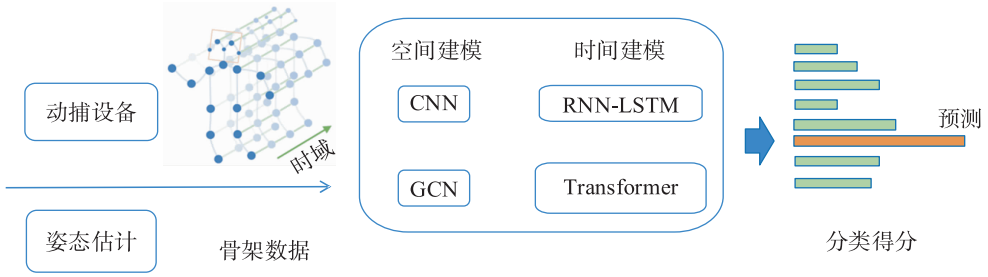

图 1　骨架动作识别

Fig.1　Skeleton-based action recognition

1.2　骨架数据集

数据集的质量与规模深刻影响着深度学习技术的发展，动作识别算法的不断更新与应用领域泛化也对数据集提出了更高要求，推出新的数据集以应对新的挑战。表 1 整理了经典数据集和一些新发布的多视角骨架数据集，对数据集的发展做一个简要回顾。动作识别技术在不断的发展，同时也推动着动作识别数据集的不断演变。2016 年之前受硬件设备与技术限制，骨架数据来源于提取的点云视频且数据集样本数量很少[42-43]，后面依赖于动作捕捉设备的发展出现了比较完善的 NTU RGB + D 数据集，再到高效的姿态估计算法提出，如 YAN 等使用 OpenPose 算法在普通视频上进行骨架数据提取。在样本场景方面，逐渐从日常动作、实验场景向着

其他场景、其他视角拓展，如文[44－45]分别构建了老年人动作数据与无人机视角下的动作数据，拓展了动作识别的应用场景。同时，在一些非视觉模态上，如雷达[46－47]、音频[48－49]等模态上进行动作识别，也取得了一些进展。虽然还未有像 Kinetics-400 同等规模的数据集提出，场景也往往局限于实验阶段，但是都可以提取出人体骨架信息使用骨架动作识别的方法进行初步分析。因而研究多种数据集在同一动作类别在不同数据模态上的对齐也是值得关注的方向。

骨架动作数据集通常使用动作捕捉设备在实验环境下获取，或是使用姿态估计技术在其他模态数据集上获取。这两类数据集的典型代表分别有 NTU RGB + D[22]与 Kinetics-skeleton[50]数据集。

NTU RGB + D 系列由南洋理工大学的 Rose Lab 实验室在 2016 年发布，命名为 NTU RGB + D，有 25 个身体关键点，包含 56 880 个样本数据、60 种动作类别，包含日常动作与交互动作，且使用交叉受试者评估(Cross-Subject，CS)和交叉视角评估(Cross-View，CV)两种评估方案对算法进行评价。而后在 2019 年原团队提出了上一版本的增强版 NTU RGB + D 120[51]，涵盖之前的所有数据，并增加了额外的 60 个类别，总的样本数为 114 480。这两个数据集都包含每个样本的 RGB 视频、深度序列、3 维骨骼数据和红外(IR)视频。每个数据集由 3 个 Kinect V2 摄像头同时捕获。

表 1 常用骨架数据集

Tab.1 Commonly used skeleton datasets

名称	发表年份	来源	样本数	类别	场景
UWA3D Multiview[42]	2014	3 维点云视频	900	30	日常动作
UWA3D Multiview II[43]	2015	3 维点云视频	1 075	30	日常动作
NTU RGB + D[22]	2016	Kinect v2	56 880	60	实验场景、日常动作
Kinetics-Skeleton[50]	2018	YouTube + OpenPose	306 245	400	实际场景、日常动作
NTU RGB + D 120[51]	2019	Kinect v2	114 480	120	实验场景、日常动作
ETRI-Activity3D[44]	2020	Kinect v2	112 620	55	中老年人的日常活动
UAV-Human[45]	2021	无人机拍摄 + 姿态估计	67 428	155	无人机角度日常生活

Kinetics-skeleton 是 2018 年 YAN 等在视频动作数据集 Kinetics 系列的 Kinetics-400 数据集[52]上，使用 OpenPose 技术提取骨架数据得到的。Kinetics-400 是在 YouTube 网站上收集切片后得到的，包含 400 种动作类别，涵盖了大部分日常生活的动作，不仅包括人与人的日常交互动作(拥抱、握手)，还有一系列人与物的交互。同时，由于拍摄方法差异，动作视角差异大，存在更多遮挡截断的困难，类内差异明显，识别更加具有挑战性。

1.3 评价指标

动作识别与图像分类任务输出相同，都依靠分类的准确率来评价，其定义可以表示为

$$A_{\mathrm{P}}=\frac{N_{\mathrm{True}}}{N_{\mathrm{All}}} \tag{1}$$

其中，A_{P} 称之为预测准确率，N_{True}代表模型预测正确的个数，N_{All}代表预测的所有样本的个数。需要注意的是 NTU RGB + D 与 NTU RGB + D 120 使用多视角、多机位采集数据，构建了交叉验证评估方案。在 NTU RGB + D 数据集中分为交叉受试者(CS)评价和交叉视角(CV)两种方法进行综合评价。交叉受试者将类别相同但是不同实验者来划分训练集与测试集，交叉视角将不同机位调整为不同角度来划分训练集与测试集。NTU RGB + D 120 也划分为交叉受试者(C-sub)评价和交叉设置(Cross-setup，C-set)两种评价方案，交叉设置与交叉视角有所不同，在高度和距离都做了调整。Kinetics-400 采用 Top-1(模型预测分数最高的类别是真实类别的概率)和 Top-5(模型预测分数前 5 最高的类别包含真实类别的概率)进行评价，为了统一指标比较，后文皆使用 Top-1 进行评价。

2 基于深度学习的骨架动作识别

随着人工智能技术的不断发展，深度神经网络成为当前的研究热点。在骨架动作识别方面，深度学习通过使用各种神经网络自动学习如何在数据中提取特征，能够将人体骨架数据模态提取出抽象程度更高的特征信息，进而表征丰富的人体运动。应用在动作识别领域的深度学习方法根据技术路线的不同可以分为 4 类：RNN、CNN、GCN 及基于

Transformer 的方法。本节梳理了每种方法的发展历史并总结各自特点，将在第 3 节对多种方法进行比较，深入讨论各种技术发展的内在逻辑。

2.1　RNN 及扩展方法

最初因为 RNN 在处理序列数据方面具有巨大的优势，RNN 被广泛应用于自然语言处理(NLP)领域。表征人体动作的骨架序列也是在时间维度上序列化的人体关键点信息，因此，早期动作识别研究者们采用各种方法在识别任务中调整 RNN 和长短时记忆(Long Short-term Memory Networks，LSTM)网络[53-66]，期望有效模拟 HAR 的骨架序列内的时间上下文信息，进而实现骨架动作识别和建模。本节介绍了 RNN 方法在骨架动作识别上的应用及其改进工作，并且整理了各种算法在 NTU RGB + D 与 NTU RGB + D 120 数据集上的性能比较，见表 2。

通过将人体动作视为骨架关节轨迹，RNN 可以很好地模拟轨迹时间序列的上下文关系。DU 等[57]使用 RNN 网络构建了基于骨架的动作识别的端到端解决方案，提出了一种名为 HBRNN-L 的层级双向循环神经网络模型，用于骨架动作识别。如图 2 所示，HBRNN-L 将人体骨架分成 5 个部件(1 个主躯干 +4 个肢体)分别进行特征提取，然后通过层级双向循环神经网络结构将各部件的特征信息进行融合，接着在 BRNN 中的最后一层添加 LSTM 神经元对输入数据进行建模，以便更好地捕捉时间序列数据中的长期依赖关系，达到提高模型对时间序列信息的建模能力，使模型能够更好地理解和预测动作序列中的相关模式和特征。

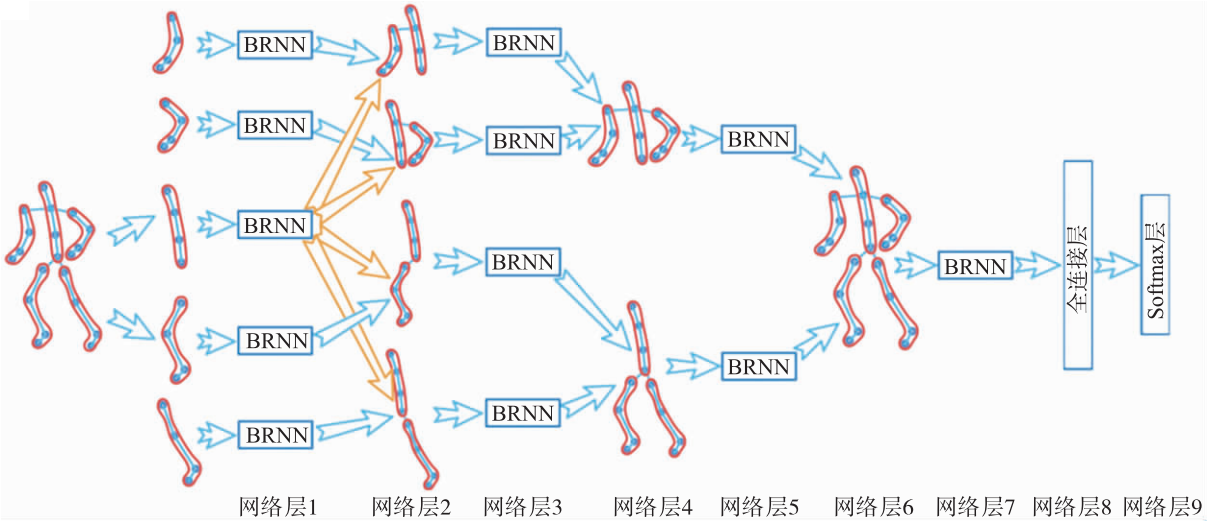

图 2　HBRNN-L 网络模型[57]

Fig.2　HBRNN-L network model[57]

RNN 方法往往只能对时间较短的时序特征信息进行有效建模，且因其简单结构会导致梯度消失或爆炸的问题。相比之下 RNN 网络的变体 LSTM 因其输入门、遗忘门和输出门等的结构，能够更好地控制信息的流动和存储，在动作识别任务中的时序建模方面比普通 RNN 更有优势。因此，LSTM 网络的相关方法在动作识别中得到了广泛的研究和关注。

为了更好地提取不同阶段的信息，许多研究者将时空特征分为短期、中期和长期特征。LEE 等[58]考虑到具有不同时间步长的 LSTM 网络可以很好地模拟不同的运动属性，提出了基于骨架动作识别的集成时间滑动 LSTM(TS-LSTM)网络。该网络通过对比实验验证了使用 LSTM 来融合高级空间和时间特征，可以学习到跨时间的隐藏特征，相比其他网络获得了更好的动作识别效果。为了充分整合时空领域信息，CUI 等[59]对于短期、中期和长期特征提出了一种基于时空模型的多源模型，将时间模型分为 3 个分支，并设计了一个空间序列分支，如图 3 所示。3 个时间分支各自处理全部的关节点、各大小不同的躯体分级，最后将分别用于感知全局信息、局部信息和细节信息的 3 个分支分数融合，得到最终结果，3 个时间模型分支处理不同层次的时间信息有助于动作识别的准确性。实验结果表明，各分支模型的准确度低于 3 分支的融合模型，证明了更丰富的时序信息可以为动作分类提供更强大的支撑。

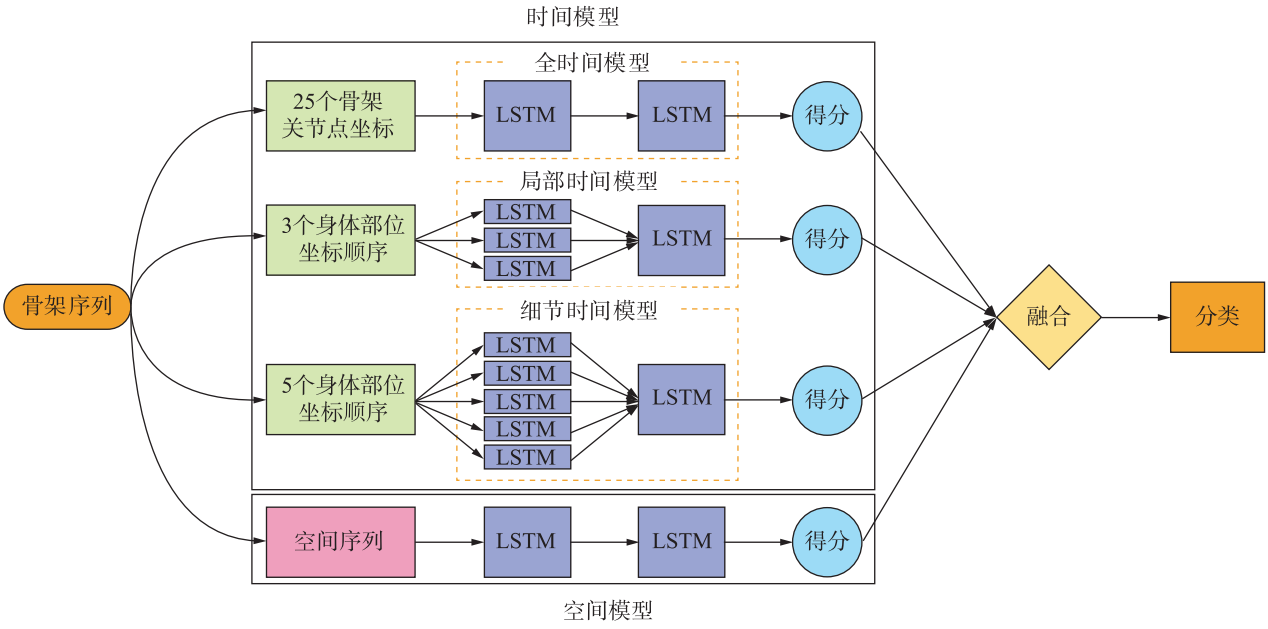

图 3 多源 LSTM 网络模型[59]

Fig.3 Multi-source LSTM networks model[59]

虽然 LSTM 网络具有良好的时间建模能力，但是对于骨架模态空间信息的利用有所欠缺，无法充分发挥骨架本身结构的优势。将所有的关节点都作为输入时，一些不相干的关节点会影响模型的性能，需要使模型更关注于包含重要信息的关节点变化，但是 LSTM 本身对于空间信息利用就比较差，也没有很强的“注意力”关注到重要关节点。针对 RNN/LSTM 方法这一弱点，通常会将其与其他方法（CNN、GCN、注意力）结合使用，来增强 RNN/LSTM 模型对空间信息的利用，同时发挥其帧间信息、长时间信息的时间建模能力。LIU 等[60]引入全局情景感知注意力机制，与 LSTM 循环的思想结合，提出了名为 GCA-LSTM（Global Context-Aware Attention LSTM），用来选择性的关注不同重要性的关节，如图 4 所示。该网络由两层 LSTM 网络组成：第 1 层对骨架序列进行编码并输入到全局上下文记忆器中，第 2 层输出注意力关系对全局上下文关系进行优化。

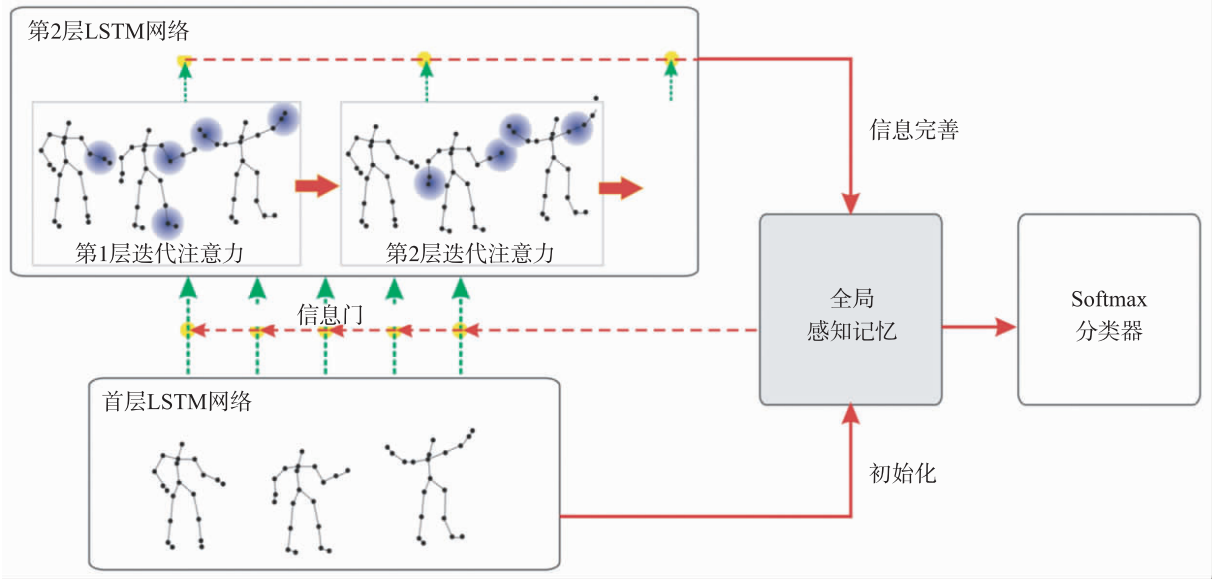

图 4 GCA-LSTM 网络结构[60]

Fig.4 Global context-aware attention LSTM network structure[60]

CNN 网络在提取空间特征方面是高效的，但是无法很好地融合时间信息。LI 等[64]有效地结合了 CNN 高效空间信息建模和 LSTM 高效时间信息建模的优点，由 CNN 和 LSTM 模型分别识别，最后融合分数，获得更准确动作分类。ZHENG 等[65]提出了一种注意力递归关系网络 LSTM（ARRN-LSTM），可以同时模块化建模空间布局和时间运动特征，递归关系网络学习单个网络中的空间特征，LSTM 用于提取视频序列中的时间特征。如图 5 所示，该网络设计了关节点与连线两种输入，首先进行统一编码送入到递归关系网络中提取特征信息；结合注意力选取更重要的特征；然后使用 LSTM 网络建模时间特征；最后将两种分数融合，实现互补位置信息与结构信息，帮助确定具体的姿势和动作。

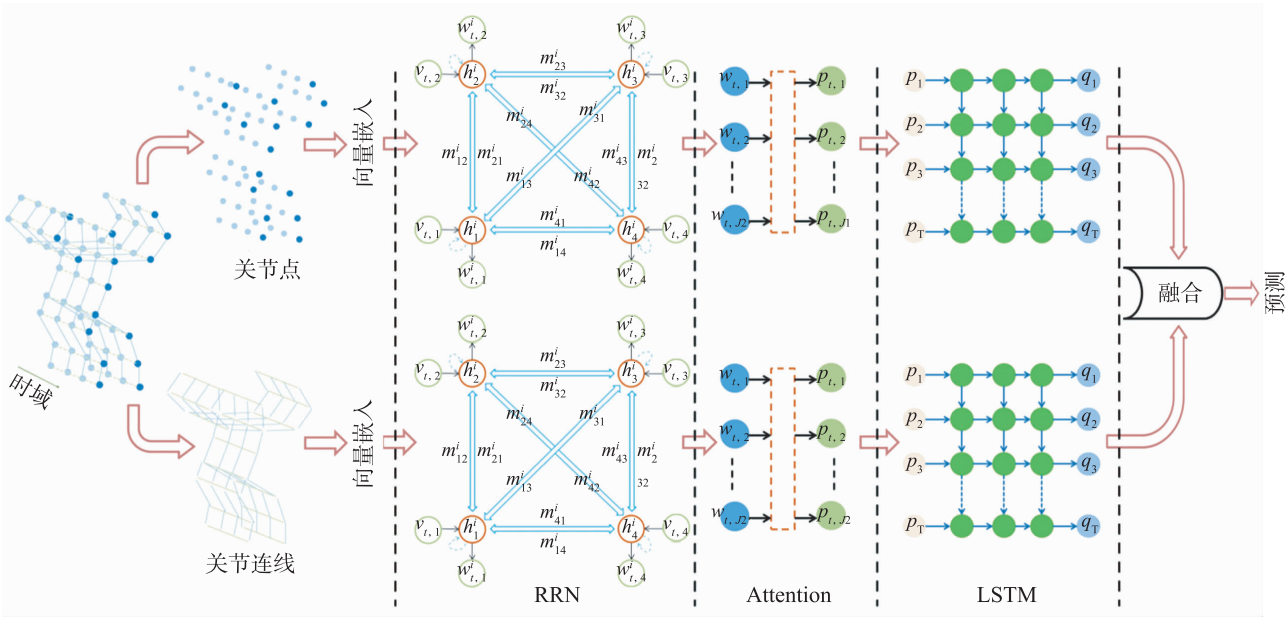

图 5　ARRN-LSTM 网络结构[62]

Fig.5　ARRN-LSTM network structure[62]

动作识别在时间上和空间上虽然有非常不同的特征表达，但终究是一个统一整体，在不同维度上存在着共现特征，运用该共现特征可以更好地表现时空特征，强化网络模型的表达能力。SI 等[66]结合图卷积方法提出了一种注意力增强图卷积 LSTM 网络（Attention enhanced Graph Convolutional LSTM，AGC-LSTM），其单元结构见图 6。该单元结合图卷积对人体骨架进行特征提取将其融入到 LSTM 中，同时捕获空间与时间信息，挖掘时空之间的共现特征，并在捕获时空信息中的共性特征后使用注意力机制改善关键节点的特征。

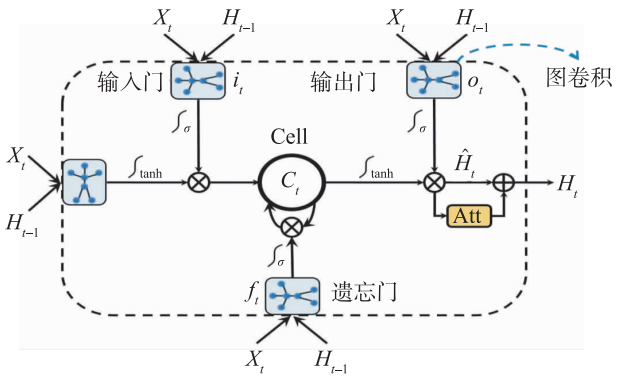

图 6　AGC-LSTM 单元[63]

Fig.6　AGC-LSTM block[63]

骨架动作识别依赖于人体骨架信息进行分类，而人体骨架的运动在图空间表达上是非线性的。对于同一种动作，不同的观察视角会得到不同的姿态变化数据，将其变换为合适的视角将有利于运动识别。ZHANG 等[64]分析了动作识别中不同的视角对于识别准确率的影响，设计了自适应视点的方案，将各种视图的骨架转换为更加一致的视图，同时保持动作的连续性，实现识别过程中自动确定“最佳”观察视角。同时，使用融合递归神经网络和卷积神经网络的双流方案，减少视角变化对模型性能的影响，并在训练过程中引入视角丰富化的方法，来增强视角变化的鲁棒性。FRIJI 等[65]认为以前主要的深度学习方法都是在特征空间上进行设计的，而忽略了形状空间的影响。因此他们引入 Kendall 形状空间[66]，构建了 KshapeNet 模型，如图 7 所示，将

原本的骨架序列重新构造了 Kendall 形状空间上的轨迹，映射到线性切线空间，并将生成的结构化数据输入到深度学习架构中，通过空间变化的方法找到合适的视角输入。

综合来看，RNN 及其变体方法 LSTM 在骨架动作识别中的发展过程，经历了方法引入、长短期时间关系引入、空间信息增强融合等多个时期。受限于自身结构限制，对时间处理效果好，而对空间利用效果差，在最近的应用中通常结合其他模型使用。表 2 列出了几个经典模型及其比较，可以看出随着对时空特征建模能力的提升，模型性能也越来越好。

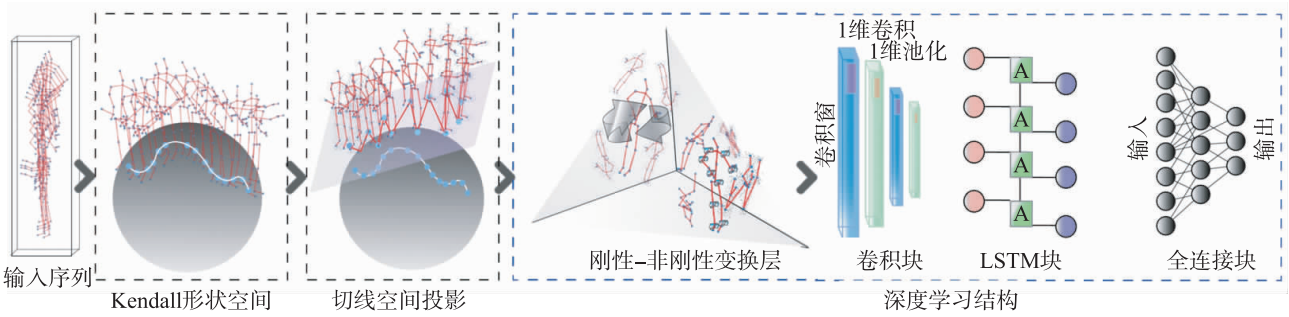

图 7 KShapeNet[68]

Fig.7 KShapeNet[68]

表 2 RNN/LSTM 指标比较

Tab.2 RNN/LSTM index comparison

模型	发表年份	NTU RGB + D		NTU RGB + D 120		创新工作
		CS	CV	C-sub	C-set	
HBRNN-L[57]	2015	59.1	64.0			端到端分层 RNN
Trust Gate ST-LSTM[67]	2016	69.2	77.7	58.2	60.9	2 维 LSTM
GCA-LSTM[60]	2017	74.3	82.8	58.3	59.2	多层 LSTM + 全局感知记忆单元
LI[61]	2017	85.0	92.3			平移尺度不变图像映射和多尺度深度卷积神经网络
CUI[59]	2018	66.6	78.1			多分支时间融合
ARRN-LSTM[62]	2018	80.7	88.8			双流 RNN + LSTM
dense-IndRNN-aug[68]	2018	86.7	93.9			独立递归神经网络
AGC-LSTM[63]	2019	89.2	95.0			GCN + LSTM
KShapeNet[65]	2021	97.0	98.5	90.6	86.7	几何感知网络

2.2 CNN 方法

卷积神经网络(CNN)因其强大的特征提取能力，广泛应用于计算机视觉领域。在基于骨架的动作识别中，CNN 也是一种广泛使用的方法。一些基于 CNN 的骨架动作识别方法首先将骨架数据编码成伪图像，然后提取这些图像的特征，这样可以充分借鉴 CNN 在图像识别领域的应用经验。本小节梳理了基于 CNN 方法及其改进的骨架动作识别算法与模型，并在 NTU RGB + D 与 NTU RGB + D 120 数据集上进行性能比较，最后简要概述了模型的创新之处，如表 3 所示。

由于 CNN 在图像数据上提取特征能力强，而骨架序列是关节点及其连线组成的关节构成的人体骨架信息，难以直接使用卷积进行特征提取。因此，DING 等[69]将骨骼数据转换为纹理颜色图像，并使用卷积提取高级特征，该方法主要包括从输入的骨架序列中提取空间特征、选择关键特征，然后从关键特征生成纹理彩色图像，最后基于图像进行 CNN 模型训练。该方法提取了 5 种关节组合特征：关节间的距离(Joint Distance)、关节的方向(Joint Orientation)、关节的速度(Joint Velocity)、关节和躯干之间的距离(Joint to Limb Distance)、四肢之间

的角度(Limb Angle)，然后选择关节组合的关键特征进行颜色编码，生成 13 种类型的图像。CNN 在每种图像上进行训练，并将 CNN 的输出得分融合后进行最终识别，其构建的模型如图 8 所示。WANG 等[70]通过颜色编码将关节轨迹的空间配置和动态表示为 3 个纹理图像，称为关节轨迹图(JTMs)，骨架序列编码为包含每帧空间结构信息和帧间时间动态信息的伪图像，然后馈送到微调在ImageNet 上训练过的 CNN 模型中进行动作识别。

表 3　CNN 方法性能比较

Tab.3　CNN method performance comparison

模型	发表年份	NTU RGB + D		NTU RGB + D 120		Kinetics-skeleton	创新工作
		CS	CV	C-sub	C-set		
SkeleMotion[71]	2019	76.5	84.7	67.7	66.9		不同时间尺度显示编码运动信息
VA-CNN[72]	2019	88.7	94.3				多视角 + 数据扩充
VA-fusion[72]	2019	89.4	95				RNN + CNN
Avinandan Banerjee[73]	2021	84.2	89.7	74.8	76.9		4 种互补角度的特征
POSE-C3D (3D Heatmap)[74]	2022	94.1	97.1	86.9	90.3	47.7	3 维热图体积
POSE-C3D (RGB + pose)[74]	2022	97	99.6	95.3	96.4	85.5	RGB + skeleton

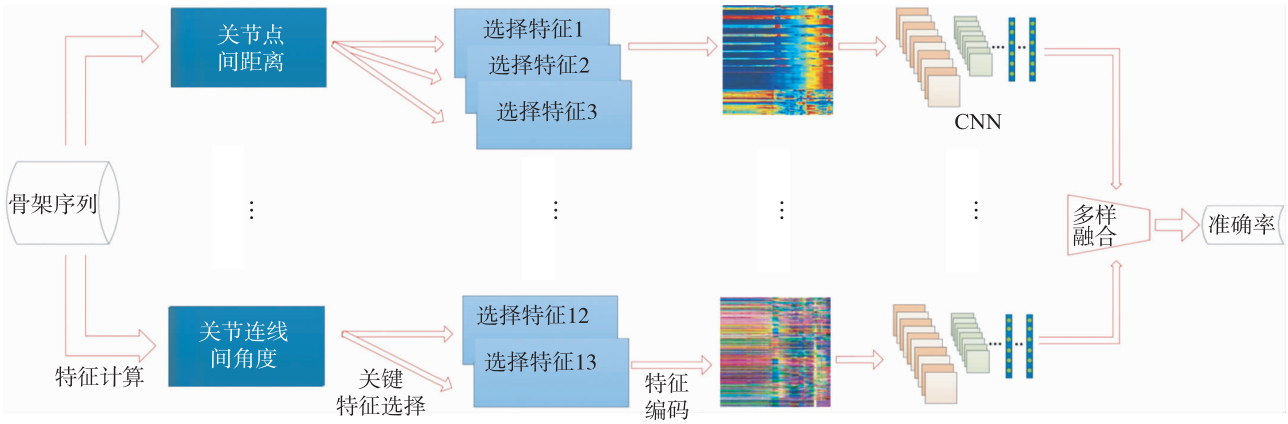

图 8　编码纹理图像[69]

Fig.8　Encoded texture image[69]

CNN 在空间特征建模方面拥有卓越能力，但在时间建模方面并不擅长，尤其是在远距离时间运动信息的提取方面，于是许多学者尝试解决 CNN 方法中长期运动建模问题。由于 3 维卷积神经网络(3D CNN)能够同时学习空间和时间维度特征，LIU 等[75]将其应用于骨架动作识别，提出使用 3 维 CNN 构建时间流与空间流的双流网络。通过将骨架关节映射到 3 维坐标空间，分别编码空间和时间信息，然后采用两路 3 维 CNN 模型分别提取深度特征，最后将每个流扩展成多时间版本，以增强深度特征捕获全局关系的能力，通过时空流增强学习全局运动信息。为了使骨架信息更准确、更全面地表示人体动作，CARLOS 等[71]改进了以往骨架数据编码为骨架图像的表示方法，提出了通过显式计算骨架关节大小和方向值来编码时间动力学的骨架图像表示法——SkeleMotion。如图 9 所示，SkeleM-

otion 将骨架关节的运动信息(速度、方向和运动范围等)进行编码和分析，采用不同的时间尺度来计算运动值，将更多的时间动态聚集到表示中，进而捕捉动作中涉及的长范围关节交互。同时，使用过滤方法去掉一些运动噪声，以更有效地捕捉和表达动作的时间动态特征。BANERJEE 等[73]旨在提高基于骨架动作识别中 CNN 模型的分类能力，提出了从骨骼标记中提取信息特征的方法。首先提取距离和角度及其向量变化的时空动态特征，然后将这些特征编码为单通道灰度图像传入 CNN 分类器，最后将 4 种特征 CNN 分类结果进行模糊积分融合得到最终结果。

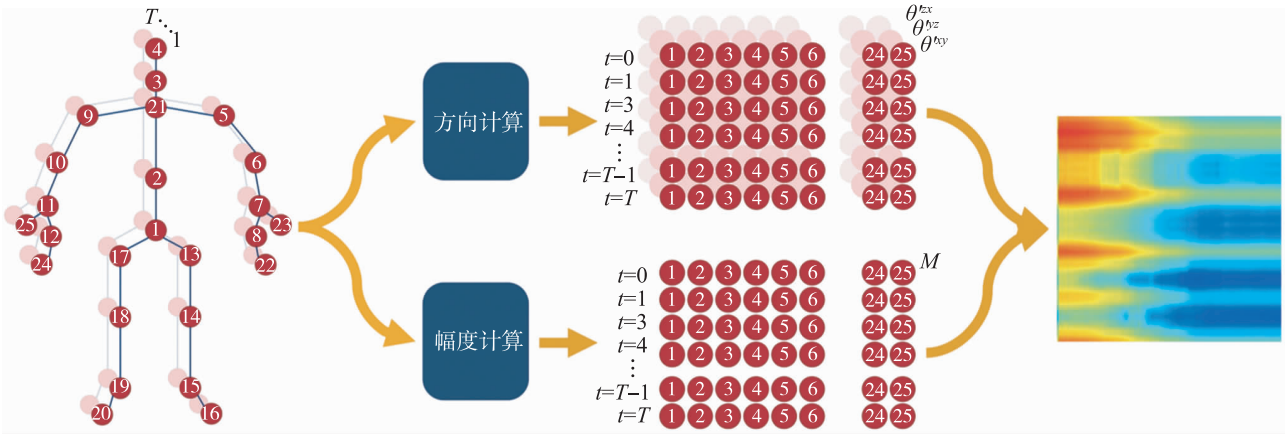

图 9 SkeleMotion 骨架编码方式[71]

Fig.9 SkeleMotion skeleton encoding method[71]

上述工作中，都是将骨架编码转化为颜色空间，这种转变方法在一定程度上失去了关节间的结构信息，并使得数据可视化效果变差。DUAN 等[74]重新思考了骨架模态的特点(只包括姿态信息，骨骼序列只捕获动作信息，同时不受上下文干扰)，采用了将骨骼数据膨胀为热图，然后将不同时间步长的热图沿着时间维度堆叠，最后形成热图体积的方法来表示骨架模态，而不是直接对骨架坐标进行处理，并提出了一个新的框架 PoseConv3D，使用了包括以动作主体为中心进行裁剪以保留尽可能多的信息，采用均匀采样来捕获视频的整个动态，使用伪热图来提升 3 维 CNN 的效果等一系列方法来提高基于骨架的动作识别性能，其具体的模型结构如图 10 所示。PoseConv3D 对于视频中的每一帧，首

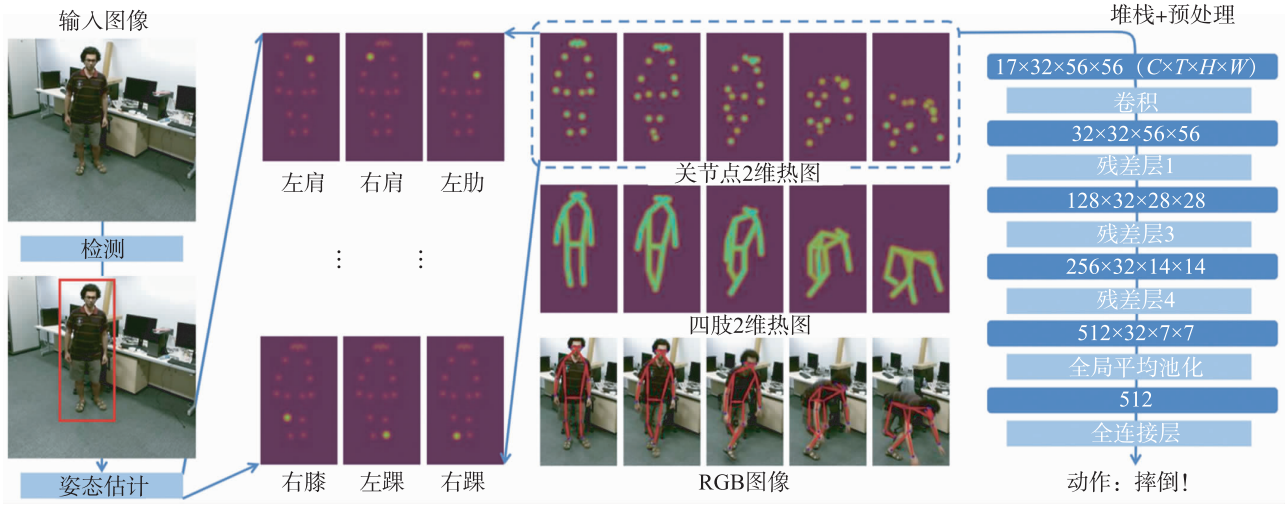

图 10 PoseConv3D 网络[74]

Fig.10 PoseConv3D network[74]

先使用两阶段姿态估计器(检测 + 姿态估计)进行 2 维人体姿态提取；然后沿着时间维堆叠关节或肢体的热图，并对生成的 3 维热图体进行预处理；最后使用 3 维 CNN 对 3 维热图体积进行分类，同时这种方法可以方便地结合 RGB 数据取得更好的识别结果。

基于 CNN 的方法背后的逻辑是利用 CNN 强大的图像特征提取能力，通过将骨架数据转换为图像或热图形式，从而使 CNN 能够高效地提取时空特征。同时，通过引入 3D CNN、多时间尺度编码、热图表示法等技术进行多动态特征编码，增强对时间动态特征和全局关系的捕捉能力，提高基于骨架的动作识别性能。表 3 列出了基于 CNN 的方法中典型算法在 NTU RGB + D 与 NTU RGB + D 120 数据集上的性能及其核心创新点。

2.3　GCN 方法

基于 GCN 的骨架动作识别方法源于对人体骨架信息的重新思考。GCN 是一种针对图数据进行学习和推理的神经网络模型。传统的神经网络主要针对规则结构的数据，如图像和文本，而 GCN 则专门设计用于处理图形数据，这些数据的结构可以是任意的图结构，例如社交网络、分子结构、交通网络等。对于骨架模态的数据来说，骨架自然就是图结构的形式(由关节点与其接线构成)。之前的研究工作简单地将骨骼数据表示为由 RNN 处理的向量序列，或由 CNN 处理的 2 维/3 维特征图，不能完全模拟身体关节的复杂时空配置和相关性，而使用拓扑图可能更适合于表示骨架数据。不同于传统的 CNN 和 RNN 网络模型，GCN 拥有处理具有广义拓扑结构数据[83]的能力，并深入挖掘其特征。

YAN 等[50]指出基于骨骼的动作识别方法通常将骨骼看作整体进行建模，引入了 GCN 的思想来处理人体骨骼序列的空间结构和时间动态。通过引入空间-时间 GCN(ST-GCN)从骨架数据自动学习空间和时间模式，开发了用于基于骨架 HAR 的 GCN。如图 11 所示，该模型将关节点作为图节点、骨架作为图的边，进行图卷积空间建模，在时间维度上连接不同帧中的同一属性关节点进行时间建模，克服了骨骼序列在 2 维或 3 维网格数据形式上表示的困难。ST-GCN 模型能够自动地处理骨骼序列的空间配置和动态变化，而不再需要手动指定对象部分，提高了模型的灵活性和泛化能力。

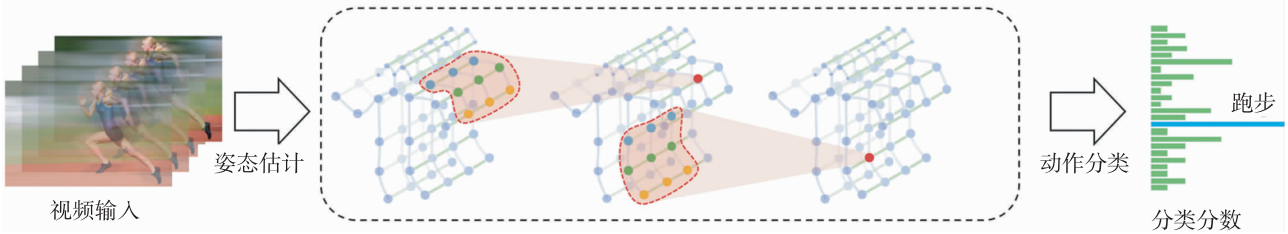

图 11　ST-GCN 流程结构图[50]

Fig.11　ST-GCN process structure diagram[50]

ST-GCN 仅包含直接连接关节的结构信息，因而只能捕捉关节之间的局部物理依赖性。LI 等[77]针对这一缺点，提出一种动作-结构 GCN(AS-GCN)以提高识别效果。在 AS-GCN 中使用了编码器-解码器网络，建立非直接连接关节之间的关联，称之为动作链接；对相邻关节点构建高级多项式建立更好的依赖关系，并在整个骨架图上进行拓展，称之为结构链接；最后将动作链接和结构链接结合成一个广义骨架图。动作链接用于捕获特定于动作的潜在依赖关系，结构链接用于表示高阶依赖关系。SHI 等[78]认为之前的图拓扑是手动设置的，并且在所有层和输入样本上是固定的，这种固定的拓扑结构对于动作识别任务中的分层 GCN 和不同样本来说可能不是最佳的。同时提出骨骼数据的二阶信息(骨骼的长度和方向)对于动作识别来说更具信息量和判别性，设计了具有残差结构的自适应图卷积层，组合非直接连接的关节之间的关联信息，使得图的拓扑结构更加灵活。总体结构上建立关节和骨架的信息识别流，通过融合给出最终识别结果。

为了从大量信息中快速筛选出有价值信息，一部分 GCN 使用了注意机制[79-80]。为了更好地探索骨架隐含的联合相关，CHENG 等[81]开发了移位图卷积网络(shift-GCN)在空间图与时间图上灵活地进行移位图运算获取感受野(图 12)，突破了传统 GCN 时空域缺少灵活性的局限性，并使用轻量级逐点卷积大大减少了计算量。

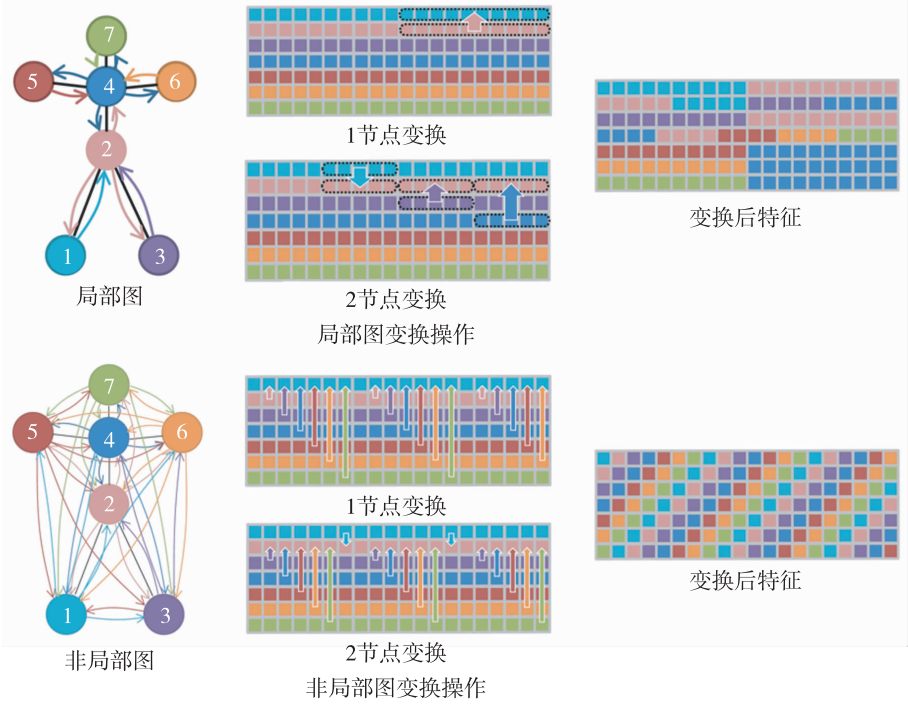

图 12 空间移位图操作[81]

Fig.12 Space shift map operation[81]

LIU 等[82]认为之前空间信息与时间信息提取只能先空间提取再时间提取(或反之)的方式不够灵活，一个关节点应当能够直接聚合其不同相邻帧节点的信息，提出了一种新的 G3D 图卷积方式，如图 13。该卷积方法不仅能在关节点周围距离为 1 的邻接点聚合特征信息，还能在距离为 $K(K>1)$ 的邻接中聚合特征，同时结合时间维度以增强时空信息流动的效率。一些学者关注在完整骨架结构中捕捉信息，然后将完整骨架结构分级为部件信息的方法[83]，将部件的特征分别进行学习，达到人体动作的有效识别。由于 GCN 在处理非欧几里得空间信息方面的优势，成为基于骨架动作识别技术中一个非常热门的研究方向。

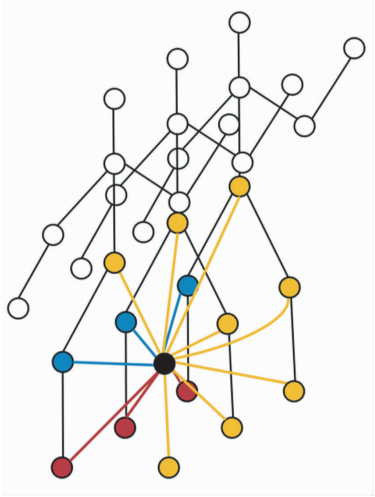

图 13 G3D 图卷积[81]

Fig.13 G3D graph convolution[81]

图拓扑的结构虽然能够有效的处理骨架信息，但是也受限于图结构有限的点与点之间的连接。另一些研究旨在解决 GCN 模型的局限性，尝试使用超图来捕获人体骨架的动作信息，构造局部超边和全局超边提取高阶特征信息，并使用超图注意力机制获得相邻节点的不同权值，进而捕获骨架数据中的非连接区域更丰富信息，而不受典型人体骨架图结构的限制。费树岷等[84]重新审视了身体躯干配合对于不同动作的影响，认为应该重点关注同一动作各关节骨架之间的配合，提出了时序拓扑非共享图卷积(temporal topology unshared graph convolution network，TTU-GC)和多尺度时间卷积相结合的动作识别方法。该方法采用时序拓扑非共享图卷积为每帧骨架建模独立的样本动态变化空间拓扑，采用多尺度时间卷积提取不同尺度空间上的动作特征，然后将这种方法作用于关节、骨骼、关节运动和骨骼运动多流数据，最终融合得分做出预测。该方法在略微提升模型复杂度的情况下，实现了图卷积在骨架动作识别上更灵活的使用。

综上所述，GCN 在基于骨架的动作识别中，通

过利用其对图数据的强大处理能力，能够更自然地表示和分析骨架的时空结构和动态变化。通过不断改进和引入新的图卷积方法，如自适应图卷积、移位图卷积和超图模型等，研究者们提升了模型的灵活性、泛化能力和识别效果，使得 GCN 在基于骨架的动作识别领域成为一个非常热门的研究方向。表 4 整理了基于 GCN 模型的骨架动作识别算法性能比较和创新性工作。

表 4　GCN 模型识别性能比较

Tab.4　Comparison of GCN model recognition performance

模型	发表年份	NTU RGB + D		NTU RGB + D 120		Kinetics-skeleton	创新工作
		CS	CV	C-sub	C-set		
ST-GCN[50]	2018	81.5	88.3	70.7	73.2	30.7	骨架时空图卷积开山之作
AS-GCN[80]	2019	86.8	94.2	78.3	79.8	34.8	强化全局关节连接 与未来动作帧预测
2s-AGCN[78]	2019	88.5	95.1			36.1	自适应图卷积
Shift-GCN[81]	2020	90.7	96.5	85.9	87.6		移位图卷积
MS-G3D[82]	2020	91.5	96.2	86.9	88.4	38	多尺度统一时空图卷积
CTR-GCN[85]	2021	92.4	96.8	88.9	90.6		通道拓扑细化网络
ST-GCN + +[86]	2022	92.6	97.4	88.6	90.8		更好的数据预处理 与代码实践
InfoGCN[87]	2022	93.0	97.1	89.8	91.2		注意力机制学习潜在表征
STEP CAT-Former[88]	2023	93.2	97.3	90	91.2		时间注意力变换器
LA-GCN[89]	2023	93.5	97.2	90.7	91.8		大语言模型监督骨架
TTU-GC[84]	2024	92.3	96.6	88.8	90.2		时序拓扑非共享图卷积

2.4　Transformer 方法

Transformer[90]结构完全基于注意力机制，拥有层叠的编码器解码器，可以并行处理而不是依赖于序列数据的逐步处理，因此在长期依赖建模、多模态融合和多任务处理方面表现良好[91]。Google 团队将 Transformer 应用于图像分类模型 VIT（VIsion Transformer）[92]之后，引爆了 Transformer 在计算机视觉领域的应用研究。由于 Transformer 可以并行使用自注意力机制来捕捉各分割图片块之间的关系，有效理解并聚合图像的局部及全局特征，同时具有良好的时空建模能力，许多研究者也将 Transformer 应用于骨架动作识别。

对于骨架建模，Transformer 可以使用自注意力机制实现任意关节点的关系建模，进而改善图卷积网络全局建模的短板。因此专注于骨架模态序列建模的编码器研究成为了 Transformer 在骨架动作识别领域的主要任务。为了模拟每个骨架帧的姿势及其在整个时间跨度内的运动，ZHANG 等[93]提出了一种空间-时间专用编码器，名为 STST-Encoder。该编码器由 8 个空间-时间专门化 Transformer 块堆叠而成，每个块由一个空间 Transformer 块（STB）和一个方向时间 Transformer 块（DTTB）组成。STB 块用于在帧级处理姿态信息，建模在同一帧中各链接关节点的运动关系；DTTB 模块用于捕捉时间维度中的长动态，建模关节点在时间维度上的变化。同时为了保持模型在学习不完全骨架时的性能，设计了一个多任务自我监控模块，自监督学习姿势一致性（PC）、骨架部位拓扑关系（PS）和骨架姿势变换（PT）等任务，改善时间维度中建模姿势变化对于动作识别任务的影响，增强识别模型对运动噪声的鲁棒性。

为了更好地建模局部关节的依赖关系与不同身体部位之间的帧内相互作用，PLIZZARI 等[94]引入了时空 Transformer 网络（ST-TR），提出了空间自注意力模块（Spatial Self-Attention module，SSA）和时间自注意力模块（Temporal Self-Attention module，TSA）来代替传统的图卷积方法。两种模块结构如图14 所示。空间和时间自注意力模块分别学习帧内关节交互和帧间运动，实现更灵活和全局性的特征提取。虽然自注意力机制能够建立全局关节点关系模型，但是并非所有的关节点都需要建立联系。

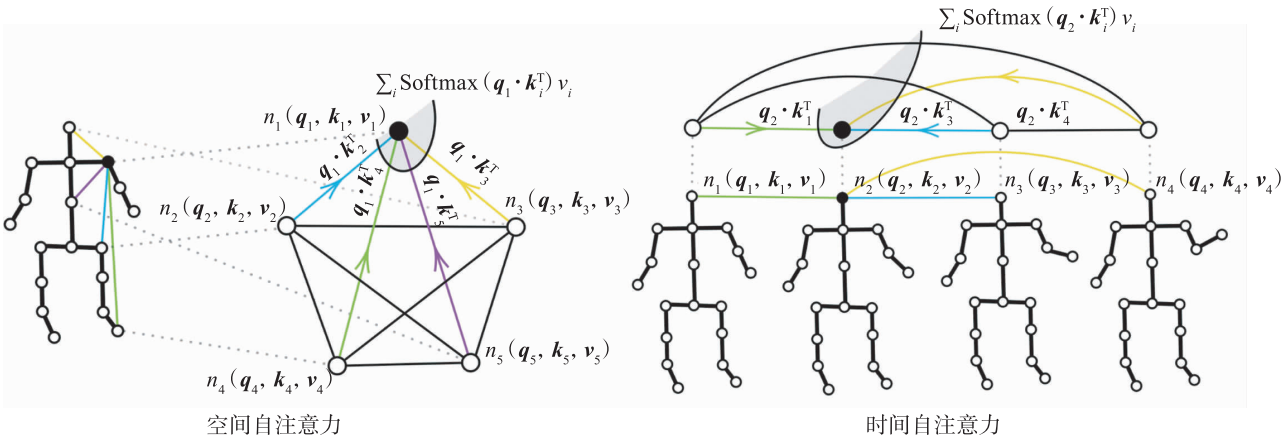

图 14 ST-TR 中的空间自注意力模块(SSA)和时间自注意力模块(TSA)[94]

Fig.14 Spatial Self-Attention (SSA) and Temporal Self-Attention (TSA) in ST-TR[94]

捕获全局和局部依赖关系会带来大量的冗余计算，因此高效地判断各关节的依赖关系、动作变化导致关节点改变的逻辑关系是提高 Transformer 模型在骨架动作识别应用中的关键。WANG 等[95]为了降低计算复杂性同时保持其高效性，将身体部位的概念引入到 Transformer 模型中，采用自动分区策略将关节编码分区，如图 15 所示。他们将人体骨架分为 5 个部位，各部位内部着意捕获内部关节间的依赖关系，同时探索各部位之间对人体运动的影响。该方法使模型更专注于身体部位而非各个独立的关节，从而降低单个噪音关节对信息编码的影响，使模型能够从部位内部和部位间捕获具有辨别性的特征，同时降低了计算成本。

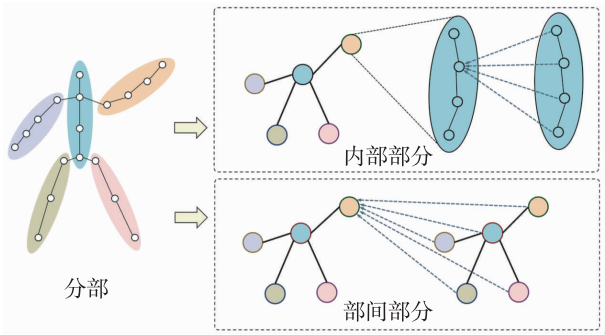

图 15 人体骨架部位划分[95]

Fig.15 Division of the human skeleton[95]

综上所述，Transformer 模型在骨架动作识别中，通过其强大的自注意力机制和并行处理能力，有效改善了传统图卷积网络在全局建模方面的不足。通过引入专用编码器、多任务自我监控模块、时空 Transformer 网络和高效的依赖关系建模策略，进一步提升了模型的性能和鲁棒性，使 Transformer 在骨架动作识别领域成为一个非常具有潜力的研究方向。表 5 列出了基于 Transforer 方法的骨架动作识别模型比较。

表 5 Transformer 动作识别模型性能比较

Tab.5 Performance comparison of action recognition model in Transformer

模型	发表时间	NTU RGB + D		NTU RGB + D 120		Kinetics-skeleton	创新工作
		CS	CV	C-sub	C-set		
ST-TR[94]	2021	90.3	96.3	85.1	87.1	38	时空 Transformer
STST[93]	2021	91.9	96.8			38.3	长时间运动、自监督数据增强
IIP-Transformer[95]	2021	92.3	96.4	88.4	89.7		多层级关节依赖关系，部分级数据增强
STTFormer[96]	2022	92.3	96.5	88.3	89.2		帧间关节交互信息
3Mformer[97]	2023	94.8	98.7	92	93.8	48.3	图节点超边超图建模

3　关键技术分析与讨论

骨架动作识别与传统的计算机视觉任务有着明显不同。人的动作是由整体姿态随时间变化构成的，所以人体动作识别需要对时间维度的变化进行建模，更多地依赖于视频数据而不是图片数据。动作识别任务普遍存在两个难点：一是如何高效地对时间维度进行建模，二是如何利用姿态变化提供的语义信息。

3.1　骨架建模

基于骨架的动作识别从人体姿态变化形成的轨迹入手，使用人体骨架状态的变化信息表征人体的运动信息。纵观基于骨架的动作识别发展历程，其背后的逻辑可以梳理为一个核心观点：构建能够概括人体运动的表征信息，为模型提供更丰富、更全面的输入信息。而构建更丰富更全面输入信息的两个解决方向是围绕时间维度与空间维度上进行人体运动信息建模展开。于是人们发展了两个技术路线实现以上目标，如图 16 所示：一个是先从时间维度入手，然后改进其空间建模能力；另一个是先从空间建模入手，接着提升其时间建模能力。

刚开始引入深度学习到动作识别领域时，人们试图从时间维度入手，采用擅长处理序列信息的 RNN 网络，将关节点序列变化为向量序列，再到后来的 LSTM 方法，更好地解决了 RNN 可能会产生梯度爆炸的问题。如图 16 上部所示，RNN/LSTM 经历了划分长短期、骨架分布等方法的发展，进一步丰富了空间和时间信息的利用。受限于本身结构，RNN/LSTM 方法虽然在处理时间维度上取得了不错的效果，但是依然无法完全模拟全身关节的复杂时空关系。后来，有学者将 LSTM 模块与其它方法组合起来使用以增强空间建模能力，如与 CNN 组合成双流网络、使用注意力机制关注重要节点、与 GCN 结合发现时间与空间的共现特征，进一步为模型识别提供了更丰富更全面的输入信息。

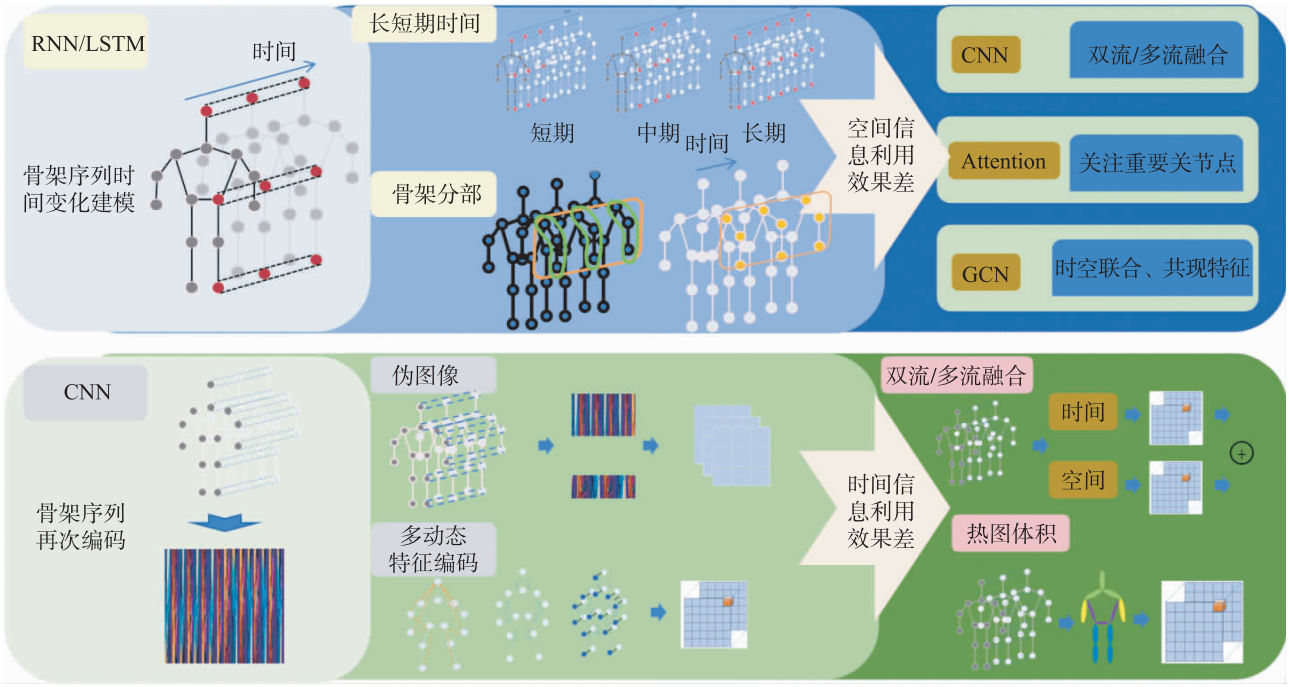

图 16　RNN/LSTM、CNN 方法骨架建模发展脉络

Fig.16　Skeleton modeling development lineage of RNN/LSTM and CNN methods

另一技术路线的思想是充分利用 CNN 强大的空间特征提取能力实现对骨架模态建模。为满足 CNN 的输入要求，首先将骨架模态处理为伪图像，然后对伪图像进行特征提取，在时间维度上与其他方法结合，如注意力机制、双流网络等。但是该方法一方面对长时间序列处理效果不佳，另一方面也没有很好地发挥骨架模态的独有优势。为了提升 CNN 在动作识别领域的性能，如图 16 下部所示，人们使用 3D CNN 将空间模态建模、时间建模统一起来。3D CNN 首先被用在基于 RGB 的动作识别上，且出现了许多改进版本，如 C3D[98]、I3D[99]。由于需要处理很多不必要的背景信息，这一类方法

会产生很大的计算量。之后研究者将 3D CNN 应用在骨架模态的时空建模上面，同时在更细粒度识别上进行改进。LIU[100]通过建立时空金字塔，融合时间维度上骨骼关键点的相关性。BANERJEE 等[73]编码多种动态特征变化为图像表示，然后使用 3D CNN 进行分类。后来 DUAN 等[74]开发出了 PoseConv3D 的新方法，与之前方法不同，PoseConv3D 在热图体积上进行特征提取，取得了良好的识别精度与速度。同时，因为其数据结构不受关节点个数的限制，在拓展性与泛化性上表现良好，且易与其他模态进行融合，拓展了骨架动作识别在各种场景下的适用性。

CNN 方法在骨架动作识别方面的进步总是围绕着对骨架序列进行重新编码或扩张，目的是将其转化为 CNN 方便处理的数据模式，这与 CNN 本身就是为图像数据设计有关。在骨架序列重新编码过程中，也是围绕不断获取更丰富、更全面的输入数据展开，从将简单的变化转化为伪图像，再到将距离变化、角度变化等动态信息编码进来，到最后将骨架序列进行热图膨胀转化为空间体积，逐步将骨架序列变换为可以表达丰富信息的数据形式，传输到对空间特征学习优秀的 CNN 网络之中。

骨架序列数据是由关节点变化构成的，传统的 RNN 和 CNN 往往无法对骨架模态关节点之间的结构关系进行很好的建模，导致它们的表现受限。因此一些学者将骨架模态中存在的关节点与连接线视为拓扑图结构，然后引进 GCN 方法进行建模。骨架数据天然形成图结构(关节点和连接线)，使用 GCN 可以更自然地建模关节点之间的关系，克服了其他方法在处理复杂时空关系上的不足。YAN 等[50]重新思考对于骨架模态的利用，指出了骨架和关节轨迹对光照变化和场景变化具有鲁棒性，且由于高精确度的深度传感器和姿态估计算法的进步，骨架信息很容易获得。同时，骨架是图形的形式，而不是 2 维或 3 维网格，以往方法[54]限制了对“局部区域”内关节轨迹的建模。于是引入 GCN，构建空间-时间 GCN(ST-GCN)从骨架数据自动学习空间和时间模式，来捕捉关节空间变化模式以及它们的时间动态。该模型可以在图和时间维度上整合信息，从而处理骨架模态的动作识别。

因为 ST-GCN 是手动设计的骨架，对于不相邻关节点的交互，动作识别效果不佳。之后的一些改进工作，如 AS-GCN[80]、2s-AGCN[78]，深化了骨架与关节点的依赖关系，更好地将骨骼的二阶信息(人体骨骼的长度和方向)与一阶信息(关节的坐标)结合起来改进拓扑共享方法[101]建模动态信道拓扑，动态地推断每个样本的通道式拓扑，从而有效地捕捉不同通道内关节点之间微妙的关系，提高了对骨架建模的表示能力。虽然 GCN 的拓扑结构可以根据不同样本动态变换，但是依然受限于固定的关节点连接，对于一些非直接连接关节点的交互动作，难以提取更高阶的语义信息。因此，HAO 等[102]设计了一个包含超图注意力机制和一个用于判别特征表示的改进残差模块的超图神经网络(Hypergraph Neural Network，Hyper-GNN)捕捉关节之间的非物理依赖关系。Hyper-GNN 通过超边缘(即非物理连接)结构来提取局部和全局结构信息，用于基于骨架的动作识别。

基于 GCN 的骨架模态动作识别方法发展脉络如图 17 上部所示。GCN 可以直接对骨架序列中的关节点进行处理，基于 GCN 的骨架动作识别方法也就围绕对关节点的处理进行发展。从形式上看，其发展过程由简单到复杂，即从局部区域直接连接的关节点拓展到全局非直接连接的关节点，再到对全身的关节点划分部位乃至多维度关节点构建超图。其发展过程与图像识别中 CNN 特征融合发展过程有相似之处，如表 6 所示。

CNN 主要在 2 维或 3 维像素图像空间上捕捉图像像素的空间关系，对图像进行网格化处理，将像素排列成 2 维或 3 维网格结构。通过图结构捕捉节点(如骨架关节)之间的空间关系，将数据表示为图结构，其中节点表示数据点(如关节)，边表示数据点之间的关系(如骨骼连接)，这种结构化表示能够更自然地处理非规则数据。这些方法皆是为建模更加复杂和多样的连接关系，而通过捕捉复杂、高阶关系和全局上下文信息来实现。

Transformer 最初在自然语言处理(NLP)领域取得了巨大成功，特别是在处理序列数据方面。再后来开始探索将自注意力机制与 Transformer 技术应用到骨架动作识别中，将序列数据表示为嵌入向量，利用位置编码捕捉时序信息，通过自注意力机制捕捉不同数据点之间的关系，在时间维度上捕捉动作的动态变化，实现结构化表示，增强对长时间依赖关系的建模能力，然后将自注意力机制扩展到时间和空间两个维度。一方面利用编码器解码器注意机制用于捕捉时间维度中的长动态与更好地利用帧间信息，另一方面利用时空自注意模块在空间和时间两个维度分别学习帧内关节交互和帧间

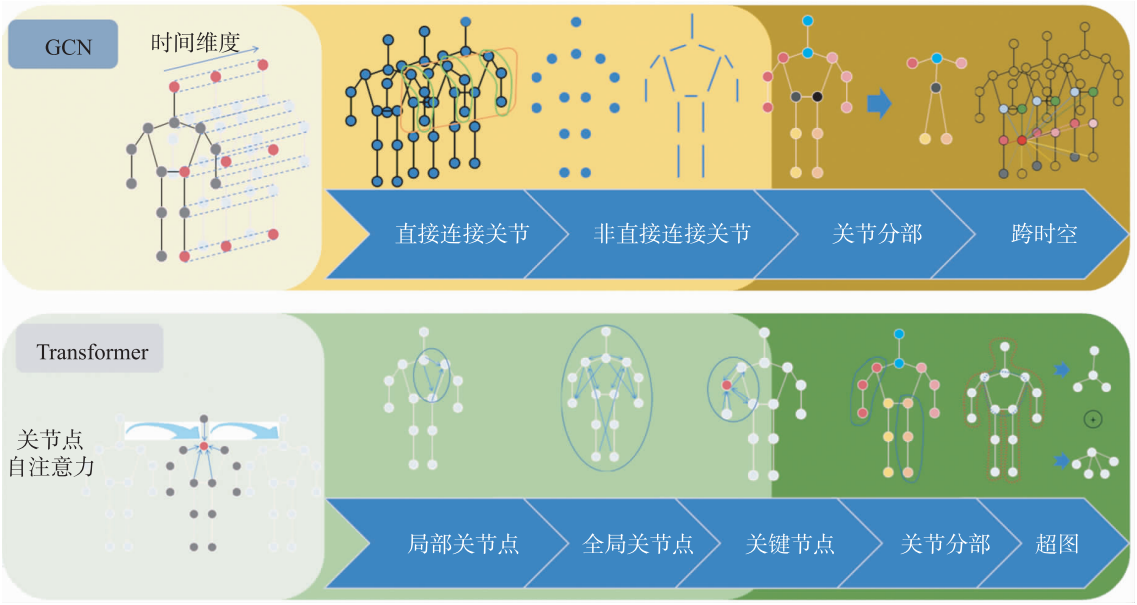

图 17　GCN、Transformer 骨架建模发展脉络

Fig.17　GCN, Transformer skeleton modeling development lineage

表 6　CNN、GCN、Transformer 方法相似性比较

Tab.6　Comparison of similarities among CNN, GCN and Transformer methods

相似性方面	CNN	GCN	Transformer
局部特征提取	利用卷积核在图像的局部区域滑动，提取局部特征；卷积操作通过多个滤波器在空间上提取图像的局部特征，从而捕捉图像的空间模式	在图结构上，利用邻接节点的信息进行特征提取，相当于在节点的局部邻域内滑动一个卷积核；通过聚合节点及其邻居节点的信息，捕捉图的局部结构特征	通过自注意力机制，关注局部特征，并在序列中捕捉局部信息
层级特征表示	通过多层卷积和池化操作，逐步提取更高层次的特征表示，捕捉图像的复杂模式和结构	通过多层图卷积操作，逐层聚合更多范围内的节点信息，逐步形成更高层次的特征表示，从而捕捉骨架数据中更复杂的时空关系	通过多层 Transformer 编码器堆叠，逐层提取更高层次的特征，捕捉复杂的时空依赖关系
非局部操作	非局部神经网络(Non-local Neural Networks)[103]通过计算特征图中所有位置之间的相似性来捕捉非局部特征，从而提高了模型的全局特征表示能力	在一个操作中连接多个节点，捕捉非局部的特征和关系	一次计算中捕捉输入序列中所有位置之间的依赖关系
特征融合	卷积层提取的特征通过层与层之间的连接和融合，形成更丰富的图像表示；这种融合可以通过跳跃连接、残差连接等方式实现，以保持特征的多样性和丰富性	通过聚合节点及其邻居节点的特征，实现特征融合；通过多层 GCN 的堆叠，可以融合更多层次的特征信息，从而捕捉骨架关节之间的复杂关系	通过多头自注意力机制，不同的注意头关注不同的特征子空间，实现特征的多样性和融合
全局上下文信息捕捉	使用全局卷积、全连接层或注意力机制来捕捉全局特征；全局卷积或全连接层用于在最后几层中整合全局信息；注意力机制，如自注意力(self-attention)和非局部神经网络(non-local neural networks)，也用于捕捉图像中的全局上下文	通过图卷积聚合更大范围内的节点信息，逐层形成全局特征表示；连接非直接相邻的节点，从而捕捉全局上下文信息，使用注意力机制来动态调整节点及其邻居的权重，捕捉节点之间的高阶关系，从而有效地提取重要特征	通过全局自注意力机制，捕捉序列中所有元素之间的依赖关系，实现全局上下文信息的捕捉
高级特征融合	多尺度特征提取和特征融合技术，如 FPN(Feature Pyramid Network)[104]和 Inception 模块[105]，通过多层次和多尺度的卷积操作捕捉丰富的特征表示	Hyper-GNN 能够捕捉节点之间的高阶依赖关系，超边(hyperedge)能够连接多个节点，从而在一个操作中捕捉复杂的局部和全局结构信息	通过多头自注意力机制、残差连接与层归一化、位置编码、前馈神经网络、编码器-解码器结构等多种机制，有效捕捉和融合复杂的时空关系

运动。

纵观骨架动作识别的发展历程，研究者们试图解决的两个问题，同时也是两个重大挑战，就是如何对时间与空间进行良好的建模。RNN/LSTM 在处理时间序列上比 GCN、CNN 能力更强，在动作识别任务中更擅长在时间维度进行建模；GCN 将骨架模态处理为拓扑结果能够构建空间模型中各关节点的依赖关系；普通 CNN 在空间维度上提取特征，后来 3D CNN 能够统一时间建模与空间建模；到了视觉 Transformer 在骨架动作识别上的应用，由于强大的注意力机制实现了更好的长时间序列与更好的空间依赖关系。

3.2 多模态融合技术

各种不同动作识别方法和模型不断刷新在各种数据集上的动作分类准确率，但是也越来越发现针对单一数据模态的方法存在识别性能上限。已经有不少研究表明，多模态融合、多技术融合能够有效提高综合识别性能。单一方法与单一模态的技术已经无法应对动作识别更深层次的挑战，需要向多技术、多模态融合方向发展。

不同模态的数据各有其特点：RGB 图像能够更好地包含识别所需要的各种信息，但是同时也容易受到环境变化的影响；红外模态能够在夜间或者是光照条件欠佳时发挥更好的成像作用；深度传感器能够提供更准确的 3 维位置信息；骨架模态可以在上述模态中提取出来，或通过动作捕捉设备获得，用来关注人体的细微动作，且具有较好的鲁棒性。研究者们也尝试将多种模态数据融合起来进行动作识别。骨架数据方便从其他多模态数据中提取出来，而且运算效率高，不容易受到环境因素干扰。因此研究如何在多模态融合中充分发挥骨架模态的特点，探索多技术、多模态融合受到了研究者们的广泛关注。前期的一些工作中已经出现了模态融合的方法，且取得了不错的效果，见表 7。

表 7 同时期融合模型与非融合模型性能比较

Tab.7 Comparison of performance between fusion model and non-fusion model in the same period

融合模型	发表年份	NTU RGB + D		融合方法	融合模态	同时期非融合模型	NTU RGB + D	
		CS	CV				CS	CV
ZHAO[106]	2017	83.7	93.7	特征融合	RGB 图像 + 骨架	GCA-LSTM[60]	74.3	82.8
SONG[107]	2018	92.6	97.9	特征融合		ST-GCN[50]	81.5	88.3
VPN(RNX3D101)[108]	2020	95.5	98	特征融合		Shift-GCN[81]	90.7	96.5
SGM-Net[109]	2020	89.1	95.9	分数融合		MS-G3D[82]	91.5	96.2
Pose-Conv3D[74]	2022	97	99.6	特征 + 分数融合		ST-GCN++[86]	92.6	97.4
LA-GCN[89]	2023	93.5	97.2	语言模型与图卷积融合	语言大模型 + 骨架	3M former[97]	94.8	98.7

已有许多研究在多模态融合角度开展。文[30]从多种数据模态角度对动作识别进行了综述调查，指出多模态机器学习通过聚合各种数据模态的优势和能力，处理和关联来自多模态的感知信息，通常可以提供更鲁棒和准确的 HAR。也有研究提出了多流网络，从骨骼、深度和 RGB 图像中提取特征，然后通过融合方法来获取最终的分类结果[110-111]。另一些研究探索了视觉和非视觉模态的协同学习，包括了使用注意机制进行知识蒸馏[112]、跨模态互补性模块和空间一致性模块[113-114]等方法，以实现视觉和非视觉模态之间的知识传输和融合。

综上所述，多模态融合技术在人体动作识别领域有着广泛的应用和研究。通过结合不同模态的信息，集合不同技术的识别优势，可以提高动作识别的准确性和鲁棒性，为相关领域的研究和实践提供重要的参考。

4 未来展望

基于人体骨架的动作识别在数据集构建、识别算法和模型以及应用等方面取得了相当引人瞩目的进展，但与人们的期望仍有一定差距。如目前的骨架数据只能表达四肢和躯干的动作，对于包括面部表情或手势等更细节、更全面的动作无法表达；现有骨架动作识别模型性能仍有很大局限，这可能源于模型本身的能力局限，也可能来源于训练不足或模型参数量不够；目前骨架动作识别的实际应用效果还很不理想。针对以上不足开展以下研究将是学

术界和产业方的共同期待。

1）全身姿态估计。

大部分骨架数据依赖于姿态估计技术获得，当前姿态估计越来越深化细粒度识别，已经发展到统一手部、脸部的全身姿态估计。基于骨架的动作识别往往依托于全身基本骨架，所以通常在 NTU_RGB + D 这种全身动作数据集上效果较好。而在面临只有部分基本骨架，譬如只有手部动作的场景下，因为之前的骨架序列还没有包含如此细节的运动信息，所以往往无法提供准确的动作识别。事实上，标准骨架数据只有 25 个关键点，对于手部细节、脸部细节无法表现出来，因为无法获得如此详解动作信息。

如图 18 所示，虽然已经有研究者在全身姿态估计面进行了探索[115]，但是还没有一个完善的全身姿态动作数据集提出，这个方向还是鲜有人探索。因此当在仅有一部分身体或只有手部动作的识别上，较全身动作识别效果较差。全身姿态估计能够更有效率地提供包含手部、脸部等更细节的姿态信息，为基于更细粒度的骨架动作识别提供基础。探索如何构建如此细粒度的动作识别模型是未来值得研究的方向。

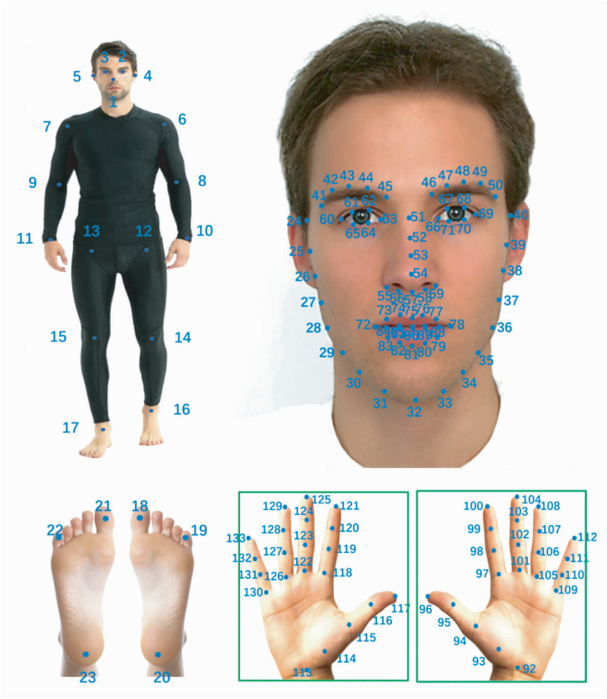

图 18　全身姿态估计[115]

Fig.18　Whole body pose estimation[115]

2）多模态学习。

不同数据模态在不同的场景下各有优势，因此使用多模态融合和跨模态迁移学习方法，通常可以实现互补，提高 HAR 的检测性能。目前最常用的融合方案是特征融合与分数融合（如表 7）。然而如文[116]所述，许多现有的多模态方法并不像预期的那样有效，依然面临着诸多挑战，如存在过度拟合、缺少高效率的融合策略等。

在早期模态融合中，像光流、骨架、红外等多模态数据在大方向上说可以归列于视觉模态内部融合的多模态学习。而如今的多模态学习希望实现在更多信息模态上进行融合学习，如在动作识别方面，有应用加速度传感器、雷达、WIFI 等信号进行动作识别的技术方案。这些技术也取得了一些进展，相对于视觉模态表现不会如图像那么直观，但是本质上也是关注人体关键关节的运动变化（即人体姿态变化）进行动作识别与分类的。这与骨架动作识别的出发点相同，那么是否可以使用“类姿态估计”方法将上述技术方案统一到骨架动作识别中呢？同理处理其他模态信号的方法能否应用在图像上呢？因此提高多模态学习在动作识别应用效率、思考如何发挥骨架模态计算效率高、专注于人体姿态变化的优势、找到更好处理骨架数据集的方法，依然对动作识别技术的发展产生着重要作用。

3）大模型技术。

如果将神经网络视为一种高效的信息压缩方式[117]，那么大模型显而易见比小模型拥有更多的参数量，可以压缩更丰富的信息，同时更深的网络层能够提取更抽象的特征信息，表征更广泛的共性信息，更宽的网络层能够保存更多的特征表现形式，提取到的特征信息更丰富。大模型已经展现了在开放世界中的零样本视频理解能力[118]，说明大模型拥有在开放世界中识别未训练动作类别的能力。一些研究者注意到了这一点，进行了基于大模型的动作识别研究[119－120]。但是在使用开放数据集测评模型识别能力的时候，依然只是在有限且固定类别的数据集中进行评估。这样的评估显然是不充分的，目前给出的评价结果只有平均正确率（Average Precision，AP）与全类平均正确率（mean Average Precision，mAP），对于整个动作识别任务来说，还只能做有限的对比，无法对模型整体进行完善的评价。基于以上分析，笔者认为基于大模型技术的 HAR，不能一味追求更大模型、更多数据的训练，而是需要设计更完善的大模型动作识别评价体系，构建多种测试输出与可视化方案，优化大模型对特定任务的结构设计。

4）新架构模型。

回溯骨架动作识别的发展历史，新模型架构的提出起到突出作用，3 维卷积实现了对时间空间两个维度的建模，将骨架动作识别两个研究方向统一了起来；图卷积神经网络 GCN 引入图的拓扑结构，将骨架数据作为拓扑图的进行处理，增强了骨架动作识别的可解释性并取得了良好的结果，开辟了骨架动作识别方法新的道路；Transformer 作为现在大模型的基础模型现在依然蓬勃发展，但是也已经有了很长历史了。当前骨架动作识别模型依然是一个热门的研究方向，但是似乎缺少像 3D 卷积、图卷积这种方法的重大突破，更多的是在对原有方法修修补补。

最近，ZHU 等[121]构建了统一框架来表达人的动作，在人体信息的基础上统一了多种下游任务。如图 19 所示，双流时空 Transformer（DSTformer）网络被用于构建运动编码器，从而全面、自适应地捕获骨骼关节之间的远程时空关系，然后将提取到的运动表征转移到多种下游任务中，实现了先进的性能。其核心贡献是提出了一个统一视角，从大规模、多样化的数据中学习人体运动的通用表征，进而为人体运动识别研究构建一个统一范式。从形式上看，这似乎统一了以人为核心的识别任务，但是究其原理依然只是量变而非质变（依靠运动编码器从大规模多样化的数据中提取运动表征）。因此，继续探究骨架动作识别方法可行的更深层次原因，构建骨架动作识别更具有代表的创新架构，将推动骨架动作识别向着更广阔的应用场景、更通用且详细的识别任务、更可靠更高效的识别效果发展。

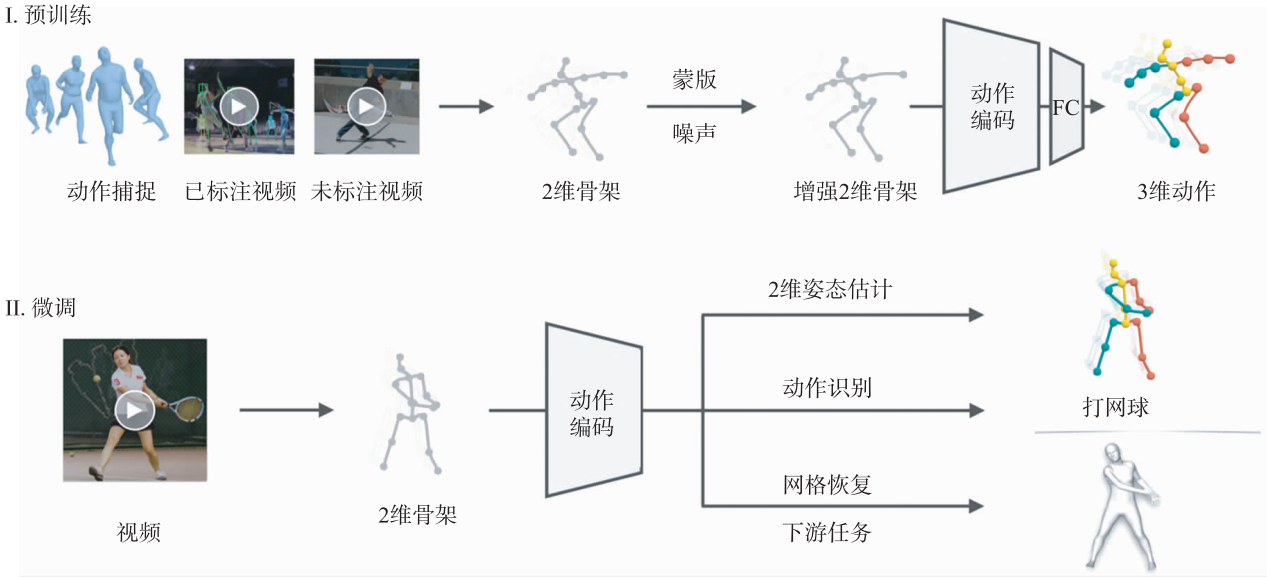

图 19 MotionBERT 模型框架[121]

Fig.19 MotionBERT model framework[121]

5）实际应用。

通过观察模型测试结果不难发现，大部分骨架动作识别模型在 NTU RGB + D 数据集识别效果较好，但是在 Kinetics-skeleton 识别结果较差。其原因在于 Kinetics 视频数据来源于用户上传，取自于真实场景，背景变化更多，动作会更多面临遮挡、截断等复杂情况，同时动作的类内差异会更加突出，模型识别效果下降是预料之中的。这也提醒研究者需要更多思考，如何在真实场景数据上增强识别的准确率与可靠性。在骨架信息不完整的情况下，如何有效发挥已知骨架位置等先验知识在预测动作中的关键作用，应该是未来研究者需要关注的一个核心问题。一方面需要更多的数据集去训练模型能够识别更多种类与更强泛化性，另一方面也要探索如何在保持模型性能的同时能够胜任更多面向特定场景、特定领域下的动作识别。

对于实际应用中信息缺失问题，多模态融合似乎是很好的解决办法，因为模态之间的优势互补可以明显提高动作识别的准确率，但也面临着计算量过大、模型需要的数据过多等限制。目前一些非视

觉模型与视觉方法的结合，以及预训练语言模型正在用于动作识别先验知识的生成，为该领域的研究提供新的可能。如文[122]使用预训练的大型语言模型作为知识引擎，自动生成动作身体部分运动的文本描述，为动作识别提供先验知识，并提出了一种多模态训练方案，通过利用文本编码器生成不同身体部分的特征向量，并监督骨骼编码器进行动作表示学习。

当前骨架动作识别已经在许多数据集上卓有成效。相对于其他模态技术或者多模态融合方案，基于骨架的动作识别通常有着更快速高效的性能，或许是动作识别场景落地最可行的方案之一。但是目前在现实中的应用还鲜有看到，因为优越性能通常是建立在高计算复杂度的基础上的。尤其是基于 Transformer 方法使用自注意力机制捕捉所有帧中所有关节之间的相关性问题必然消耗大量计算资源。因此，如何降低计算成本和资源消耗[123]（如 CPU、GPU 和能耗），实现高效快速的 HAR，值得进一步研究。在这方面与 RGB 数据模态相比，骨架模态数据量更小，计算效率更高。此外，相对于 RGB 数据集易受环境变化影响，骨架数据对光照变化和背景噪声具有鲁棒性，并且在现实世界中，骨骼数据集也可以保护我们的隐私。另一方面，由于许多实际应用中无法使用端到端的算法模型解决实际识别需求，需要将单一解决方案与其他技术结合起来使用，如吴胜昔等[124]结合目标检测与骨架姿态进行护具佩戴检测。

5　小结

动作识别是计算机视觉核心任务之一，在广泛的场景下具有重要研究意义与实际价值。基于骨架的动作识别方法因其对人体姿态变化的聚焦，具有独特的优势。基于骨架建模方法主要有 RNN、CNN、GCN 和 Transformer 四种技术路线，从其发展过程来看，就是在空间建模与时间建模两大核心问题上的不断探索和完善。面对动作识别发展更困难的挑战，未来的研究重点在于继续探索骨架建模背后的技术发展逻辑，建立更丰富的动作表达范式和数据集，研发更有效的动过识别模型，以及如何通过多模态融合来克服当前面临的计算量和数据需求问题，同时完善实际应用中模型计算量过大的问题。

目前算法模型的发展与改进思路总是向着种类更多、范围更广的方向发展，而忽略了在特定场景下解决特定动作识别需求，即如何在有限的数据以及硬件资源的限制下实现有着实际价值的动作识别应用。在这方面需要各行业研究者们的不断尝试，同时也希望不断提出和完善相关数据集，促进研究与应用两方面的更好发展。

参考文献

[1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2016: 770 – 778.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84 – 90.

[3] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. New York, USA: ACM, 2019: 6105 – 6114. .

[4] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2019: 1314 – 1324.

[5] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//International Conference on Computer Vision. Piscataway, USA: IEEE, 2017: 2961 – 2969.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2016: 779 – 788.

[7] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]//European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 21 – 37.

[8] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 213 – 229.

[9] 刘欢. 面向智能交通场景基于人体姿态的行人行为识别研究[D]. 吉林：吉林大学, 2023.

LIU H. Research on pedestrian behavior recognition based on human pose in intelligent transportation scenes[D]. Jilin: Jilin University, 2023.

[10] 李甜甜. 基于计算机视觉的学生课堂行为识别研究与应用[D]. 太原: 太原师范学院, 2023.
Li T T. Research and application of classroom behavior recognition based on computer vision[D]. Taiyuan: Taiyuan Normal University, 2023.

[11] 郑益群. 基于人体姿态估计的课堂教学行为识别[D]. 长春: 东北师范大学, 2023.
ZHENG Y Q. Teaching behavior recognition based on human pose estimation[D]. Changchun: Northeast Normal University Academic, 2023.

[12] 姚磊岳. 面向智慧安全管控的复杂人体行为识别技术研究及应用[D]. 南昌: 南昌大学, 2022.
YUE L Y. Research on complex human behavior recognition and its application in intelligent security management and control [D]. Nanchang: Nanchang University, 2022.

[13] KHALID N, GOCHOO M, JALAL A, et al. Modeling two-person segmentation and locomotion for stereoscopic action identification: A sustainable video surveillance system[J]. Sustainability, 2021, 13(2): 970 – 973.

[14] SEEMANTHINI K, MANJUNATH S S. Human detection and tracking using HOG for action recognition[J]. Procedia Computer Science, 2018, 132: 1317 – 1326.

[15] SINGH R, KUSHWAHA A K S, SRIVASTAVA R. Multi-view recognition system for human activity based on multiple features for video surveillance system[J]. Multimedia Tools and Applications, 2019, 78: 17165 – 17196.

[16] 张学琪. 复杂工业制造环境下的动作识别技术研究[D]. 杭州: 杭州电子科技大学, 2023.
ZHANG X Q. Research on motion recognition technology in complex industrial manufacturing environment[D]. Hangzhou: Hangzhou Dianzi University, 2023.

[17] 梁健. 基于人体骨架的工作流程识别系统研究[D]. 徐州: 中国矿业大学,2022.
LIANG J. Research on workflow recognition system based on human skeleton[D]. Xuzhou: China University of Mining and Technology, 2022.

[18] SHOTTON J, FITZGIBBON A, COOK M, et al. Real-time human pose recognition in parts from single depth images[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2011: 1297 – 1304.

[19] COLLOBERT R, WESTON J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//25th international conference on Machine Learning. Piscataway, USA: IEEE, 2008: 160 – 167.

[20] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding[J]. Advances in Neural Information Processing Systems, 2022, 35: 36479 – 36494.

[21] MAO Y, YOU C, ZHANG J, et al. A survey on mobile edge computing: The communication perspective[J]. IEEE Communications Surveys & Tutorials, 2017, 19(4): 2322 – 2358.

[22] SHAHROUDY A, LIU J, NG T T, et al. Ntu RGB + D: A large scale dataset for 3D human activity analysis[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2016: 1010 – 1019.

[23] ZHANG Z. Microsoft kinect sensor and its effect[J]. IEEE multimedia, 2012, 19(2): 4 – 10.

[24] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2017: 7291 – 7299.

[25] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019: 5693 – 5703.

[26] LIU J, DING H, SHAHROUDY A, et al. Feature boosting network for 3D pose estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(2): 494 – 501.

[27] MAJI D, NAGORI S, MATHEW M, et al. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2022: 2637 – 2646.

[28] ZHANG P, LAN C, ZENG W, et al. Semantics-guided neural networks for efficient skeleton-based human action recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 1112 – 1121.

[29] 张会珍，刘云麟，任伟建，等. 人体行为识别特征提取方法综述[J]. 吉林大学学报(信息科学版)，2020，38(3)：360－370.
ZHANG H Z，LIU Y L，REN W J，et al. Human behavior recognition feature extraction method：A survey[J]. Journal of Jilin University (Information Science Edition)，2020，38(3)：360－370.

[30] SUN Z，KE Q，RAHMANI H，et al. Human action recognition from various data modalities：A review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2022，45(3)：3200－3225.

[31] 毕春艳，刘越. 基于深度学习的视频人体动作识别综述[J]. 图学学报，2023，44(4)：625－639.
BI C Y，LIU Y. A survey of video human action recognition based on deep learning [J]. Journal of Graphics，2023，44(4)：625－639.

[32] 刘宝龙，周森，董建锋，等. 基于骨架的人体动作识别技术研究进展[J]. 计算机辅助设计与图形学学报，2023，35(9)：1299－1322.
LIU B L，ZHOU S，DONG J F，et al. Research progress in skeleton-based human action recognition[J]. Journal of Computer-Aided Design & Computer Graphics，2023，35(9)：1299－1322.

[33] WANG C，YAN J. A comprehensive survey of RGB-based and skeleton-based human action recognition[J]. IEEE Access，2023，11：53880－53898.

[34] ELMAN J L. Finding structure in time[J]. Cognitive Science，1990，14(2)：179－211.

[35] ALBAWI S，MOHAMMED T A，AL-ZAWI S. Understanding of a convolutional neural network[C]//International Conference on Engineering and Technology. Piscataway，USA：IEEE，2017：1－6.

[36] WU S，SUN F，ZHANG W，et al. Graph neural networks in recommender systems：A survey[J]. ACM Computing Surveys，2022，55(5)：1－37.

[37] LIU X，PENG H，ZHENG N，et al. Efficientvit：Memory efficient vision transformer with cascaded group attention[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，USA：IEEE，2023：14420－14430.

[38] XIN W，LIU R，LIU Y，et al. Transformer for skeleton-based action recognition：A review of recent advances[J]. Neurocomputing，2023，537：164－186.

[39] WANG J，LIU Z，WU Y，et al. Mining actionlet ensemble for action recognition with depth cameras[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，USA：IEEE，2012：1290－1297.

[40] HUSSEIN M E，TORKI M，GOWAYYED M A，et al. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations[C]//Twenty-third International Joint Conference on Artificial Intelligence. New York，USA：ACM. 2013：327－348.

[41] LECUN Y，BENGIO Y，HINTON G. Deep learning[J]. Nature，2015，521(7553)：436－444.

[42] RAHMANI H，MAHMOOD A，HUYNH D，et al. HOPC：Histogram of oriented principal components of 3D pointclouds for action recognition[C]//European Conference on Computer Vision. Berlin，Germany：Springer，2014：742－757.

[43] RAHMANI H，MAHMOOD A，HUYNH D，et al. Histogram of oriented principal components for cross-view action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，38(12)：2430－2443.

[44] JANG J，KIM D，PARK C，et al. ETRI-activity3D：A large-scale RGB-D dataset for robots to recognize daily activities of the elderly[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway，USA：IEEE，2020：10990－10997.

[45] LI T，LIU J，ZHANG W，et al. UAV-HUMAN：A large benchmark for human behavior understanding with unmanned aerial vehicles[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，USA：IEEE，2021：16266－16275.

[46] QIAN Y，CHEN C，TANG L，et al. Parallel LSTM-CNN Network with radar multispectrogram for human activity recognition [J]. IEEE Sensors Journal，23(2)：1308－1317.

[47] ZHAO P，LU C X，WANG B，et al. CubeLearn：End-to-end learning for human motion recognition from raw mmwave radar signals[J]. IEEE Internet of Things Journal，10(12)：1302－1321.

[48] OWENS A，EFROS A A. Audio-visual scene analysis with self-supervised multisensory features[C]//European Conference on Computer Vision . Berlin，Germany：Springer，2018：631－648.

[49] GAO R, OH T H, GRAUMAN K, et al. Listen to look: Action recognition by previewing audio[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 10457 – 10467.

[50] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//AAAI Conference on Artificial Intelligence. Menlo Park, USA: AAAI, 2018: 7444 – 7452.

[51] LIU J, SHAHROUDY A, PEREZ M, et al. NTU RGB + D 120: A large-scale benchmark for 3D human activity understanding [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10): 2684 – 2701.

[52] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[EB/OL]. (2017 – 05 – 19)[2023 – 10 – 08]. http: //arXiv. org/abs/1705. 06950v1.

[53] LI C, WANG P, WANG S, et al. Skeleton-based action recognition using LSTM and CNN[C]//IEEE International Conference on Multimedia & Expo Workshops. Piscataway, USA: IEEE, 2017: 585 – 590.

[54] ZHANG S, LIU X, XIAO J. On geometric features for skeleton-based action recognition using multilayer LSTM networks[C]// Winter Conference on Applications of Computer Vision. Piscataway, USA: IEEE, 2017: 148 – 157.

[55] ZHANG S, YANG Y, XIAO J, et al. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks[J]. IEEE Transactions on Multimedia, 2018, 20(9): 2330 – 2343

[56] ZHAO R, WANG K, SU H, et al. Bayesian graph convolution LSTM for skeleton based action recognition[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2019: 6882 – 6892.

[57] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2015: 1110 – 1118.

[58] LEE I, KIM D, KANG S, et al. Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks[C]//International Conference on Computer Vision. Piscataway, USA: IEEE, 2017: 1012 – 1020.

[59] CUI R, ZHU A, ZHANG S, et al. Multi-source learning for skeleton-based action recognition using deep LSTM networks[C]// International Conference on Pattern Recognition. Piscataway, USA: IEEE, 2018: 547 – 552.

[60] 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, 2017, 27(4): 1586 – 1599.

[61] LI B, DAI Y, CHENG X, et al. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN[C]//IEEE International Conference on Multimedia & Expo Workshops. Piscataway, USA: IEEE, 2017: 601 – 604.

[62] ZHENG W, LI L, ZHANG Z, et al. Relational network for skeleton-based action recognition[C]//IEEE International Conference on Multimedia and Expo. Piscataway, USA: IEEE, 2019: 826 – 831.

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

[64] ZHANG P, LAN C, XING J, et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data[C]//International Conference on Computer Vision. Piscataway, USA: IEEE, 2017: 2117 – 2126.

[65] FRIJI R, DRIRA H, CHAIEB F, et al. Geometric deep neural network using rigid and non-rigid transformations for human action recognition[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2021: 12611 – 12620.

[66] KENDALL D G. Shape manifolds, procrustean metrics, and complex projective spaces[J]. Bulletin of the London Mathematical Society, 1984, 16(2): 81 – 121.

[67] 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. Berlin, Germany: Springer, 2016: 816 – 833.

[68] LI S, LI W, COOK C, et al. Independently recurrent neural network (INDRNN): Building a longer and deeper RNN[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 5457 – 5466.

[69] DING Z, WANG P, OGUNBONA P O, et al. Investigation of different skeleton features for CNN-based 3D action recognition [C]//IEEE International Conference on Multimedia & Expo Workshops. Piscataway, USA: IEEE, 2017: 617 – 622.

[70] WANG P, LI Z, HOU Y, et al. Action recognition based on joint trajectory maps using convolutional neural networks[C]// ACM International Conference on Multimedia. New York, USA: ACM, 2016: 102 – 106.

[71] CAETANO C, SENA J, BRÉMOND F, et al. SkeleMotion: A new representation of skeleton joint sequences based on motion information for 3D action recognition[C]//IEEE International Conference on Advanced Video and Signal Based Surveillance. Piscataway, USA: IEEE, 2019: 1 - 8.

[72] ZHANG P, LAN C, XING J, 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.

[73] BANERJEE A, SINGH P K, SARKAR R. Fuzzy integral-based CNN classifier fusion for 3D skeleton action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(6): 2206 - 2216.

[74] DUAN H, ZHAO Y, CHEN K, et al. Revisiting skeleton-based action recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2022: 2969 - 2978.

[75] LIU H, TU J, LIU M. Two-stream 3D convolutional neural network for skeleton-based action recognition[EB/OL]. (2017 - 05 - 23)[2023 - 07 - 12]. http://export.arxiv.org/abs/1705.08106.

[76] WU Z, PAN S, CHEN F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(1): 4 - 24.

[77] LI M, CHEN S, CHEN X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019: 3595 - 3603.

[78] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019: 12026 - 12035.

[79] CHO S, MAQBOOL M, LIU F, et al. Self-attention network for skeleton-based human action recognition[C]//IEEE Winter Conference on Applications of Computer Vision. Piscataway, USA: IEEE, 2020: 635 - 644.

[80] SONG Y F, ZHANG Z, WANG L. Richly activated graph convolutional network for action recognition with incomplete skeletons [C]//IEEE International Conference on Image Processing. Piscataway, USA: IEEE, 2019: 1 - 5.

[81] CHENG K, ZHANG Y, HE X, et al. Skeleton-based action recognition with shift graph convolutional network[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 183 - 192.

[82] LIU Z, ZHANG H, CHEN Z, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 143 - 152.

[83] HUANG L, HUANG Y, OUYANG W, et al. Part-level graph convolutional network for skeleton-based action recognition[C]// AAAI Conference on Artificial Intelligence. Menlo Park, USA: AAAI, 2020: 11045 - 11052.

[84] 费树岷，赵宏涛，杨艺等. 基于时序拓扑非共享图卷积和多尺度时间卷积的骨架行为识别[J]. 信息与控制，2023，52(6)：758 - 772.
FEI S M, ZHAO H T, YANG Y, et al. Temporal Topology unshared graph convolution and multiscale temporal convolution for skeleton-based action recognition[J]. Information and Control, 2023, 52(6): 758 - 772.

[85] CHEN Y X, ZHANG Z Q, YUAN C F, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2021: 13339 - 13348

[86] DUAN H, WANG J, CHEN K, et al. PYSKL: Towards good practices for skeleton action recognition[C]//ACM International Conference on Multimedia. New York, USA: ACM, 2022: 7351 - 7354.

[87] CHI H, HA M H, CHI S, et al. INFOGCN: Representation learning for human skeleton-based action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2022: 20186 - 20196.

[88] LONG N H B. STEP CATFormer: Spatial-temporal effective body-part cross attention transformer for skeleton-based action recognition[EB/OL]. (2023 - 11 - 06)[2023 - 12 - 03]. http:// arXiv.ag/abs/2312.03288.

[89] XIANG W, LI C, ZHOU Y, et al. Generative action description prompts for skeleton-based action recognition[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2023: 10276 - 10285.

[90] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//International Conference on Neural Information Processing Systems. New York, USA: ACM, 2017: 6000 - 6010.

[91] HAN K, WANG Y, CHEN H, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 87 - 110.

[92] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[EB/OL]. (2020 - 03 - 09)[2023 - 05 - 12]. http://arxiv.org/abs/1012.07825050.

[93] ZHANG Y, WU B, LI W, et al. STST: Spatial-temporal specialized transformer for skeleton-based action recognition[C]// ACM International Conference on Multimedia. New York, USA: ACM, 2021: 3229 - 3237.

[94] PLIZZARI C, CANNICI M, MATTEUCCI M. Spatial temporal transformer network for skeleton-based action recognition[C]// ICPR International Workshops and Challenges: Virtual Event. Berlin, Germany: Springer, 2021: 694 - 701.

[95] WANG Q, PENG J, SHI S, et al. IIP-Transformer: Intra-inter-part transformer for skeleton-based action recognition[EB/OL]. (2020 - 03 - 09)[2023 - 05 - 12]. http://arxiv.org/abs/2110.13385.

[96] QIU H, HOU B, REN B, et al. Spatio-temporal tuples transformer for skeleton-based action recognition[EB/OL]. (2022 - 01 - 080[2023 - 06 - 12]. http://arxiv.org/abs/2201.02489.

[97] WANG L, KONIUSZ P. 3Mformer: Multi-order Multi-mode Transformer for Skeletal Action Recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2023: 5620 - 5631.

[98] HARA K, KATAOKA H, SATOH Y. Learning spatio-temporal features with 3D residual networks for action recognition [C]// International Conference on Computer Vision workshops. Piscataway, USA: IEEE, 2017: 3154 - 3160.

[99] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2017: 6299 - 6308.

[100] LIU J, AKHTAR N, MIAN A. SKEPXELS: Spatio-temporal image representation of human skeleton joints for action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition workshops. Piscataway, USA: IEEE, 2019: 10 - 19.

[101] IMRAN J, RAMAN B. Deep residual infrared action recognition by integrating local and global spatio-temporal cues[J]. Infrared Physics & Technology, 2019, 102: 217 - 232.

[102] HAO X, LI J, GUO Y, et al. Hypergraph neural network for skeleton-based action recognition[J]. IEEE Transactions on Image Processing, 2021, 30: 2263 - 2275.

[103] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 7794 - 7803.

[104] LIN T Y, DOLL? R P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE,2017: 2117 - 2125.

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

[106] ZHAO R, ALI H, VAN DER SMAGT P. Two-stream RNN/CNN for action recognition in 3D videos[C]//IEEE International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2017: 4260 - 4267.

[107] SONG S, LAN C, XING J, et al. Skeleton-indexed deep multi-modal feature learning for high performance human action recognition[C]//IEEE International Conference on Multimedia and Expo. Piscataway, USA: IEEE, 2018: 1 - 6.

[108] DAS S, SHARMA S, DAI R, et al. VPN: Learning video-pose embedding for activities of daily living[C]//European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 72 - 90.

[109] LI J, XIE X, PAN Q, et al. SGM-Net: Skeleton-guided multimodal network for action recognition[J]. Pattern Recognition, 2020, 104: 1103 - 1126.

[110] WANG L, DING Z, TAO Z, et al. Generative multi-view human action recognition[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2019: 6212 - 6221.

[111] WANG P, LI W, WAN J, et al. Cooperative training of deep aggregation networks for RGB-D action recognition[C]//AAAI Conference on Artificial Intelligence. Menlo Park, USA: AAAI, 2018: , 32(1): 923 - 939.

[112] RADEVSKI G, GRUJICIC D, BLASCHKO M, et al. Multimodal distillation for egocentric action recognition[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2023: 5213 - 5224.

[113] LIU Y, WANG K, LI G, et al. Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition[J]. IEEE Transactions on Image Processing, 2021, 30: 5573 - 5588.

[114] PEREZ A, SANGUINETI V, MORERIO P, et al. Audio-visual model distillation using acoustic images[C]//IEEE Winter

Conference on Applications of Computer Vision. Piscataway, USA: IEEE, 2020: 2854 – 2863.

[115] JIN S, XU L, XU J, et al. Whole-body human pose estimation in the wild[C]//European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 196 – 214.

[116] WANG W, TRAN D, FEISZLI M. What makes training multi-modal classification networks hard? [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 12695 – 12705.

[117] TISHBY N, ZASLAVSKY N. Deep learning and the information bottleneck principle[C]//Information Theory Workshop. Piscataway, USA: IEEE, 2015: 812 – 838.

[118] PU S, ZHAO K, ZHENG M. Alignment-uniformity aware representation learning for zero-shot video classification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2022: 19968 – 19977.

[119] KE Q, BENNAMOUN M, AN S, et al. Learning clip representations for skeleton-based 3d action recognition[J]. IEEE Transactions on Image Processing, 2018, 27(6): 2842 – 2855.

[120] NI B, PENG H, CHEN M, et al. Expanding language-image pretrained models for general video recognition[C]//European Conference on Computer Vision. Berlin, Germany: Springer, 2022: 1 – 18.

[121] ZHU W, MA X, LIU Z, et al. MotionBERT: A unified perspective on learning human motion representations[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2023: 15085 – 15099.

[122] SHE D, LAI Y K, YI G, et al. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2021: 8475 – 8484.

[123] LIN J, GAN C, HAN S. TSM: Temporal shift module for efficient video understanding[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2019: 7083 – 7093.

[124] 吴胜昔，咸博龙，冒鑫鑫，等. 基于姿态估计的护具佩戴检测与动作识别[J]. 信息与控制，2021，50(6)：722 – 730，739.
WU S X, XIAN B L, MAO X X, et al. Protective wearing detection and action recognition based on pose estimation[J]. Information and Control, 2021, 50(6): 722 – 730, 739.

作者简介

孟祥璞(2000 –)，男，硕士生。研究领域为姿态估计，目标检测，动作识别。

李　硕(1996 –)，男，硕士生。研究领域为目标检测，动作识别，缺陷检测。

苑明哲(1971 –)，男，研究员，博士生导师。研究领域为智能控制，边缘计算。