基于深度学习的群体行为识别: 综述与展望

朱晓林[1], 王冬丽[2], 欧阳万里[3], 李抱朴[4], 周 彦[2†], 刘金富[2]

(1. 湘潭大学 数学与计算科学学院, 湖南 湘潭 411105; 2. 湘潭大学 自动化与电子信息学院, 湖南 湘潭 411105;
3. 悉尼大学 电气与信息工程学院, 悉尼 澳大利亚 2006; 4. 百度美国研究院, 森尼韦尔 美国 94086)

摘要: 群体行为识别是计算机视觉领域中备受关注的研究方向, 在智能监控系统和体育运动分析等领域中具有广泛的应用推广价值. 本文对过去七年来基于深度学习的群体行为识别方法进行了全面综述, 有助于更好推动群体行为识别的发展. 首先, 介绍群体行为的定义、通用识别流程以及主要的挑战; 其次, 从群体行为识别的建模方法和内在机理进行划分, 并进一步细分类、讨论和分析这些方法的优缺点; 然后, 给出群体行为识别的常用数据集, 列举了相关的开源代码库和评估指标; 最后, 对该领域未来的研究方向进行了展望.

关键词: 群体行为识别; 深度学习; 层级时序建模; 交互关系推理; Transformer

引用格式: 朱晓林, 王冬丽, 欧阳万里, 等. 基于深度学习的群体行为识别: 综述与展望. 控制理论与应用, 2024, 41(12): 2207 – 2223

DOI: 10.7641/CTA.2023.20375

Group activity recognition based on deep learning: Overview and outlook

ZHU Xiao-lin[1], WANG Dong-li[2], OUYANG Wan-li[3], LI Bao-pu[4], ZHOU Yan[2†], LIU Jin-fu[2]

(1. School of Mathematics and Computational Science, Xiangtan University, Xiangtan Hunan 411105, China;
2. School of Automation and Electronics Information, Xiangtan University, Xiangtan Hunan 411105, China;
3. School of Electrical and Information Engineering, The University of Sydney, Sydney 2006, Australia;
4. Baidu Research (USA), Sunnyvale, CA 94086, USA)

Abstract: Group activity recognition has attracted much attention in the computer vision community, and it is widely applied in intelligent monitoring systems and sports video analysis. This paper provides a comprehensive review of the group activity recognition methods based on deep learning over the past seven years, which will help to promote the development of group activity recognition. First, the definition, the general recognition process, and the main challenges of group activity are introduced; Secondly, we classify the group activity recognition methods in modeling and internal mechanism, subdivide them, and further discuss the advantages and disadvantages of these methods; Thirdly, we present the common datasets of group activity recognition, the relevant open-source code libraries, and the evaluation index; Finally, we analyze the future research directions in group activity recognition.

Key words: group activity recognition; deep learning; hierarchical temporal modeling; interaction relationship reasoning; Transformer

Citation: ZHU Xiaolin, WANG Dongli, OUYANG Wanli, et al. Group activity recognition based on deep learning: Overview and outlook. *Control Theory & Applications*, 2024, 41(12): 2207 – 2223

1 引言

近年来, 随着社会智慧化和人工智能技术的发展, 面向大规模场景的群体行为识别(group activity recognition, GAR)已成为计算机视觉领域中备受关注的研究方向, 同时也是群体视频分析与理解的核心技术, 更是后续群体行为预测、异常行为预警的关键基础性

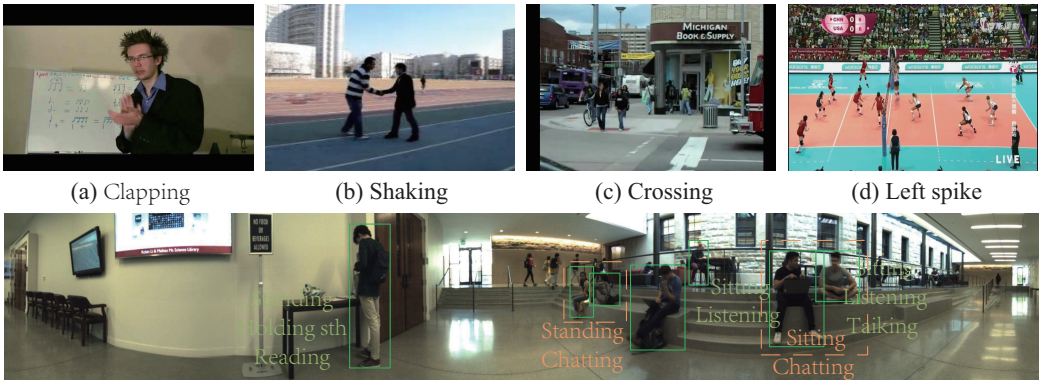
收稿日期: 2022−05−10; 录用日期: 2023−05−11.
†通信作者. E-mail: yanzhou@xtu.edu.cn; Tel.: +86 15200377751.
本文责任编委: 柯良军.
国家重点研发计划项目(2020YFA0713503), 国家自然科学基金项目(61773330), 国家航空科学基金项目(20200020114004), 湖南省科技创新计划项目(2020GK2036), 湖南省自然科学基金项目(2023JJ30598), 湖南省研究生科研创新项目(CX20220652)资助.
Supported by the National Key Research and Development Project of China (2020YFA0713503), the National Natural Science Foundation of China (61773330), the Aeronautical Science Foundation of China (20200020114004), the Science and Technology Innovation Program of Hunan Province (2020GK2036), the Natural Science Foundation of Hunan Province (2023JJ30598) and the Postgraduate Research Innovation Project of Hunan Province (CX20220652).

技术. 群体行为识别以视频中运动人群的行为分析和理解为研究目的, 旨在推断复杂场景中一群人进行的整体活动.

目前单人行为识别已取得令人满意的性能, 且研究状态趋于饱和. 而群体行为更加广泛地存在人类社会中, 具有重要的理论价值和广阔的应用前景. 其最突出的两个应用是智能监控系统和体育运动分析. 1) 监控设备的广泛应用迅速增加了视频数据量, 分析和理解复杂的视频内容已经成为智能视频监控领域的重要研究热点, 构建自动化、智能化的视频监控分析平台成为迫切需求; 2) 体育运动的规模化和职业化进程推动了群体行为分析研究的发展, 例如一群人和一个物体不断地进行交互(22人的足球比赛、12 人的排球比赛、10人的篮球比赛等). 识别体育运动中的群体活动(足球比赛中的射门、排球比赛中的扣杀和传球、篮球比赛中的投篮和抢断等)能够帮助评估团队战术策略, 制定合理的球员训练方案以及向媒体提供相关资讯.

国内外研究者根据人体行为活动中执行动作的个体数量不同, 可分为一个行为者的单人行为识别、两个行为者的交互行为识别、多个行为者的群体行为识别. 其中, 单人行为是指单个人的基本运动行为(跑步、过马路等), 或者可以视为多个简单姿势动作(挥手、抬脚等)的组合; 交互动作一般指的是人与人的交互或者人与物的交互, 如握手、看书等, 也是目前较受关注的人体行为识别类型; 群体行为是指一个场景中包含多人和多物的活动, 如排球比赛、团体会议等, 是最复杂也是目前亟需解决的人体行为识别类型. 为对拥挤场景中的行为类型进行更全面的理解, Han等人[1]首次提出全景行为识别 (panoramic human activity recognition, PAR), 它集成了3个子任务: 个体行为识别(任务I)旨在识别场景中每个个体的动作; 社交群组活动识别(任务II)旨在将人群划分为社交群组并识别他们的活动; 全局行为识别(任务III)侧重于对整个场景的整体活动理解. 图1是不同类型的人体行为识别示例, 其图左上角的标签表示行为类型, 图1(e)中个体边界框(绿色实线)的标签表示个体行为, 群组边界框(橙色虚线)的标签表示社交活动.

由于场景中一般有多个个体, 甚至有多个群组, 且群组和个体的位置存在相对变化, 使得群组间、个体间存在相互遮挡以及不同行为类别间存在相似性等挑战. 如图2(a)–(b)所示, 仅使用外观特征区分(a)中的walking行为和(b)中的crossing行为是存在混淆的, 因为这两种行为都是从一个点移动到另外一个点; 如图2(c)所示, 并非所有个体行为在群体行为中都同等重要, 带红星的行为者对识别右扣球行为发挥重要作用, 同时左拦网与右扣球行为者构成时空上的交互关系. 因此, 如何提高复杂非结构化场景中群体行为识别的准确性、鲁棒性是值得持续关注的研究方向之一.



(a) Clapping (b) Shaking (c) Crossing (d) Left spike

(e) Resting conversing

图 1 不同类型的人体行为识别示例图

Fig. 1 Examples of different types of human activity recognition

(a) Walking (b) Crossing (c) Right spike

图 2 不同类型的群体行为识别示例图

Fig. 2 Examples of different types of group activity recognition

目前, 已有大量学者对单人动作识别[2–3]和人体行为识别[4–5]进行了综述, 但是针对群体行为识别的综述较少. Vahora等人[6]全面回顾了不同的群体行为识别方法并进行分类, 但仅是根据手工特征和可学习的特征描述符, 并没有涉及深度学习方法. 裴利沈等人[7]从使用深度神经网络架构的不同归类梳理群体行为识别方法, 着重于讨论分析群体行为识别核心网络架构的优缺点. Wu等人[8]从手工特征和深度学习两方面进行全面的综述, 是目前较为详实的群体行为识别综述文献. 本文则着重从群体行为识别建模方法的角度进行介绍, 补充了基于图的群体行为识别关系推理方法及多模态融合策略等. 进一步探索群体行为的内在机理, 并对基于深度学习的群体行为识别方法进一步细分类, 梳理了群体行为识别近几年非常火热的Transformer以及最具代表性的(state-of-the-art, SOTA)方法.

基于上述分析, 本文从基于深度学习的群体行为识别的建模方法和内在机理进行划分, 着重介绍长短时记忆网络[9](long short-term memory, LSTM)的层级时序建模方法、图卷积神经网络[10](graph convolutional network, GCN)的交互关系推理方法、Transformer模型[11]的自注意力机制及多模态融合策略, 同时还总结了半监督方法、弱监督方法, 以及仅使用骨架数据的方法等. 本文的贡献总结如下: 1) 对基于深度学习的群体行为识别方法进行分类, 并对群体行为识别的建模方法和内在机理进行细分类; 2) 总结了群体行为识别的基准数据集、开源代码库、评价指标; 3) 指出了群体行为识别领域未来的研究方向.

2　群体行为识别的相关介绍

2.1　群体行为的定义

百度百科中“群体行为”的定义: 群体行为是团体行为的一种特殊形式, 为了实现某个特定的目标, 由两个或更多的相互影响、相互作用、相互依赖的个体组成的人群集合体. 群体行为并不是简单个体之间的相互叠加, 在群体中, 个体会受到其他个体的影响, 从而会呈现出完全不一样的群体行为. 群体行为决定着个体行为的方向, 个体行为是群体行为的体现. 群体行为识别旨在理解每个个体的行为, 以及他们在群体环境中如何相互作用.

2.2　群体行为的通用识别流程

给定一段由K个人进行群体行为的视频段, 均匀地采样T帧提取个体特征. 将第t帧中第k个人的表示记为X_t^k, 将第t帧的时序表示记为H_t, 其中$t \in \{1,\cdots,T\}$, $k \in \{1,\cdots,K\}$. 一般地, 通过骨干网络提取静态卷积神经网络(convolutional neural network, CNN)特征作为个体行为者的空间表示, 然后利用LSTM等方法从静态个体表示中进行时间动态建模. 将空间CNN特征X和时序特征$H=\{h_1,\cdots,h_T\}$拼接成时空特征$\tilde{X} \in \mathbb{R}^{T\times K\times D}$, D是特征向量的长度. 采用层级时序建模、关系推理建模等推理方法进一步增强特征表示, 融合生成最终的高层次特征表示, 用于个体行为和群体行为识别分类, 图3是群体行为的通用识别流程图.

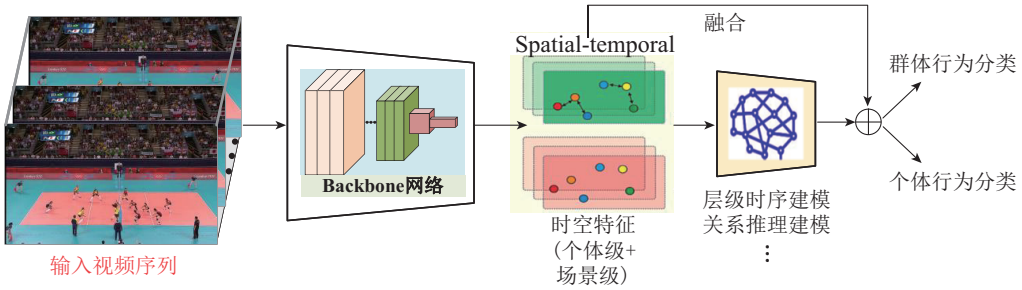


图 3　群体行为的通用识别流程图

Fig. 3　The general flowchart of group activity recognition

2.3　群体行为识别主要的挑战

目前群体行为识别仍面临着巨大挑战, 这些挑战主要与视觉特征相关[12], 如行为类内的差异性和类间的相似性、个体行为者间的相互遮挡等. 通过文献调研, 群体行为识别的主要挑战总结如下:

1) 强有力的时空关系及上下文信息表征: 群体行为识别涉及多人在一个场景中执行不同的动作及进行不同的交互. 因此, 推断群体行为识别需要理解个体行为者之间的时空关系和上下文语境信息, 抓住个体之间关系的时空演变对群体行为识别是至关重要的.

2) 复杂场景中的多群组及关键行为者: 一个场景中可能同时存在多个群组活动, 如何确定区分多个子群的行为类型及关注贡献度大的关键行为者, 这是面对大规模实际场景应用的一个挑战.

3) 高精度、鲁棒、实时的端到端检测–跟踪–识别框架: 以往大多数研究者将检测跟踪技术视为群体行为识别之前的一项解耦任务, 如何更好地将中级检测

跟踪任务和高级识别任务结合起来, 形成端到端的群体行为识别框架, 是目前群体行为识别研究的一个挑战.

4) 视觉特征存在的隐私道德及模型泛化能力差等问题: 人工智能应用程序中的隐私道德问题已经越来越凸显, 基于视觉特征容易侵犯用户的隐私[13]. 另外, 基于视觉特征容易产生场景偏差, 造成模型泛化能力差, 如何解决仅基于视觉特征输入造成模型迁移差的问题也是目前需解决的一个难题.

5) 更大、更具挑战性的统一数据集: “AI界一直缺少一套系统的理念与方法整合在不同领域不同任务不同数据集上的不同成果”, 数据驱动的深度学习技术往往需要更大、更具挑战性的统一数据集. 缺乏统一的行为类别定义及标签的标注, 限制了相关研究工作的开展.

3 基于深度学习的群体行为识别方法

随着计算机硬件性能的提升和高性能计算技术的提高, 深度学习模型在图像分类[14]、目标检测[15]、语义分割[16]、多目标跟踪[17]、行为识别[2]等领域取得了明显的性能提升, 为相关任务提供了一种准确、高效、可靠的表示学习方法. 基于深度学习的群体行为识别性能远远优于基于传统手工特征的方法, 已成为主流方法并不断地扩展群体行为识别的深度和广度. 如图4所示, 从群体行为识别的建模方法和内在机理将现有的基于深度学习的群体行为识别方法大致分为5类: 层级时序建模、注意力建模、交互关系建模与推理、多模态融合策略, 以及其他方法. 具体如下:

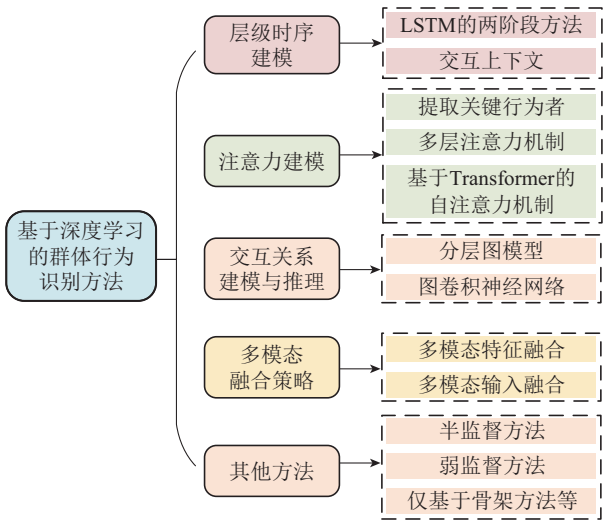


图 4 基于深度学习的群体行为识别方法分类

Fig. 4 The taxonomy of group activity recognition methods based on deep learning

3.1 层级时序建模

与传统的行为识别相比, 群体行为识别需进一步探索个体行为者之间的时序信息. 先前的工作通过设计各种手工特征的图模型[18]或AND-OR语法模型[19]学习时序特征, 但其识别精度较低且模型依赖于特定场景. 得益于深度学习的成功, 递归神经网络[20](recurrent neural network, RNN)作为一种处理可变长度序列数据的神经网络, 在群体行为识别的序列数据建模方面取得了优良性能.

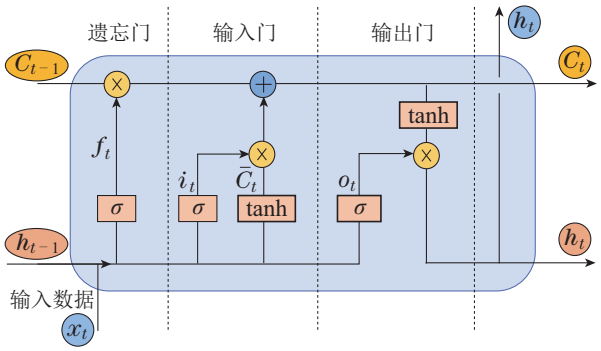


图 5 LSTM的架构图[9]

Fig. 5 The architecture of LSTM

给定一段T帧的视频$\{x^t|t=1,\cdots,T\}$, 其中x^t是第t帧的静态CNN特征[21], 利用LSTM学习运动状态序列$\{h^t|t=1,\cdots,T\}$描述该段视频中个体行为状态. 时序建模通过LSTM学习隐藏状态以描述视频序列的动态信息. 图5是LSTM的架构示意图, 主要由输入门、遗忘门、输出门、输入调制门和记忆单元状态组成, 则时刻t 的LSTM单元可以表示为

$$i^t=\sigma\left(W_{ix}x^t+W_{ih}h^{t-1}+b_i\right), \tag{1}$$

$$f^t=\sigma\left(W_{fx}x^t+W_{fh}h^{t-1}+b_f\right), \tag{2}$$

$$o^t=\sigma\left(W_{ox}x^t+W_{oh}h^{t-1}+b_o\right), \tag{3}$$

$$\tilde{C}^t=\varphi\left(W_{gx}x^t+W_{gh}h^{t-1}+b_C\right), \tag{4}$$

$$C^t=f^t\odot C^{t-1}+i^t\odot\tilde{C}^t, \tag{5}$$

$$h^t=o^t\odot\varphi\left(C^t\right), \tag{6}$$

其中: i^t, f^t, o^t, $\tilde{C}^t$和C^t分别是输入门、遗忘门、输出门、输入调制门和记忆单元状态; $\sigma(\cdot)$是sigmoid激活函数; $\odot$表示对应元素相乘; $\varphi(\cdot)$表示双曲正切$\tanh(\cdot)$函数, W_{*x}和W_{*h}是权重矩阵; b_*是偏置向量. 具体地说, 输入门i^t在时间步t控制新输入数据的贡献, 以更新记忆单元. 而遗忘门f^t则决定前一状态C^{t-1}的内容有多少保留到当前状态C^t. 输出门o^t则学习如何从当前状态C^t的记忆单元中导出LSTM单元在时间步t的输出.

1) LSTM的两阶段方法: 层级时序建模最初是基于LSTM的两阶段解决方案, 该方法利用CNN和LSTM作为基网络, 捕获个体行为者的时间动态特征. Ibrahim等人[22]首次引入深度学习框架, 设计基于LSTM的两阶段分层深度时间动态模型(hierarchical deep temporal model, HDTM). 如图6所示, 该方法首

先提取场景中每个个体行为者的时间动态特征, 然后聚合个体的特征表示识别群体行为. 具体来说, 第1阶段将个体级LSTM应用于每个个体的轨迹以提取空间和时间变化特征. 第2阶段采用群体级LSTM, 将个体层次的信息进行池化聚合形成群体层次特征. 该方法是首次应用CNN-LSTM框架解决群体行为识别的工作, 此后许多基于“CNN-LSTM”的群体行为识别方法[23–26] 如雨后春笋般涌现. Li等人[25]提出一种新颖的基于语义的群体行为识别两阶段方案, 即基于语义的群体行为识别(semantics based group activity recognition, SBGAR), 采用CNN-LSTM框架比先前的方法具有更高的准确性. Al-Habib等人[26]提出基于LSTM网络的统一深度学习框架, 用于检测多个群组并识别其相应的群体行为. 使用场景级CNN和LSTM学习场景中深层次的群体级特征以推断群体行为.

针对LSTM直接级联的脆弱性以及底层LSTM产生的误差会不断积累传播到更高层的问题. Shu等人[23]设计能量层代替softmax层进行行为预测, 同时利用p值(即事件发生的可能性)来计算最具置信度的能量值. 通过设计置信度–能量循环网络(confidence-energy recurrent network, CERN)扩展了现有的LSTM两阶段结构, 额外增加置信度测量和能量机制, 从不同的语义层次对个人行为、个体之间的交互行为和群体行为进行识别. 由于以上基于LSTM的两阶段方法忽略了一个基本事实, 即个人层面的行为和群体层面的活动是随着时间的推移而发生的. 为此, Shu等人[27]提出图嵌套LSTM(graph LSTM in LSTM, GLIL)网络, 该网络联合建模个体层面和群体层面的行为.

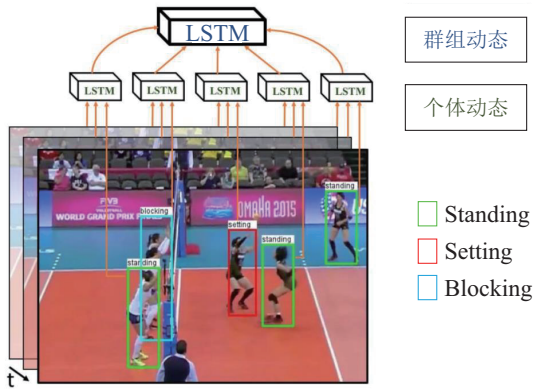


图 6　基于LSTM的两阶段模型[22]

Fig. 6　Two-stage model based on LSTM[22]

LSTM两阶段方法总体上可划分为3步[28]: 1) 首先将场景中的行为者检测区域块进行导出, 依据人体轨迹构建行为图像序列, 随后将该行为图像序列输入CNN和低层LSTM, 提取个体行为者对应的空间特征和时序变化特征; 2) 在得到个体时空特征的基础上, 对所有个体特征进行池化聚合, 并把聚合结果输入高层LSTM以提取群体行为特征; 3) 将高层LSTM输出的群体行为特征送入全连接层, 经过Softmax函数得到不同的类别得分.

2) 交互上下文: 为建模行为者之间的交互上下文信息, Wang等人[29]提出一种基于LSTM网络的分层循环交互上下文建模框架(recurrent modeling of interaction context, RMIC). 如图7所示, RMIC主要通过个体级LSTM网络捕获单个行为者的动态, 然后利用上下文编码器对群组级和场景级交互上下文进行建模, 将编码结果送入群组级和场景级LSTM进行群体行为识别. 该方法的主要特色是使用上下文编码器生成多级交互上下文, 处理单个个体、群组内个体交互和群组间的交互, 进而构建高阶上下文建模方案并产生更具判别力的群体交互特征.

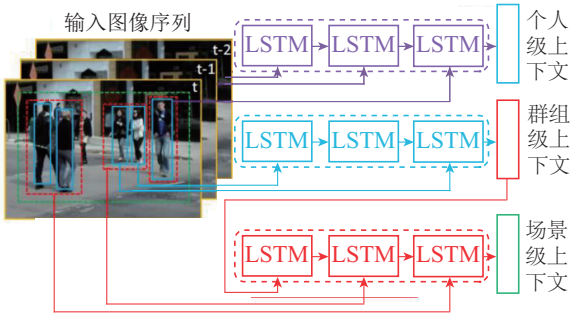


图 7　分层循环交互上下文建模框架[29]

Fig. 7　The modeling framework of hierarchical recurrent interactional context[29]

为解决相似的个体运动在不同群体行为中造成混乱的问题, Kim等人[30]提出一种考虑显著子事件的判别性群体上下文特征(discriminative group context feature, DGCF), 由个体属性和行为者轨迹的显著子事件属性组成. 所提出的DGCF描述符可以在场景中生成子组以减少相似运动的影响, 同时行为者之间的关系通过轨迹数据进行区分, 并采用轨迹增强方法减少深度网络中的过拟合问题.

在时序上下文方法的基础上, 研究者将行人检测跟踪与行为识别看作耦合任务, 进而设计统一的端到端框架. Zhuang等人[31]提出一种端到端的差分递归卷积神经网络(differential recurrent CNN, DRCNN). 该方法不依赖于轨迹检测, 利用CNN的语义表示和堆叠差分长短时记忆网络的记忆状态对时空信息进行建模, 对时空信息进行深度融合以实现端到端的群体行为识别. Bagautdinov等人[32]提出统一的社交场景理解框架(social scene understanding, SSU), 在未剪辑的视频序列上解决3个任务: 多人检测、个人行为识别和群体行为识别. SSU框架对原始图像序列进行预处理, 并且依赖于多尺度特征融合以及特征提取层微调, 使模型能够捕获上下文和交互关系. 由于生成对抗网络[33](generative adversarial networks, GAN)能够学习

数据的原始分布, 可有效克服深度网络的局限性. Gammulle等人[34]首次尝试将GAN引入到群体行为识别任务中, 提出一种基于LSTM的半监督、多层序列生成对抗网络(Multilevel sequence GAN, MLS-GAN)结构. 该网络允许模型学习中间动作级表示, 以发现个体之间交互的上下文信息. 半监督MLS-GAN框架通过学习个体级和场景级特征, 引入LSTM网络和门控单元以实现特征融合, 同时考虑长时间依赖性, 探索个体行为者间的交互关系实现群体行为识别.

3.2 注意力建模

Deng等人[35]验证了一些个体行为者与整个群体行为可能无关, 将所有的个体行为进行等效建模会带来混淆信息, 应重点关注少数关键参与者, 即对群体行为识别起作用的关键行为者和关键帧. 针对上述问题, 可通过注意力建模提高群体行为识别的性能. 基于注意力建模的群体行为识别方法可分为: 提取关键行为者、多层注意力机制, 以及基于Transformer的自注意机制.

1) 提取关键行为者: 为捕获关键行为者的动态信息, 抑制无关行为者的信息, Yan等人[36]提出基于参与贡献度的时间动态模型(participation contributed temporal dynamic model, PCTDM). 图8是行为者运动强度图, PCTDM重点关注在整个过程中稳步移动(移动时间较长)或在重要时刻剧烈移动(与群体活动密切相关)的行为者, 进而抑制无关的个体行为, 提高群体行为识别的准确率. 在[36]的框架上进行扩展, 提出一种新的位置感知参与贡献时间动态模型[37](position-aware PCTDM, P^{2}CTDM), 设计位置感知交互模块, 同时考虑特征相似性和位置信息用于群体行为识别.

Tang等人[38]首次通过探索语义域中的先验知识, 显式建模群体行为者的交互注意力. 提出一种新颖的语义注意力保持师生(semantics preserving teacher-student, SPTS)模型, 利用教师网络中的注意力知识引导学生网络, 旨在挖掘语义保持注意力, 自动关注关键行为者, 忽略无关的个体行为者. 针对传统群体特征建模忽略不同群组之间的内在依赖性, 继续扩展SPTS框架, 分别在语义域和外观域构造两类图网络[39](节点表示提取的个体特征, 边用于描述不同个体之间的关系), 以解释不同群组之间的依赖关系. 由于以往方法中认为所有个体行为者对群体活动的贡献相等, Tang等人[40]提出时空上下文一致性(spatio-temporal context coherence, STCC)约束和全局上下文一致性(global context coherence, GCC)约束, 以捕获关键行为者并量化它们对群体活动的贡献. 文献[41]提出基于空间通道和混合通道的双重注意力模型, 动态地为每个特征分配注意力权重, 并关注成员间的相互依赖性以及提取关键行为者.

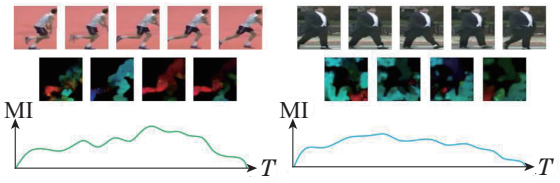


图 8 行为者运动强度图[36]

Fig. 8 The motion intensity (MI) of actor[36]

2) 多层注意力机制: 群体行为识别的关键在于明确地捕捉不同层次的复杂时空交互, 在个体层面上, 每个行为者不仅依赖于自身的时空特征, 还依赖于场景中其他个体提供的互动信息. 针对视频中群体行为识别的两个主要挑战: 1)如何重点关注对群体行为有关键贡献的行为者; 2)如何对行为者之间的上下文结构进行有效建模. Kong等人[42]通过设计分层注意力网络(hierarchical attention networks, HAN)以关注不同交互的行为者. 进一步构建分层上下文网络(hierarchical context networks, HCN), 反复建模组内/组间上下文结构. 通过整合视觉注意力和上下文结构, 该框架提升了群体行为识别模型的判别力.

以“群体行为识别的关键在于明确地捕捉不同层次的复杂时空交互”为出发点, Lu等人[43]提出基于时变注意机制的两级注意互动模型: 1) 个体级注意机制以姿势特征为输入, 获取场景中个体之间不同程度的互动; 2)场景级注意机制采用池化策略, 以探索个体行为与高层次群体行为之间的交互, 通过改进的两级门控递归单元(gated recurrent unit, GRU)处理长时间的时变性和一致性, 在个体级和场景级上明确地对交互关系进行建模. 针对以往研究方法没有联合探索个体和群体层面的不平衡互动关系, Lu等人[44]提出一种嵌入图注意力模块的图注意交互模型(graph attention interaction model, GAIM), 在统一框架中自适应地学习个人和群体层面上不同程度的交互关系, 并进一步自动编码个人行为和群体行为的时空演化.

3) 基于Transformer[11]的自注意机制: 自注意机制能够对群体行为识别任务中行为者的互动和交互关系进行推理.

在自注意力层中, 输入向量首先被转换成3个不同的向量: 查询向量q、键值向量k和值向量v, 维度为$d_q = d_k = d_v = d_{\text{model}} = 512$. 来自不同输入的向量被压缩到$Q$, K, V矩阵中, 不同输入向量之间的注意力函数计算如下:

步骤 1 计算不同向量之间的分数$S = QK^{\mathrm{T}}$.

步骤 2 归一化分数$S_n = S/\sqrt{d_k}$.

步骤 3 使用softmax函数将归一化分数转换为概率$P = \text{softmax}\ (S_n)$.

步骤 4 获得权重值矩阵$Z = VP$.

上述过程可简化为

$$A(Q,K,V)=\mathrm{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V, \tag{7}$$

多头注意力机制, 通过叠加多层自注意力层提升模型性能. 给定一个输入向量和头数h, 与自注意力层处理机制类似, 生成3组矩阵$\{Q_i\}_{i=1}^{h}$, $\{K_i\}_{i=1}^{h}$和$\{V_i\}_{i=1}^{h}$, 则多头注意力机制的计算如下:

$$\begin{cases}\mathrm{MultiHead}(Q',K',V')=\\ \mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_h)W^{\mathrm{o}},\\ \mathrm{head}_i=A(Q_i,K_i,V_i),\end{cases} \tag{8}$$

式中: W^{o}是投影权重, Q_i是$\{Q_i\}_{i=1}^{h}$的串联操作.

Gavrilyuk 等人[45]首次将Transformer引入群体行为识别任务中, 提出行为者Transformer框架(actor-transformers for group activity recognition, AT-GAR)细化以及聚合个体级特征, 并选择性地提取与群体行为识别相关的信息. 如图9所示, AT-GAR 采用 2D pose 网 络 (high-resolution network, HRNet[46]) 和 3D CNN (I3D[47])提取特定行为者的静态和动态特征表示, 并送入Transformer学习行为者之间的交互关系. 为充分探索个体之间的时空交互, 产生合理的群体表征, Li 等人[48]提出一种新颖的群体行为识别框架GroupFormer, 联合捕获时空上下文信息, 通过聚类时空Transformer有效地增强个体和群体表示. 图10是GroupFormer框架图, 主要贡献如下: 1) GroupFormer对时空依赖关系进行综合建模, 利用交叉解码器建立时空信息之间的联系; 2) 利用聚类自注意机制动态地将所有个体划分为多个子群, 以更高效地学习具有行为感知能力的语义表征. 文献[49]针对上述方法[45, 48]没有细化个体级特征直接输入到自注意机制中, 设计以动作为中心的聚合策略来学习全局的动作级特征, 从而弥补个体级特征和群组表示之间的差距.

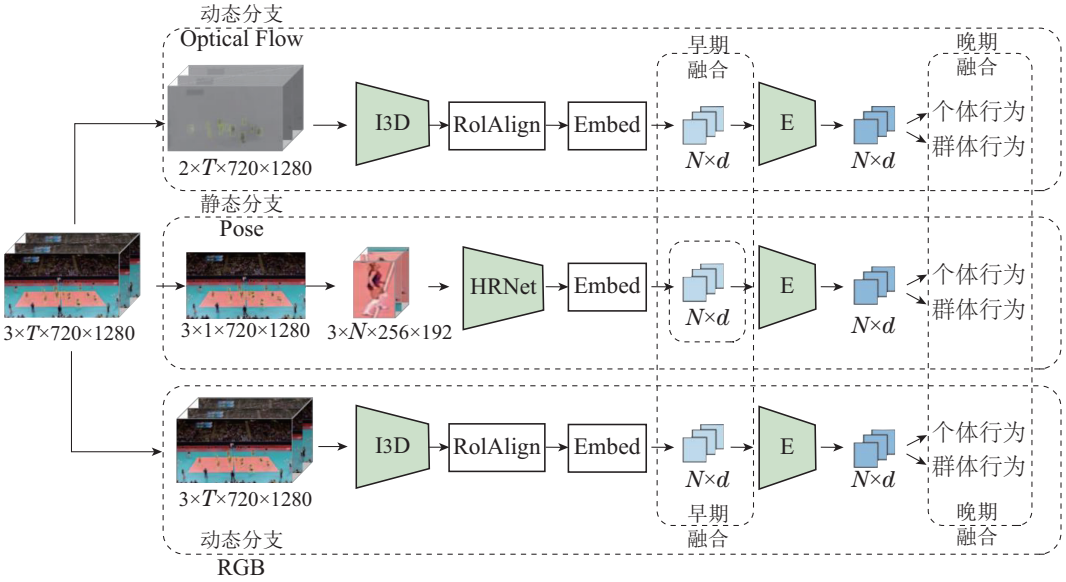


图 9　行为者Transformer框架[45]

Fig. 9　The framework of Actor-Transformer[45]

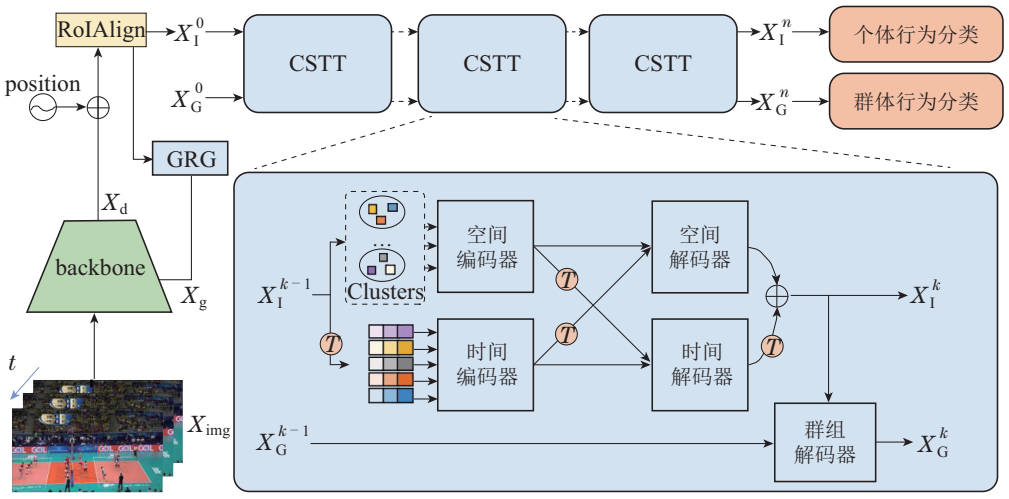


图 10　GroupFormer框架图[48]

Fig. 10　The framework of GroupFormer[48]

以上群体行为识别方法并没考虑场景的多尺度信息, 同时还存在隐私和道德层面的问题. Zhou等人[12]只使用关键点模态数据, 提出一种多尺度Transformer框架, 在不同场景下利用基于自注意力的推理, 构建多尺度组合推理的细粒度群体行为识别方法Composer, 同时使用对比聚类、辅助预测和数据增强技术改进中间表示, 从而提高群体行为识别的整体性能.

Han等人[50]提出双路径行为者交互框架(dualpath actor interaction, Dual-AI), 以两种互补的顺序灵活地堆叠空间Transformer和时间Transformer, 同时设计自监督一致性的多尺度(即帧(frame)级别和视频(video)级别)行为者对比损失函数以增强互补性. Yuan等人[51]为充分探索视觉上下文在群体行为识别中的作用, 提出基于Transformer的视觉上下文编码器用于群体行为识别(learning visual context GAR, LVC-GAR), 提取行为者的视觉上下文以整合全局上下文信息, 并设计时空双线性池化策略挖掘丰富的成对交互关系, 证明视觉上下文在群体行为识别中的有效性.

为进一步将自注意力机制与群体行为识别结合, Pramono等人[52]首次将条件随机场(conditional random field, CRF)和自注意机制结合, 即SACRF, 共同建模群体行为识别中个体交互者的时空关系. 接着, Pramono等人在文献[52]的框架上扩展, 利用HRNet[46]捕获姿势特征, 并在空间自注意力结构中引入计算两个节点距离的线性函数, 以提高模型学习局部关系的能力, 进一步采用对比损失来处理个体与群体关系的上下文信息, 有效提高群体行为识别的精度[53].

3.3 交互关系建模与推理

1) 分层图模型: 针对场景中个体间的结构化关系建模问题, Ibrahim等人[54]提出分层关系网络(hierarchical relational networks, HRN), 计算个体行为者之间的关系表示, 利用图结构描述行为者潜在的交互关系. 其关系层的输入是个体初始特征表示和潜在的关系图, 输出是个体间的关系表示. 堆叠多个关系层以学习紧凑的关系表示用于群体行为识别. Yan等人[55]提出基于层次图的交叉推理网络用于群体行为识别, 即分层图交叉推理网络(hierarchical graph-based cross inference network, HiGCIN). 用端到端的方式构建、学习和推断多层次(身体区域、个体级和群组级)的识别模型.

为充分挖掘群体行为识别任务中丰富的语义关系, 例如个体间、群组间的互动和空间交互关系, Deng等人[35]提出图模型和深度神经网络集成的框架, 基于RNN的顺序推理机制在节点之间的连接上施加门函数进行深层次结构关系推理. 采用上述结构推理机制(structure inference machines, SIM)进行迭代学习, 能够对低级别的网络输出进行增强处理, 对更高级别推理任务实现有效学习.

为充分考虑个体交互信息以建模行为者之间的空间关系, Azar等人[56]提出端到端训练的深度卷积神经网络, 称为卷积关系机(convolutional relational machine, CRM). 通过引入中间表示(活动图), 学习并提取视频帧中个体间复杂的空间关系. 受文献[57]的启发, 进一步使用多阶段细化模块减少活动图中的错误预测, 通过聚合模块集成各阶段的细化特征用于识别群体行为. Qi等人[58]提出基于语义图和时空注意的语义递归神经网络框架stagNet: 采用语义图提取动态场景中行为者间的时空关系, 通过nodeRNNs和edgeRNNs进行推理, 该模型使用“因子共享”和“消息传递”机制同时预测场景标签和行为者关系; 进一步整合时空注意机制关注视频中的关键行为者, 从而提高行为识别性能; 在文献[58]的框架上进行扩展, 提出行为者身体区域注意机制以及全局–局部特征池策略提高个体动作识别的性能, 同时在4个公共数据集上表现出优越的性能[59].

2) 图卷积神经网络: 近年来, 将图模型与深度神经网络相结合是深度学习研究中的新兴课题. 图卷积网络是CNN在图上的推广, 可以处理非欧几里德数据, 已广泛应用于计算机视觉, 例如点云分类、动作识别和交通预测.

Wu等人[60]首次将GCN引入群体行为识别任务中, 构建灵活且有效的行为者关系图(actor relation graphs, ARG). 图11是ARG用于群体行为识别, 其中每个节点表示一个行为者, 每条边表示行为者之间的关系, 并同时捕捉行为者之间的外观和位置关系用于群体行为识别. 采用ARG在时空图上学习个体间的交互, 进一步应用空间局部化和稀疏时间采样策略进行交互关系推理, 以实现群体行为识别. ARG 模型利用图结构显式地建模交互关系, 图中的节点表示行为者 $A=\{(x_i^{\rm a},x_i^{\rm s})|i=1,\cdots,N\}$, 其中$N$为行为者的数量, $x_i^{\rm a}\in\mathbb{R}^d$是第$i$个行为者的外观特征, $x_i^{\rm s}\in(t_i^{\rm x},t_i^{\rm y})$是第$i$个行为者边界检测框的中心坐标. 构建图$G\in\mathbb{R}^{N\times N}$表示行为者间的交互关系, 关系值$G_{ij}$表示第个$j$行为者对第$i$个行为者的重要性.

为获得强有力的表征以捕捉行为之间的潜在关系, 需考虑外观特征和位置信息. 此外, 注意到外观关系和位置关系具有不同的语义属性. 关系值定义为以下复合函数:

$$G_{i,j}=h\left(f_{\rm a}\left(x_i^{\rm a},x_j^{\rm a}\right),f_{\rm s}(x_i^{\rm s},x_j^{\rm s})\right),\tag{9}$$

其中$f_{\rm a}(x_i^{\rm a},x_j^{\rm a})$表示两个行为者之间的外观关系, 同样地, 通过$f_{\rm s}(x_i^{\rm s},x_j^{\rm s})$计算位置关系. 函数$h$将外观和位置关系进行归一化.

Yuan等人[61]提出用于视频群体行为识别的时空

动态推理网络(dynamic inference network, DIN), 如图12所示, 在时空图的推理过程中引入可变形卷积的思想, 通过在局部时空交互域上对中心人物的全局交互图进行预测并更新特征, 从而解决在群体行为识别中出现的过平滑和高计算量的问题. 通过设计动态推理(dynamic relation, DR)模块和动态游走(dynamic walk, DW)模块用于捕获交互场特征以增强行为者之间的交互语境, 在群体行为识别中达到最新技术(state of the art, SOTA)的效果. 由于图卷积只考虑成对的交互关系(一对一), 文献[62]提出多超边超图结构捕获多行为者的高阶关系(多对一). 超边包含灵活数量的超节点, 能够对非成对关系进行建模.

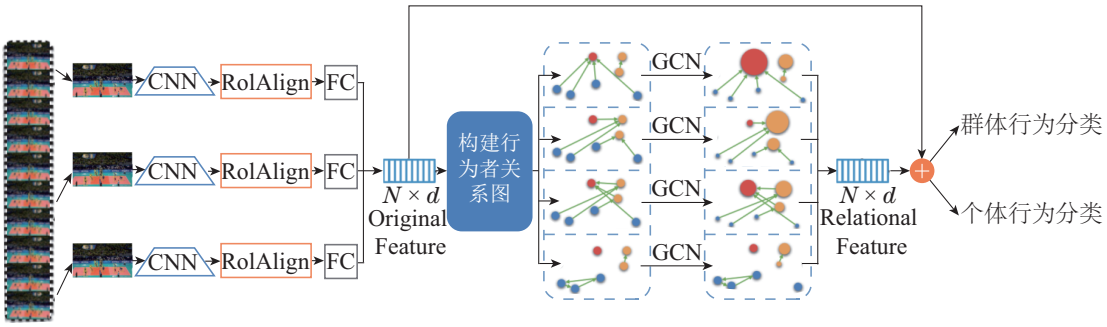


图 11　行为者关系图用于群体行为识别[60]

Fig.11　The actor relation graphs for group activity recognition[60]

为对拥挤场景中的活动进行更全面的理解, Han等人[1]提出全景人体行为识别的解决方案, 旨在同时实现个体行为、社会群体活动和全局活动的识别. 该方案采用一种改进的层次图神经网络, 以逐步表示和建模细粒度的人体行为和群体行为的交互社会关系. 在不同图之间进行知识迁移具有广阔的应用前景, 但鲜有研究. Tang等人[63]提出图交互网络(graph interaction networks, GINs)模型在图的权重矩阵之间迁移知识以保留交互关系信息, 探索基于骨架的跨数据无监督动作识别和基于多模态输入的全监督群体行为识别, 实验结果表明GINs的有效性. Hu等人[64]首次利用强化学习策略解决群体行为识别, 提出渐进关系学习(progressive relation learning, PRL)以提取与群体行为相关的交互关系. 基于时空特征和个体交互图显式地建模群体行为的语义关系, PRL采用关系门控层逐步细化语义关系图以提高群体行为识别的精度.

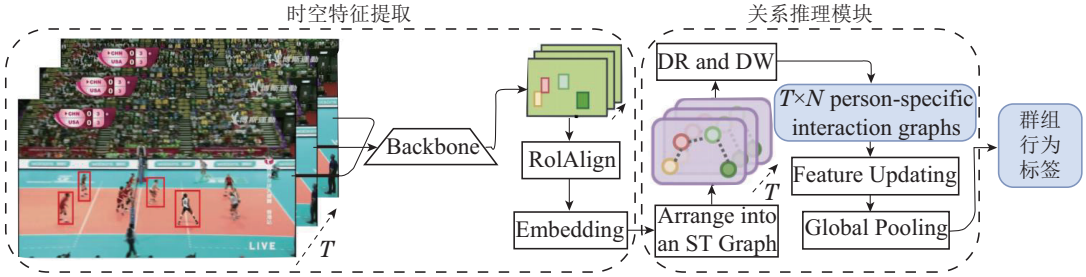


图 12　时空动态推理网络[61]

Fig. 12　The spatio-temporal dynamic inference network[61]

3.4　多模态融合策略

多模态学习已逐渐发展为视频内容分析与理解的主要手段, 通过将多种不同模态的信息进行融合, 进而用于分类任务或回归任务. 在群体行为识别任务中, 应用多种模态的数据特征可以有效提升识别的精度和鲁棒性. 多模态融合可分为多模态特征融合和多模态输入融合, 具体如下:

1) 多模态特征融合: 在群体行为识别中, 多模态特征融合是指从不同模态中提取与行为相关的特征(RGB帧、光流、骨架、位姿等)并进行融合. Azar等人[65]提出一个基于多流CNN的群体行为识别框架, 设计多流分支并采用不同模态(RGB帧、光流、位姿、扭曲光流)进行训练, 在融合层进行行为预测. Rossi等人[66]提出多模态深度信念网络的群体行为识别解决方案, 利用深度信念网络观察群体行为是否存在共同的时间/空间动态.

针对上述方法严重依赖行为者的外观特征, 忽略可用的语境信息等问题. Dasgupta等人[67]提出一种双流上下文感知结构(context aware GAR, CA-GAR), 该结构不仅考虑个体层面的外观特征, 同时利用上下文信息进行群体行为识别. 作者设计姿势上下文和场景上下文两种互补模块, 结合外观特征, 丰富个体行为者的编码方式. Zalluhoglu等人[68]认为在群体行为识别任务中, 研究者除了分析整个输入图像, 还须考虑多子区域特征, 首次使用多流卷积神经网络和多子区

域实现群体行为识别任务, 该结构使用单个RGB帧、光流信息及子区域特征, 扩展文献[69]提出的双流架构, 采用4个CNN架构捕获更细粒度的时空信息.

Liu等人[70]针对多模态特征融合无法捕获不同模态之间潜在交互的问题, 提出多模态语义上下文感知图神经网络(multimodal-semantic context aware graph neural network, MSCA-GNN). 该网络在多模态视觉图和语义图上进行表征学习, 以获取行为者的上下文感知特征, 并发现其潜在的交互模式. 同时设计双向映射学习机制, 利用多模态视觉图和语义图中可学习的连接权重进行融合, 以便通过语义上下文细化每个模态的视觉表示. 并在此框架上进行扩展, 设计了基于姿势–位置注意机制学习的视觉语义图神经网络, 采用基于注意力的聚合层将姿势信息和位姿信息融合用于群体行为识别[71].

2) 多模态输入融合: 多模态输入主要包括视觉模态和非视觉模态, 在群体行为识别任务中常用的是视觉模态(RGB帧), 而研究非视觉模态特征的文章较少, 这是未来的一个热点研究方向. 如图13所示, Das等人[72]利用骨架、RGB、加速度计和陀螺仪4种输入数据, 建立一个全新的基于深度学习的多模态人体行为识别集成网络(multi-modal human activity recognition, MMHAR). 将加速度和陀螺仪数据转换为相应的信号图, 进而融合多种模态数据输入CNN模型, 聚合到集成模型进行人体行为识别. 文献[73]利用RGB图像和人体骨架作为包含互补信息的输入, 设计双分支的时空网络进行群体行为识别.

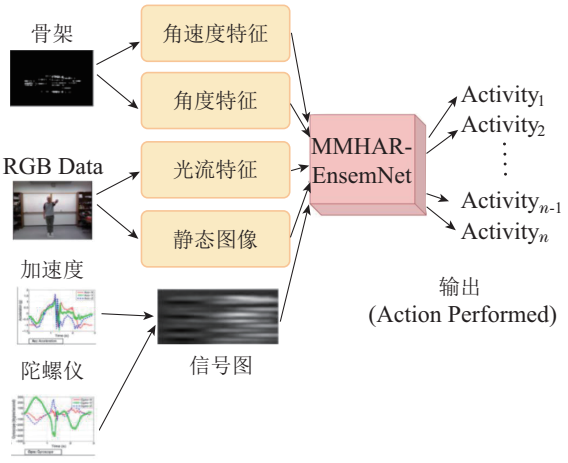


图 13 多模态人体行为识别集成网络[72]

Fig. 13 Multi-modal human activity recognition ensemble network

针对运动场景中不可避免的遮挡、摄像机角度变化和人体姿势变化等挑战性问题, Zhou等人[74]提出深度多模态输入融合算法, 通过将视觉特征、骨架姿势、概率图和音频信号融合为高级特征用于人体行为识别. 研究者将语义信息[25, 38]、骨架姿势信息[12]与视觉特征进行融合用于提高群体行为识别的精度和鲁棒性. 从使用骨架信息和基于关系网络的方法进行交互行为识别中受到启发, Perez等人[75]针对以往方法大多依赖于标准的双流方法(RGB和光流)作为输入特征, 提出一种群组交互关系网络(group interaction relational network, GIRN), 直接在关节空间坐标上生成特征表示, 从而推断不同个体之间的交互. GIRN还嵌入注意力机制, 使推理模块能抓住群体场景中的关键行为者.

3.5 群体行为识别的其他方法

这部分介绍的群体行为识别其他方法不属于上述的4种方法, 包括弱监督方法、半监督方法, 以及仅使用骨架数据的方法等. 这些均为近年提出的新方法, 处于发展上升的阶段.

Yan等人[76]提出一种弱监督群体行为识别的解决方案, 仅使用视频级别的标签, 这种设置不仅适用于真实场景, 而且为新基准的标注提供一种更简单、成本更低的方法. 如图14所示, 为挖掘有效的行为者交互信息, 构建关键实例进行交互关联, 进而设计社交自适应模块(social adaptive module, SAM), 从噪声数据中推理出关键人物和关键帧, 用于群体行为识别. 针对以往方法将行为者检测作为群体行为识别任务之前的一个独立步骤, 容易忽略两者任务之间的相互关系, Zhang等人[77]提出一种用于群体行为识别的弱监督深度学习体系, 通过行为者检测器和群体行为分类器在训练阶段相互完善和逐步加强. 仅需要个体行为者边界框和群体行为标签实现群体行为识别, 使得群体行为识别任务比以往的解决方法更适用于实际场景.

Kim等人[78]提出一种弱监督的群体行为识别模型(detector-free weakly supervised, DFWS), 既不依赖于边界框标签, 也不依赖于目标检测. 利用Transformer的自注意力机制来定位和编码群体活动的组件上下文进行群体行为识别. Bian等人[79]首次探索自监督表征学习下基于骨架的群体行为识别方法, 对不同级别行为者间的群体交互进行建模.

Zappardino等人[80]在不使用个体行为标签的情况下, 提出仅使用骨架数据的端到端半监督群体行为识别方法, 通过从预训练的特征提取器中计算伪标签以达到具有竞争力的识别性能. 为提高模型的泛化能力, Thilakarathne等人[13]首次提出仅使用骨架姿势作为输入进行群体行为识别(pose only group activity recognition system, POGARS)的框架. 文献[81]提出新颖的Zoom Transformer模型实现基于骨架的群体行为识别, 这项工作将会促进基于骨架的群体行为识别的研究. 基于骨架数据的群体行为识别方法不容易产生场景偏差, 便于模型的迁移和泛化, 后续研究骨架数据和视觉数据的多模态融合将会是一个热点方向.

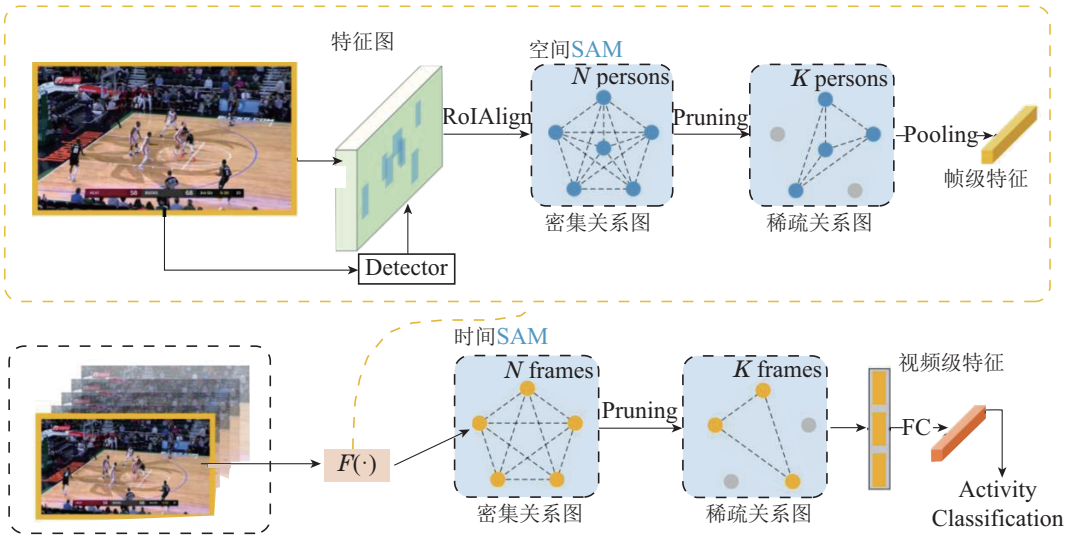


图 14　基于社交自适应模块的弱监督群体行为识别[76]

Fig. 14　Social adaptive module for weakly-supervised group activity recognition

3.6　讨论与分析

图15展示了群体行为识别领域中最具代表性工作的时间线, 标注了2016年到2022年间基于深度学习的群体行为识别方法的发展过程.

通过上述对基于深度学习的群体行为识别方法的介绍发现, 群体行为识别算法大多涉及个体行为者的检测、个体行为识别, 以及群体行为识别的实现, 群体行为识别相关综述文献如表1所示. 基于层级时序建模的群体行为识别方法使用两阶段LSTM模型学习个体级别动作的时序表示, 并通过池化函数将个体特征生成群组级别的表示. 两阶段方法启发了大量后续工作, 但它的局限性在于平等对待所有行为者. 通常在体育运动分析等场景中, 群体行为类别往往是由几个关键人物定义的, 通过注意力建模重点关注少数关键参与者, 提取对群体行为识别起作用的关键人和关键帧. 基于Transformer的方法可进一步捕获行为者之间高贡献度的时空信息, 这种方法的性能相比较于两阶段的层级时序建模有进一步的提升, 但是建立鲁棒的关键行为者模型仍是一个需要解决的难题. 在多人场景中, 行为者之间的交互关系建模对于识别群体行为至关重要. 基于交互关系建模与推理的群体行为识别方法, 其核心思想在于自适应地联合推断群体活动中不同组成部分之间丰富的互动关系, 从而推断出个体行为和群体行为. 基于多模态融合策略进行群体行为识别, 这可以提高群体行为识别算法的精度和鲁棒性, 但难以实际应用. 基于弱监督方法的多模态群体行为识别或许是未来一个值得研究的方向.

表 1　群体行为识别相关综述文献

Table 1　Related surveys about group activity recognition

方法类别	细分类方法	代表性论文	讨论与分析
层级时序建模	LSTM的两阶段方法 交互上下文	HDTM[22], CERN[23], SBGAR[25], GLIL[27] RMIC[29], DGCF[30], SSU[32], MLS-GAN[34]	局限性在于平等对待 所有行为者
注意力建模	提取关键行为者 多层注意力机制 基于Transformer的 自注意力机制	PCTDM[36], P^{2}CTDM[[37], SPTS[38], STCC-GCC[40] HAN-HCN[42], GAIM[44] AT-GAR[45], GroupFormer[48], LVC-GAR[51], SACRF[52], Composer[12], Dual-AI[50]	联合时空建模, 构建更为鲁棒的 行为者注意力模型
交互关建模 与推理	分层图模型 图卷积神经网络	HRN[54], HiGCIN[55], SIM[35], CRM[56], stagNet[58] ARG[60], DIN[61], GINs[63]	自适应推理行为者的交互和 时空演变关系
多模态融合	多模态特征融合 多模态输入融合	CA-GAR[67], MSCA-GNN[70] MMHAR[72], GIRN[75]	多模态鲁棒性强, 但难以实际应用
其他方法	半监督方法 弱监督方法 仅基于骨架方法等	Zappardino et al[80] SAM[76], DFKS[78], Zhang et al[77] POGARS[13], Zoom Transformer[81]	适用于实际场景, 但均为近年提出的新方法, 处于发展上升的阶段

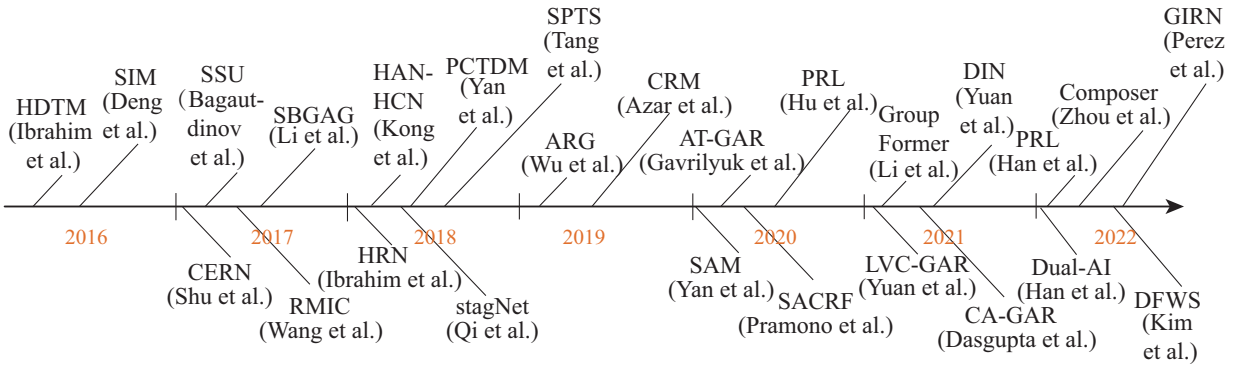


图 15 群体行为识别方法发展的时间线

Fig.15 Timeline of the development in group activity recognition methods

4 群体行为识别数据集及评估指标

此部分对常用的群体行为识别公共数据集进行介绍, 方便群体行为识别相关研究学者更好地开展工作. 如表2所示, 在群体行为识别研究中常用的数据集主要分为两大类, 一类是面向智能监控系统的公共监控视频数据集; 另一类是面向体育运动分析的大型体育视频数据集, 其主要动机是构建智能化、自动化的安防系统和体育分析系统.

表 2 群体行为识别数据集

Table 2 The datasets of group activity recognition

数据集名称	视频数量	群体行为类别数	个体行为类别数	年份	所属类型
CAD[82]	44	5	6	2009	监控视频数据集
CAED[83]	75	6	8	2011	监控视频数据集
NCAD[84]	32	6	3	2012	监控视频数据集
UCLA[85]	1	6	10	2012	监控视频数据集
NHD[18]	1	2	6	2012	监控视频数据集
VD[22]	55	8	9	2016	体育运动数据集
NBA[76]	181	9	N/A	2020	体育运动数据集
NCAA[86]	257	11	N/A	2016	体育运动数据集
BFHD[87]	58	3	11	2012	体育运动数据集
HARD[88]	12	4	N/A	2020	体育运动数据集
Soccer[89]	74	4	N/A	2020	体育运动数据集

4.1 监控视频数据集

所有的监控视频数据都是在实际的场景中(街道、校园、医院餐厅等)进行拍摄, 从不同动态视角进行采集, 以增强数据集的多样性和挑战性. 下面对群体行为识别任务中常用的监控数据集进行简单介绍:

CAD (collective activity dataset)[82]: CAD 数据集[1]是应用最为广泛的群体行为数据集, 该数据库由一台低分辨率手持式摄像机在动态视角下拍摄视频组成, 共有44个视频. 该数据集包含5种不同的群体行为标签 (crossing, waiting, queueing, walking, talking)、6不同的个体行为标签(NA, crossing, waiting, queueing, walking, talking)和8种不同的个体级别的姿势标签 (right, front-right, front, front-left, left, back-left, back, back-right), 整个数据集的视频序列每隔10帧对个体的检测边界框、个体行为标签和个体姿势标签进行手动注释.

CAED (collective activity extended dataset)[83]: CAED数据集共有75个视频, 在CAD数据集的基础上增加两种新的行为类别(dancing, jogging), 并移除定义不清、存在混淆的行为类别Walking. 原因在于Walking更倾向于个人行为, 而不是群体行为.

NCAD (new collective activity dataset)[84]: NCAD数据集共有32个视频和6种不同的群体行为标签 (gathering, talking, dismissal, walking together, chasing, queueing)和3种不同的个体行为标签(running, walking, standing still), 定义8种不同个体级别的姿势标签, 同CAD数据集一致.

UCLA courtyard dataset[85]: UCLA数据集包含一段时长106分钟、帧率30 fps、分辨率2560*1920的视频, 采用鸟瞰视角进行拍摄采集视频. 该数据集包含6

[1] https://cvgl.stanford.edu/projects/collective/collectiveActivity.html.

种不同的群体行为类别(walking-together, standing-in-line, discussing-in-group, sitting-together, waiting-in-group, guided-tour)和10种不同的个体行为类别(riding-skateboard, riding-bike, riding-scooter, driving-car, walking, talking, waiting, reading, eating, sitting), 对每帧的群体行为、个体行为和个体的检测边界框进行手动注释.

NHD (nursing home dataset)[18]: NHD数据集由低分辨率鱼眼摄像头在疗养院的餐厅中采集视频组成, 该数据集包含6种不同的个体行为标88签(walking, standing, sitting, bending, falling)和2种不同场景级的群体行为标签(fall, non-fall), 共有2990个带注释的标签帧, 其中大约1/3是fall的标签帧.

4.2 体育视频数据集

与监控视频数据集相比, 体育数据集具有更严重的遮挡和复杂的交互. 由于体育运动分析的迫切需求, 国内外许多研究机构正在致力于体育运动(足球, 排球, 篮球, 曲棍球等)中的群体行为识别研究, 下面对群体行为识别任务中常用的体育运动数据集进行简单介绍:

VD (volleyball dataset)[22]: VD数据集[2]是应用最为广泛的体育运动分析数据集, 因其规模大、个体行为者间交互复杂、移动速度快. 该数据集由55场排球比赛视频中截取4830个视频组成, 每个视频中间帧标注了个体检测边界框、个体行为标签以及群体行为标签, 其包含9种不同的个体行为标签(waiting, setting, digging, falling, spiking, blocking, jumping, moving, standing)和8种不同的群体行为标签(right winpoint, right passing, right spiking, right setting, left setting, left spiking, left passing, left winpoint). 为了实验比较的公平性, 研究者们[45, 48, 60–61]都选择将数据集中的2/3作为训练集, 剩下的1/3作为测试集.

NBA dataset[76]: NBA数据集是目前最大、最具挑战性的群体行为识别基准数据集. 该数据集总共由181个NBA比赛视频中截取9172个视频片段组成, 视频帧分辨率为1920*1080, 采样帧率为12 fps, 其包含9种不同的群体行为标签, 但没有对个体行为进行注释.

NCAA basketball dataset[86]: NCAA数据集从YouTube视频网站中提供的296个NCAA比赛视频选取257场篮球比赛视频, 该数据集视频时长为1.5 h, 包含总共11种不同的群体行为识别标签, 考虑5种投篮类别, 每种类别都有成功和失败两种类型. 这些视频数据被随机分成212个训练视频、12个验证视频和33个测试视频.

BFHD (broadcast field hockey dataset)[87]: BFHD数据集从5个比赛视频中选取58个视频序列, 其中包含11种不同的个体行为类别(pass, dribble, shot, receive, tackle, prepare, stand, jog, run, walk, save)、5种曲棍球比赛行为类别(attacker, first defenders, defenders defend against person, defenders defend against space, other)和3种不同的场景级的群体类别(attack play, free hit, penalty corner).

HARD (hockey activity recognition dataset)[88]: HARD数据集是从国际曲棍球联合会(FIH)和YouTube(2018年曲棍球世界杯)视频网站中进行收集, 总共包括12场转播曲棍球比赛, 视频帧率为25 fps, 分辨率为1280*720, 其包含4种不同的群体行为标签(free hit, goal, long corner, penalty corner). 该数据集总共收集400个关键曲棍球活动帧, 每个类包含100 帧, 以确保数据集均匀分布.

Soccer Dataset[89]: 该数据集是Sportlogiq公司提供2018–2019年英超联赛的74场足球比赛视频, 帧率为30 fps的转播比赛视频. 使用多摄像头以及视觉跟踪系统获取足球和所有球员的轨迹, 标注每帧的群体行为标签(background, pass, reception, shot), 并在数据集上提供一种基于轨迹和视觉特征的群体行为检测方法框架.

4.3 相关的开源代码

表3列举了群体行为识别相关开源代码库以及对应算法在VD和CAD数据集上的性能. 通过提供相关经典论文的开源代码, 以便复现或更好地理解群体行为识别的内在机制, 促进群体行为识别领域的快速发展. 其中CAD数据集中“4 classes”与“5 classes”的区别是将“Walking”和“Crossing”合并成“Moving”, 由原来的5种类别减少为4种. 从2016–2022年间所召开的顶级会议中群体行为识别算法发展历程可知, 目前VD数据集的最佳识别精度为95.7%, CAD数据集的最佳识别精度为96.5%, 该最佳方法主要通过强有力的I3D骨干网络提取时空特征, 且采用RGB、光流、位姿3种模态特征作为输入.

4.4 群体行为识别数据集的评估指标

为进一步评估量化算法的有效性, 引入特定的客观评价指标对算法模型在数据集的表现进行分析. 常见的3种群体行为识别性能评估指标为: 1)多类准确度(multi-class accuracy, MCA): 正确预测群体行为类别的百分比; 2)平均每类准确度(mean per-class accuracy, MPCA): 由于每个类别的类别数不平衡, 计算最终所有类别平均的准确度; 3)混淆矩阵(confusion matrix): 评测模型性能可视化的特定矩阵, 其矩阵的每一行和每一列分别代表真值类和预测类的标签.

[2] https://github.com/mostafa-saad/deep-activity-rec.

表 3 群体行为识别相关论文的开源代码以及识别精度

Table 3 Open source code for papers and the accuracy in group activity recognition

方法	发表于	输入	骨干网络	是否开源	MCA/%		MPCA/%
					VD	CAD(5 classes)	CAD(4 classes)
HDTM[22]	CVPR 2016	RGB	AlexNet	是	81.9	81.5	89.6
SSU[32]	CVPR 2017	RGB	Inception-v3	是	89.9	N/A	N/A
SBGAR[25]	CVPR 2017	RGB, Flow	Inception-v3	是	66.9	86.1	89.9
PCTDM[36]	ACM 2018	RGB, Flow	AlexNet	是	87.7	N/A	92.2
HRN[54]	ECCV 2018	RGB	VGG19	是	89.5	N/A	N/A
ARG[60]	CVPR 2019	RGB	Inception-v3	是	92.6	91.0	93.1
HiGCIN[55]	TPAMI 2020	RGB	ResNet-18	是	91.5	N/A	93.4
Skeletons[80]	ICPR 2020	RGB	P3D	是	91.0	N/A	N/A
DIN[61]	ICCV 2021	RGB	VGG-16	是	93.6	N/A	95.9
GroupFormer[48]	ICCV 2021	RGB, Flow, Pose	I3D	是	95.7	N/A	96.3
DFWS[78]	CVPR 2022	RGB	ResNet-18	否	90.5	N/A	N/A
Composer[12]	ECCV 2022	Keypoint	HRNet	是	94.6	N/A	96.2
Dual-AI[50]	CVPR 2022	RGB, Flow	Inception-v3	否	95.4	N/A	96.5

5 未来展望

基于群体行为识别的发展现状、面临的挑战及存在的问题, 本文对群体行为识别领域未来的研究方向展望如下:

1) 构建更为高效、鲁棒的群体行为识别模型: 对于群体行为识别而言, 不仅仅需要对个体行为进行识别, 更需要通过个体行为者间的交互关系推断群体行为的标签. 对基于视频分析的行为识别问题, 2D CNN不能很好地捕获时序上的信息, 使用3D CNN(I3D[47]等)更适用于提取时空特征以及捕获连续帧间的时序运动信息. 时空图卷积神经网络和时空Transformer模型近几年在群体行为识别取得了优越的性能, 如何自适应地联合推断群体活动中丰富的时空联系和互动关系是需要进一步解决的问题. 仅基于视觉模态的群体行为识别往往容易存在场景偏差, 不便于模型泛化, 如何进行多模态融合, 构建更为高效、鲁棒的群体行为识别模型仍需进一步探索.

2) 小样本/弱监督学习: 群体行为识别需要依赖大量标注数据进行训练, 这是一项复杂且耗时的任务, 使得面向实际场景的群体行为识别方法的落地应用受到了限制. 但目前涉及小样本条件下的群体行为识别问题的成果几乎没有, 属于全新研究范畴. 因此, 如何从少量的样本中以及无标签的数据中快速学习, 并进一步延伸到新的未知行为类别中, 也是一个待研究的任务.

3) 轻量化、端到端的统一网络架构设计: 基于深度学习的群体行为识别方法由于强大的特征提取能力, 相较于传统手工特征方法具有更优秀的性能, 但代价是需要更多的计算量和存储量. 针对这一问题, 目前主要有两种解决方法: 1) 轻量化模型设计, 主要思想是设计一种简单有效的网络结构, 在不影响识别性能的情况下降低模型复杂度, 例如文献[90]替换骨干网络为MNASNet(mobile neural architecture search network)[91]和MoblieNet[92] 以减少模型的计算量. 2) 模型压缩, 通过模型剪枝和知识蒸馏等方法重新设计简单的网络结构, 从而提高模型训练速度和解决存储量过大问题. 为提高基于深度学习的群体行为识别算法的实时性和加强其落地应用, 设计轻量化、端到端的统一网络架构同样是一个值得研究的问题.

4) 大规模、更具挑战性的数据集: 目前虽然有大量的群体行为识别数据集, 但是这些数据集都是基于某类特定场景的, 并且群体行为类别和个体行为标签仅有几种. 例如最常用的排球数据集于2016年提出, 目前群体行为识别的精度已到95.7%[48]之高, 但仅限于排球比赛数据集中. 缺乏统一规范的行为类别定义以及详细标注的数据集, 限制了相关研究工作的开展, 因此, 创建大规模、更具挑战性的通用数据集是一个亟待解决的问题.

6 结束语

群体行为识别以视频中运动人群的行为分析和理解为研究目的, 旨在推断识别复杂场景中一群人进行的整体活动. 本文首先介绍群体行为的定义、通用识别流程及主要的挑战; 其次, 从群体行为识别的建模方法和内在机理进行划分, 并进一步细分类、讨论和分析这些方法的优缺点; 然后, 给出群体行为识别的常用数据集, 列举了相关的开源代码论文和评估指标; 最后, 在对现有技术方法分析和讨论的基础上, 对群体行为识别领域未来的研究方向进行了展望.

参考文献:

[1] HAN R, YAN H, LI J, et al. Panoramic human activity recognition. *ArXiv Preprint,* 2022, arXiv: 2203.03806.

[2] ZHU Y, LI X, LIU C, et al. A comprehensive study of deep video action recognition. *ArXiv Preprint,* 2020, arXiv: 2012.06567.

[3] PAREEK P, THAKKAR A. A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 2021, 54(3): 2259 – 2322.

[4] AHMAD T, JIN L, ZHANG X, et al. Graph convolutional neural network for human action recognition: A comprehensive survey. *IEEE Transactions on Artificial Intelligence*, 2021, 2(2): 128 – 145.

[5] NAIR S A L, MEGALINGAM R K. Human action recognition: A review. *International Conference on System Modeling Advancement in Research Trends*. Moradabad, India: IEEE, 2021: 249 – 252.

[6] VAHORA S A, CHAUHAN N C. A comprehensive study of group activity recognition methods in video. *Indian Journal of Science and Technology*, 2017, 10(23): 1 – 11.

[7] PEI Lishen, ZHAO Xuezhuan. A survey of collective activity analysis and recognition based on deep learning. *Journal of Frontiers of Computer Science and Technology*, 2022, 16(4): 775 – 790.
(裴利沈, 赵雪专. 群体行为识别深度学习方法研究综述. 计算机科学与探索, 2022, 16(4): 775 – 790.)

[8] WU L F, WANG Q, JIAN M, et al. A comprehensive review of group activity recognition in videos. *International Journal of Automation and Computing*, 2021, 18(3): 334 – 350.

[9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory. *Neural Computation*, 1997, 9(8): 1735 – 1780.

[10] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks. *ArXiv Preprint.* 2017, arXiv:1609.02907.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017, 30.

[12] ZHOU H, KADAV A, SHAMSIAN A, et al. COMPOSER: Compositional reasoning of group activity in videos with keypoint-only modality. *ArXiv Preprint,* 2022, arXiv: 2112.05892.

[13] THILAKARATHNE H, NIBALI A, HE Z, et al. Pose is all you need: The pose only group activity recognition system (POGARS). *ArXiv Preprint,* 2021, arXiv: 2108.04186.

[14] CHAN T H, JIA K, GAO S, et al. PCANet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, 2015, 24(12): 5017 – 5032.

[15] ZHAO Z Q, ZHENG P, XU S T, et al. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 30(11): 3212 – 3232.

[16] MINAEE S, BOYKOV Y Y, PORIKLI F, et al. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 44(7): 3523 – 3542.

[17] CIAPARRONE G, LUQUE SÁNCHEZ F, TABIK S, et al. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 2020, 381: 61 – 88.

[18] LAN T, WANG Y, YANG W, et al. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(8): 1549 – 1562.

[19] SHU T, XIE D, ROTHROCK B, et al. Joint inference of groups, events and human roles in aerial videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Boston, MA, USA: IEEE, 2015: 4576 – 4584.

[20] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent neural network based language model. *Interspeech*, 2010, 2(3): 1045 – 1048.

[21] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, 86(11): 2278 – 2324.

[22] IBRAHIM M S, MURALIDHARAN S, DENG Z, et al. A hierarchical deep temporal model for group activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016: 1971 – 1980.

[23] SHU T, TODOROVIC S, ZHU S C. CERN: Confidence-energy recurrent network for group activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017: 5523 – 5531.

[24] SHU X, TANG J, QI G J, et al. Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(3): 1110 – 1118.

[25] LI X, CHOO CHUAH M. SBGAR: Semantics based group activity recognition. *Proceedings of the IEEE International Conference on Computer Vision*. Venice, Italy: IEEE, 2017: 2876 – 2885.

[26] Al-HABIB M, HUANG D, Al-QATF M, et al. Cooperative hierarchical framework for group activity recognition: From group detection to multi-activity recognition. *Proceedings of the 8th International Conference on Software and Computer Applications*. Penang, Malaysia: ACM, 2019: 291 – 298.

[27] SHU X, ZHANG L, SUN Y, et al. Host‐parasite: Graph LSTM-in-LSTM for group activity recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, 32(2): 663 – 674.

[28] LI Ding, ZHANG Wensheng. Attentive pooling for group activity recognition. *Scientia Sinica Informationis*, 2021, 51(3): 399 – 412.
(李定, 张文生. 面向群体行为识别的注意力池化机制. 中国科学:信息科学, 2021, 51(3): 399 – 412.)

[29] WANG M, NI B, YANG X. Recurrent modeling of interaction context for collective activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017: 3048 – 3056.

[30] KIM P S, LEE D G, LEE S W. Discriminative context learning with gated recurrent unit for group activity recognition. *Pattern Recognition*, 2018, 76: 149 – 161.

[31] ZHUANG N, YUSUFU T, YE J, et al. Group activity recognition with differential recurrent convolutional neural networks. *International Conference on Automatic Face Gesture Recognition*. Washington, DC, USA: IEEE, 2017: 526 – 531.

[32] BAGAUTDINOV T, ALAHI A, FLEURET F, et al. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017: 4315 – 4324.

[33] GOODFELLOW I. NIPS 2016 tutorial: Generative adversarial networks. *ArXiv Preprint,* 2017, arXiv: 1701.00160.

[34] GAMMULLE H, DENMAN S, SRIDHARAN S, et al. Multi-level sequence GAN for group activity recognition. *Asian Conference on Computer Vision*. Perth, Australia: Springer, 2019: 331 – 346.

[35] DENG Z, VAHDAT A, HU H, et al. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016: 4772 – 4781.

[36] YAN R, TANG J, SHU X, et al. Participation-contributed temporal dynamic model for group activity recognition. *Proceedings of the 26th ACM International Conference on Multimedia*. Seoul, South Korea: ACM, 2018: 1292 – 1300.

[37] YAN R, SHU X, YUAN C, et al. Position-aware participation-contributed temporal dynamic model for group activity recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, 33(12): 7574 – 7588.

[38] TANG Y, WANG Z, LI P, et al. Mining semantics-preserving attention for group activity recognition. *Proceedings of the 26th ACM International Conference on Multimedia*. Seoul, South Korea: ACM, 2018: 1283 – 1291.

[39] TANG Y, LU J, WANG Z, et al. Learning semantics-preserving attention and contextual interaction for group activity recognition. *IEEE Transactions on Image Processing*, 2019, 28(10): 4997 – 5012.

[40] TANG J, SHU X, YAN R, et al. Coherence constrained graph LSTM for group activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(2): 636 – 647.

[41] LI Y, LIU Y, YU R, et al. Dual attention based spatial-temporal inference network for volleyball group activity recognition. *Multimedia Tools and Applications*, 2022, 82(10): 15515 – 15533.

[42] KONG L, QIN J, HUANG D, et al. Hierarchical attention and context modeling for group activity recognition. *International Conference on Acoustics, Speech and Signal Processing*. Calgary, AB, Canada: IEEE, 2018: 1328 – 1332.

[43] LU L, DI H, LU Y, et al. A two-level attention-based interaction model for multi-person activity recognition. *Neurocomputing*, 2018, 322: 195 – 205.

[44] LU L, LU Y, YU R, et al. GAIM: Graph attention interaction model for collective activity recognition. *IEEE Transactions on Multimedia*, 2020, 22(2): 524 – 539.

[45] GAVRILYUK K, SANFORD R, JAVAN M, et al. Actor-transformers for group activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA, USA: IEEE, 2020: 839 – 848.

[46] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA: IEEE, 2019: 5693 – 5703.

[47] CARREIRA J, ZISSERRMAN A. Quo Vadis, action recognition? A new model and the kinetics dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017: 6299 – 6308.

[48] LI S, CAO Q, LIU L, et al. Groupformer: Group activity recognition with clustered spatial-temporal transformer. *Proceedings of the IEEE International Conference on Computer Vision*. Montreal, QC, Canada: IEEE, 2021: 13668 – 13677.

[49] LI W, YANG T, WU X, et al. Learning action-guided spatio-temporal transformer for group activity recognition. *Proceedings of the 30th ACM International Conference on Multimedia*. Lisboa, Portugal: ACM, 2022: 2051 – 2060.

[50] HAN M, ZHANG D J, WANG Y, et al. Dual-AI: Dual-path actor interaction learning for group activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA: IEEE, 2022: 2990 – 2999.

[51] YUAN H, NI D. Learning visual context for group activity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada: AAAI Press, 2021, 35(4): 3261 – 3269.

[52] PRAMONO R R A, CHEN Y T, FANG W H. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. *Proceedings of the European Conference on Computer Vision*. Glasgow, UK: Springer, 2020: 71 – 90.

[53] PRAMONO R R A, FANG W H, CHEN Y T. Relational reasoning for group activity recognition via self-attention augmented conditional random field. *IEEE Transactions on Image Processing*, 2021, 30: 8184 – 8199.

[54] IBRAHIM M S, MORI G. Hierarchical relational networks for group activity recognition and retrieval. *Proceedings of the European Conference on Computer Vision*. Munich, Germany: Springer, 2018: 721 – 736.

[55] YAN R, XIE L, TANG J, et al. HiGCIN: Hierarchical graph-based cross inference network for group activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 45(6): 6955 – 6968.

[56] AZAR S M, ATIGH M G, NICKABADI A, et al. Convolutional relational machine for group activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA: IEEE, 2019: 7892 – 7901.

[57] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016: 4724 – 4732.

[58] QI M, QIN J, LI A, et al. stagNet: An attentive semantic RNN for group activity recognition. *Proceedings of the European Conference on Computer Vision*. Munich, Germany: Springer, 2018: 101 – 117.

[59] QI M, WANG Y, QIN J, et al. StagNet: An attentive semantic RNN for group activity and individual action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, 30(2): 549 – 565.

[60] WU J, WANG L, WANG L, et al. Learning actor relation graphs for group activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA: IEEE, 2019: 9964 – 9974.

[61] YUAN H, NI D, WANG M. Spatio-temporal dynamic inference network for group activity recognition. *Proceedings of the IEEE International Conference on Computer Vision*. Montreal, QC, Canada: IEEE, 2021: 7476 – 7485.

[62] LI W, XIE W, TU Z, et al. Multi-hyperedge hypergraph for group activity recognition. *International Joint Conference on Neural Networks (IJCNN)*. Padua, Italy: IEEE, 2022: 1 – 7.

[63] TANG Y, WEI Y, YU X, et al. Graph interaction networks for relation transfer in human activity videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, 30(9): 2872 – 2886.

[64] HU G, CUI B, HE Y, et al. Progressive relation learning for group activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA, USA: IEEE, 2020: 980 – 989.

[65] AZAR S M, ATIGH M G, NICKABADI A. A multi-stream convolutional neural network framework for group activity recognition. *ArXiv Preprint,* 2018, arXiv: 1812.10328.

[66] ROSSI S, CAPASSO R, ACAMPORA G, et al. A multimodal deep learning network for group activity recognition. *International Joint Conference on Neural Networks*. Rio de Janeiro, Brazil: IEEE, 2018: 1 – 6.

[67] DASGUPTA A, JAWAHAR C V, ALAHARI K. Context aware group activity recognition. *International Conference on Pattern Recognition*. Milan, Italy: IEEE, 2021: 10098 – 10105.

[68] ZALLUHOGLU C, IKIZLER-CINBIS N. Region based multistream convolutional neural networks for collective activity recognition. *Journal of Visual Communication and Image Representation*, 2019, 60: 170 – 179.

[69] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*. Montreal, Quebec, Canada: MIT Press, 2014.

[70] LIU T, ZHAO R, LAM K M. Multimodal-semantic context-aware graph neural network for group activity recognition. *International Conference on Multimedia and Expo*. Shenzhen, China: IEEE, 2021: 1 – 6.

[71] LIU T, ZHAO R, LAM K M, et al. Visual-semantic graph neural network with pose-position attentive learning for group activity recognition. *Neurocomputing*, 2022, 491: 217 – 231.

[72] DAS A, SIL P, SINGH P K, et al. MMHAR-EnsemNet: A multi-modal human activity recognition model. *IEEE Sensors Journal*, 2021, 21(10): 11569 – 11576.

[73] ZHAI X, HU Z, YANG D, et al. Spatial temporal network for image and skeleton based group activity recognition. *Proceedings of the Asian Conference on Computer Vision*. Macao, China: Springer, 2022: 20 – 38.

[74] ZHOU E, ZHANG H. Human action recognition toward massive-scale sport sceneries based on deep multi-model feature fusion. *Signal Processing: Image Communication*, 2020, 84: 115802.

[75] PEREZ M, LIU J, KOT A C. Skeleton-based relational reasoning for group activity analysis. *Pattern Recognition*, 2022, 122: 108360.

[76] YAN R, XIE L, TANG J, et al. Social adaptive module for? Weakly-supervised group activity recognition. *Proceedings of the European Conference on Computer Vision*. Glasgow, UK: Springer, 2020: 208 – 224.

[77] ZHANG P, TANG Y, HU J F, et al. Fast collective activity recognition under weak supervision. *IEEE Transactions on Image Processing*, 2020, 29: 29 – 43.

[78] KIM D, LEE J, CHO M, et al. Detector-free weakly supervised group activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA: IEEE, 2022: 20083 – 20093.

[79] BIAN C, FENG W, WANG S. Self-supervised representation learning for skeleton-based group activity recognition. *Proceedings of the 30th ACM International Conference on Multimedia*. Lisboa, Portugal: ACM, 2022: 5990 – 5998.

[80] ZAPPARDINO F, URICCHIO T, SEIDENARI L, et al. Learning group activities from skeletons without individual action labels. *International Conference on Pattern Recognition*. Milan, Italy: IEEE, 2021: 10412 – 10417.

[81] ZHANG J, JIA Y, XIE W, et al. Zoom transformer for skeleton-based group activity recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, 32(12): 8646 – 8659.

[82] CHOI W, SHAHID K, SAVARESE S. What are they doing? Collective activity classification using spatio-temporal relationship among people. *Proceedings of the International Conference on Computer Vision Workshops*. Kyoto, Japan: IEEE, 2009: 1282 – 1289.

[83] CHOI W, SHAHID K, SAVARESE S. Learning context for collective activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO, USA: IEEE, 2011: 3273 – 3280.

[84] CHOI W, SAVARESE S. A unified framework for multi-target tracking and collective activity recognition. *Proceedings of the European Conference on Computer Vision*. Florence, Italy: Springer, 2012: 215 – 230.

[85] AMER M R, XIE D, ZHAO M, et al. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. *Proceedings of the European Conference on Computer Vision*. Florence, Italy: Springer, 2012: 187 – 200.

[86] RAMANATHAN V, HUANG J, ABU-EL-HAIJA S, et al. Detecting events and key actors in multi-person videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, 2016: 3043 – 3053.

[87] LAN T, SIGAL L, MORI G. Social roles in hierarchical models for human activity recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Providence, RI, USA: IEEE, 2012: 1354 – 1361.

[88] RANGASAMY K, AS'ARI M A, RAHMAD N A, et al. Hockey activity recognition using pre-trained deep learning model. *ICT Express*, 2020, 6(3): 170 – 174.

[89] SANFORD R, GORJI S, HAFEMANN L G, et al. Group activity detection from trajectory and video data in soccer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. Seattle, WA, USA: IEEE, 2020: 898 – 899.

[90] KUANG Z, TIE X. Improved actor relation graph based group activity recognition. *ArXiv Preprint,* 2020, arXiv: 2010.12968.

[91] TAN M, CHEN B, PANG R, et al. Mnasnet: Platform-aware neural architecture search for mobile. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA: IEEE, 2019: 2820 – 2828.

[92] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv Preprint,* 2017, arXiv: 1704.04861.

作者简介:

朱晓林 博士研究生, 目前研究方向为群体行为识别、视频理解与分析, E-mail: xiaolin.zhu@smail.xtu.edu.cn;

王冬丽 博士, 副教授, 目前研究方向为行为识别、模式识别与分布式学习, E-mail: wangdl@xtu.edu.cn;

欧阳万里 博士, 高级讲师, 目前研究方向为计算机视觉和模式识别, E-mail: wanli.ouyang@sydney.edu.au;

李抱朴 博士, 目前研究方向为计算机视觉和模式识别及其应用, E-mail: baopuli@baidu.com;

周 彦 博士, 教授, 目前研究方向为机器视觉与集群机器人、信号处理与信息融合, E-mail: yanzhou@xtu.edu.cn;

刘金富 硕士研究生, 目前研究方向为群体行为识别、视频理解与分析, E-mail: 202021002342@smail.xtu.edu.cn.