基于深度学习的群体动作识别综述

严锐[1),2)] 葛晓静[2)] 黄捧[2)] 舒祥波[2)] 唐金辉[2)]

[1)](南京大学 计算机科学与技术系 南京 210023)

[2)](南京理工大学 计算机科学与工程学院 南京 210094)

摘 要 不同于传统的简单动作识别，群体动作识别需要理解场景中由若干人物的单人动作和他们之间的交互动作构成的复杂语义。近年来，群体动作识别在公共安全监控、体育视频分析和社会角色理解等领域的研究与应用引起了学者们的广泛关注。但是现有能够帮助学者们快速了解研究概况的中文文献很少且用于归纳和分析的依据较为笼统。为此，本文旨在综述近十年来基于深度学习的群体动作识别的研究进展。首先，本文介绍了群体动作识别的问题与定义，总结了现有解决方案的核心流程和该研究的关键挑战。然后，本文针对现有研究中的两个核心内容，即个体动作特征的提取及其关联建模，对现有文献作出了归纳与分析。具体而言，本文介绍并总结了群体动作研究中常用的人体行为特征，并将现有关联建模类型归纳为三类，即线性关联、序列关联和图关联。此外，本文还列举了现有的十二种可用于群体动作研究的视频数据集，并在三个常用数据集上对目前流行的方法进行了对比与分析。最后，本文研判了几个更具挑战的未来研究趋势。综上，本文剖析了群体动作识别的核心研究思路及未来研究趋势，有助于相关研究人员快速了解群体动作识别的研究概况。

关键词 视频理解；动作识别；群体动作识别；深度学习；注意力机制；递归神经网络；图模型

中图法分类号 TP391

A Survey Of Group Activity Recognition Based On Deep Learning

YAN Rui[1),2)] Ge Xiao-Jing[2)] HUANG Peng[2)] SHU Xiang-Bo[2)] TANG Jin-Hui[2)]

[1)] (Department of Computer Science and Technology, Nanjing University, Nanjing 210023)

[2)] (Department of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094)

Abstract Different from traditional action recognition focused on single individuals, group activity recognition aims to understand the complex semantics composed of individual actions and their interactions within a scene. In recent years, the application of group activity recognition in various domains such as public safety monitoring, sports video analysis, and social role understanding has garnered significant attention from researchers. However, there is a scarcity of Chinese literature providing a comprehensive overview of the research progress in this field, and the foundational aspects for induction and analysis remain vague. This paper aims to fill this gap by offering a thorough review of the progress in group activity recognition research over the past decade, with a particular focus on developments facilitated by deep learning technologies. To begin, we establish a clear problem definition for group activity recognition, differentiating it from individual action recognition by highlighting the significance of understanding group dynamics and interactions. Following this, we outline the basic pipeline common to most group activity recognition approaches, which typically involves the detection and tracking of individuals, the

本课题得到国家资助博士后计划(No. GZB20230302)、江苏省卓越博士后计划(No. 2023ZB256)、国家自然科学基金(No. 62302208, No. 61925204, No. 62222207, No. 62072245)、江苏省自然科学基金(No. BK20211520)资助.**严锐**，博士，主要研究领域为计算机视觉、多媒体分析.**葛晓静**，硕士，主要研究领域为计算机视觉、机器学习. **黄捧**，博士，主要研究领域为计算机视觉、多媒体分析. **舒祥波**，博士，教授，主要研究领域为计算机视觉、机器学习. **唐金辉**（**通信作者**），博士，教授，主要研究领域为多媒体分析、计算机视觉.

extraction of features pertinent to their actions, the recognition of individual actions, and the aggregation of these actions to infer group activities. Concurrently, we discuss the challenges inherent to this research field, such as the variability in group sizes, the complexity of interactions, and the diversity of possible group activities across different contexts.

Delving deeper into the core aspects of group activity recognition research, this paper then provides an in-depth analysis of two critical components: the extraction of individual action features and their association modeling. We introduce several deep learning-based methods for extracting video features that are commonly employed in the study of group activities. These methods are adept at capturing the nuances of individual actions and the contextual information necessary for understanding group dynamics. Following this, we categorize existing approaches to modeling the associations between individual actions into three distinct types: linear association, sequence association, and graph association. Each type offers a unique perspective on how individual actions interact and combine to form coherent group activities, from simple linear relationships to complex, non-linear interactions represented by graphs. Furthermore, recognizing the importance of empirical research in advancing the field, this paper provides a comprehensive list of 12 existing video datasets specifically curated for group activity research. These datasets vary in terms of the scenarios they cover, from sports and public spaces to more controlled settings, thereby offering diverse opportunities for testing and improving group activity recognition algorithms. We also conduct a comparative analysis of existing methods using the two most popular datasets, highlighting their strengths and weaknesses and providing insights into their performance.

In conclusion, this paper offers a comprehensive review of the advancements in group activity recognition based on deep learning over the past decade. It covers the problem definition, research challenges, feature extraction techniques, association modeling methods, evaluation datasets, and future research directions. By consolidating and analyzing the existing knowledge, this review provides researchers with valuable insights and guidance for further exploration and development in the field of group activity recognition.

Key words video understanding; action recognition; group activity recognition; deep learning; attention mechanism; recurrent neural network; graph model

1 引言

随着信息存储技术和互联网基础设施的不断完善，视频数据的接收与传输更加方便快捷。再加上社交媒体的兴起，每天都有大量视频数据被上传到各大社交平台（例如 YouTube、抖音、微博等）。此外，世界各国近年来越来越重视智能监控建设，导致监控视频数据也呈现爆炸式增长。由此可见，未来智能算法研究的主要推动力和应用场景将不再以图像数据分析为主导，而是着眼于视频数据的处理与理解。视频分析技术必将迎来新的机遇与挑战。

基于视频数据的人体动作识别[1][2][3][4][5][6]技术是视频分析领域中一项基础且关键的研究内容。通常情况下，根据参与动作的人物数量将人体动作分为单人动作（例如图 1-（a）所示）[7][8][9][10][11][12][13]、交互动作（例如图 1-（b）所示）[14]和群体动作（三人及以上，如图 1-(c)所示）[15][16][17][18][19][20][21][22][23]。长期以来，学者们主要聚焦于对单人动作和交互动作 [24][25][26]的研究，而忽略了真实应用场景中人物数量通常众多且变化的这一特点。近年来，比较贴近实际需求的群体动作识别[16][27]开始受到学者们的广泛关注。该任务旨在识别一群人共同完成的动作，例如图 1-（c）中所示排球比赛中“左侧二传”。特别地，群体动作中每个个体具有不同的行为状态

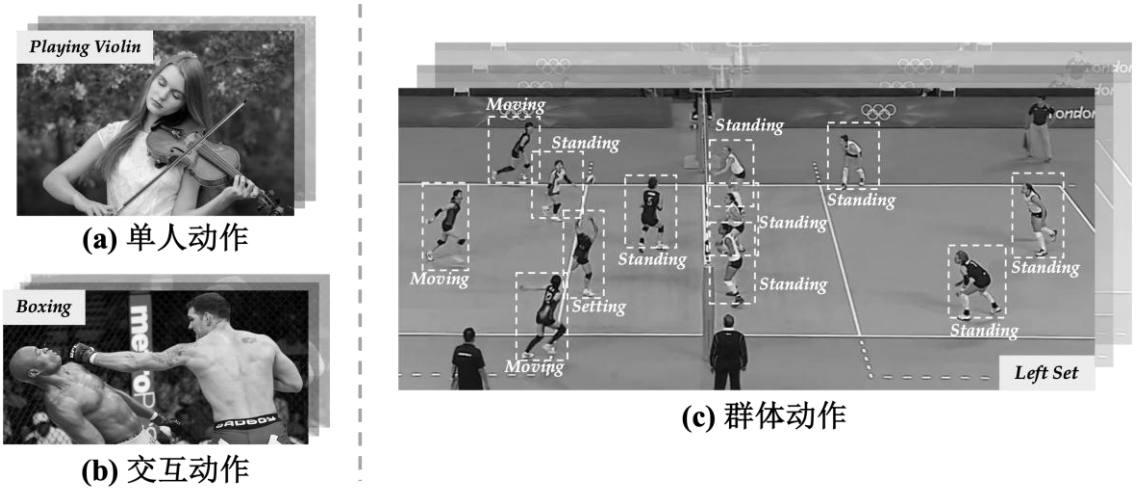

图 1　三种不同类型的人体动作示例

，且个体的行为之间存在关联交互。所以相比于传统的单人和交互动作识别，群体动作识别不仅需要理解每个人的动作，还需基于个体运动信息进行更复杂的推理建模。因此，群体动作识别极具挑战性，已成为视频分析领域的研究热点和难点。该技术在社会公共安全[28]、智能交通[29][30]和智能体育分析[31]等方面具有重要的应用意义。

随着计算机视觉领域中特征表示技术[2][26][32][33][34][35][36]的不断迭代，群体动作识别的研究方案也日新月异。但是目前相关综述陈旧，且采用的归纳依据较为笼统。例如，现有综述[37][38]只依据研究动机或视觉特征提取方式进行分类阐述。然而群体行为识别算法通常包含多个核心步骤，每个步骤涉及的挑战也不尽相同。此种分类方式缺乏组织性和全面性，对初涉此领域的读者而言，不便于梳理整个算法的框架。因此，从一个自上而下的视角对群体动作识别算法逐步骤分析与归纳，将有助于读者快速地理解本领域的研究现状。

为此，本文将群体动作识别拆解为若干阶段或步骤，并对每个阶段或步骤的设计与相关研究作出分析、回顾。具体而言，本文认为现有大多数方案都遵循同一个流程，如图 2 所示：1）个体动作特征提取；2）个体动作特征关联建模；3）个体特征聚合（为群体动作特征）；4）群体动作预测。学者们主要围绕第 1）和 2）两步展开研究。鉴于此，本文主要以常用视频特征和关联建模类型为依据，归纳现有方法并分析它们之间的联系和区别。具体而言，本文介绍了群体动作识别研究中常用的几种动作特征表示（即卷积特征、卷积-序列特征、非局部自注意力特征和双流特征）。然后，本文将个体动作间的关联建模方式归纳为三大类：线性关联、序列关联和图关联（如图 5 所示）。此外，群体动作识别研究中用到的数据集（如图 9 所示）也在不断演化中，本文具体罗列了可用于群体行为识别的多人行为数据集。然而，大部分数据集由于样本量和类别数较少，未能被广泛使用。因此，本文在最为常用的三个数据集（Volleyball、Collective Activity 和 NBA）上对现有主流方法进行了对比与分析。

本文的组织结构如下。第 2 节介绍了群体动作识别问题的定义和挑战，第 3 节回顾了该任务常用的视频特征表示。第 4 节根据关联建模类型归纳、分析和对比了现有研究工作，第 5 整理并分析了现有的一些可用数据集，第 6 节分析和对比了现有方法在三个流行数据集上的性能表现。第 7 节和第 8 节讨论了群体动作识别未来的研究趋势并进行了总结。

2 问题介绍

2.1 基本定义

群体动作识别（group activity recognition, GAR）[15][16][17][39][40]旨在识别任意场景内由一群人互动完成的动作。具体而言，给定一个分辨率为的 $H\times W$ 视频，从中采样出 T 帧作为输入数据，$V\in\mathbb{R}^{T\times 3\times H\times W}$。对应地，还会给定视频中 K 个人物的运动轨迹 $B\in\mathbb{R}^{T\times K\times 4}$（4维位置坐标表示坐标框），如图1所示。值得注意的是，这样的人物运动轨迹通常是利用成熟的跟踪算法[41]对中间帧的精准人物标注框进行跟踪生成的。部分方法[17]会使用检测

算法从中间帧提取可能的人物框，替代原有的人工标注。学者们需要在此基础上设计模型 M 去预测群体动作类别。

2.2 经典流程

本文将群体动作识别流程定义为：

$$c_i^g = M(V, B) \tag{1}$$

其中，$c_i^g \in C^g$ 为预先定义的群体动作类别集合中的某个标签类别。群体动作内包含若干个体动作，实践表明同时或分步进行个体和群体动作的识别有助于深入理解群体动作。因此，现有大部分方法将群体动作识别分解为两个识别任务来执行。公式（1）则改写为：

$$c_i^{ind}, F^{ind} = M_1(V, B), c_i^g = M_2(F^{ind}) \tag{2}$$

其中，M_1 和 M_2 分别指个体动作识别和群体动作识别阶段的子模型。c_i^{ind} 代表个体动作预测类别，F^{ind} 代表个体动作特征。多年来，大部分方法[16][42][43][44][45][46][47][48][49][50]都遵循图2中的两阶段识别流程，包含个体动作识别阶段和群体动作识别阶段。**个体动作识别阶段**：根据给定的个体位置信息，从视频中提取个体动作特征用于识别每个人的动作。**群体动作识别阶段**：对第一阶段提取的个体动作特征进行更高层级的推理建模以挖掘个体间的交互关系，并基于此交互关系获得最终的视频级特征以完成群体动作识别。这两个阶段具体包含以下四个步骤：

- **个体动作特征提取**：一般有两种常见做法，1）预先根据目标框从全图中裁剪出若干小图片，然后将其送入编码器提取个体特征[16][42][43]；2）直接对全图提取特征，再根据目标框用RoIPool[51]或RoIAlign[52]来选取出对应到每个人的个体特征[17][20]。第二种方法通常依赖较大的输入图像尺寸，会明显增加特征编码器的参数量和计算量。受限于计算资源，早期工作大多采用第一种方案。
- **关联建模**：群体动作由多人共同执行而成，个体间存在不可避免的交互。如何基于所得个体特征来进行更高层级的交互推理建模是群体识别的核心研究内容。近年来的研究中有关该部分的设计与思考主要围绕递归神经网络[42]、图神经网络[53]和注意力机制[54]来展开。
- **特征聚合**：将来自不同时刻的不同个体特征融合为一个视频级特征向量，以表示整个视频中涉及的群体动作。这一步的实现方式有很多种，比如平均池化[55]、最大池化[16]和递归聚合[42][43]等。值得注意的是，特征聚合的操作也会被设计在关联建模中[43]。
- **学习目标**：在训练阶段，绝大部分方法会同时将**个体动作识别**和**群体动作识别**作为模型的学习目标，以确保个体特征表示的有效性。也有一些工作尝试丢弃个体动作识别[55][56]任务或新增人物检测[44]任务。

本文着重讨论个体动作特征提取和关联建模两个步骤，并在章节 2 和章节 3 详细回顾和归纳了相关文献。受限于计算资源，早期部分方法[16][42]都是分步执行上述两个阶段的识别任务（即个体和群体动作识别），本文称之为**两阶段分步法**。后续大部分工作采用了端到端的训练策略[17][44][55]完成两个阶段的识别任务，本文称之为**两阶段端到端方法**。有少量工作[57][58]尝试直接使用每帧的全图表示用以群体动作识别，这类似于传统动作识别的流程，因此本综述不再赘述。

2.3 面临挑战

群体识别是一项复杂的视觉任务，所需理解的视觉信息错综复杂。这给学者们在算法设计方面带来了诸多挑战，本文总结如下。

可视信息冗余多变：群体动作场景的视野通常较大，涵盖若干人物和大量嘈杂信息，给群体动作表示的构建带来巨大挑战。其一，场景中许多人物

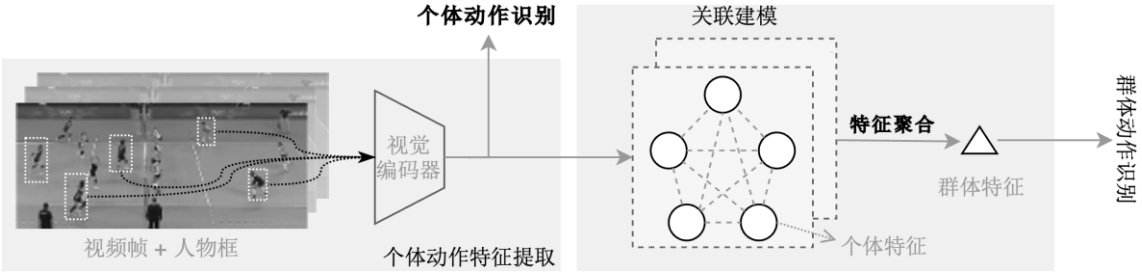

图 2 基于深度学习的群体动作识别经典流程示意图

和背景信息是冗余无效的；其二，人物数量的不断变化给模型设计带来困扰。因此，如何抑制无效视觉信息，基于多变的人物构建有效表示现有研究的主要任务。

多人物间层次化时空建模：群体动作场景内含有多个人物，他/她们之间的交互产生了不同层次的语义，例如人与人，人与群体，子群体与子群体等等。这些交互的发生于时空域中，如何对其动态地建模是群体动作识别任务的关键挑战之一。

细粒度监督成本高：细粒度的人或物的监督信息（例如位置和类别）有助于提升群体动作识别的准确性。但是对于实时应用场景而言，获取这类细粒度信息的成本极高甚至无法获取。在缺失细粒度监督信息的情况下，大部分算法性能下降明显，有些算法甚至无法使用。因此，如何缓解群体动作识别算法对细粒度监督信息的依赖是当前的研究热点。

3 常用动作特征回顾

在群体动作识别研究中，通常需要利用现有视频表示算法从视频中提取个体动作特征。因此，本节着重介绍从RGB视频帧中提取的常用动作特征，即卷积特征、卷积-序列特征、非局部自注意力特征，并在表1进行了归纳与对比。此外，少数工作将光流[57][59][60]、姿态[61][62][63]等模态数据作为补充，以增强个体动作特征的表征能力。特别地，本文将这类从多模态数据中提取的特征，统一归纳为多流特征。**需要强调的是，本节着重阐述群体识别任务中如何提取各个类的群体动作特征，并在表1进行了对比与讨论，并未展开介绍各类经典算法的具体实现。**

3.1 卷积特征

得益于卷积操作的强大空间表征建模能力，2D卷积神经网络（convolutional neural network，CNN）[64][65]在各类基于图像的视觉任务（图像分类[64]，检测[51]和分割[66]等）中被广泛使用，比如AlexNet[64]、VGG[67]、ResNet[33]和GoogLeNet[68]。为建模视频数据中独特的时序信息，学者们将2D卷积扩展为3D卷积[69][70]并提出3DCNN，例如C3D[71]、I3D[72]和S3D[73]。

现有的群体动作识别研究中，广泛使用了上述的两类卷积网络（2DCNN和3DCNN）以提取个体动作特征。由于个体动作类别有限，许多工作[17][48][55]直接利用较浅的2DCNN（如AlexNet，ResNet-18）从若干帧中提取静态空间特征并在时序维度作简单的平均池化，就足以建模个体动作。此外，最近的一些群体动作识别方法采用了更深的2DCNN（如Inception和VGG）[55][69]或者3DCNN[66][68]，但这不仅增益有限而且明显增加了算法的复杂度和计算量。

3.2 卷积-序列特征

递归神经网络（recurrent neural network，RNN）在序列建模方面的能力，已经在自然语言处理[74][75]和语音识别领域[76][77]得到了充分验证。因此，有学者提出综合利用CNN和RNN从视频数据中提取时空表示[2]。具体而言，这类方法使用卷积网络CNN从视频帧中独立提取静态空间特征，然后将若干个时刻对应的独立视频帧表示送入到RNN结构中以构建时序联系。RNN结构通常用长短期记忆网络（long short-term memory，LSTM）[78]或门控递归单元（gate recurrent unit，GRU）[79]来实现。

在群体动作识别研究中，学者们通常先使用成熟的2D卷积网络（CNN）提取每个人物的空间特征，再将同属于一个人物的不同时刻的卷积表示送入到一个RNN中以构建时序表示（具体如图3所示）[16][42][45][47][49][67]。现有工作多采用LSTM[78][80][81]来提取个体动作的序列特征[16][42][45][47]。相比于前文中提到的卷积特征，卷积-序列特征可以轻松地建模个体动作的长期时序依赖。但是递归神经网络无法并行优化，过长的序列数据会导致模型计算效率显著下降。

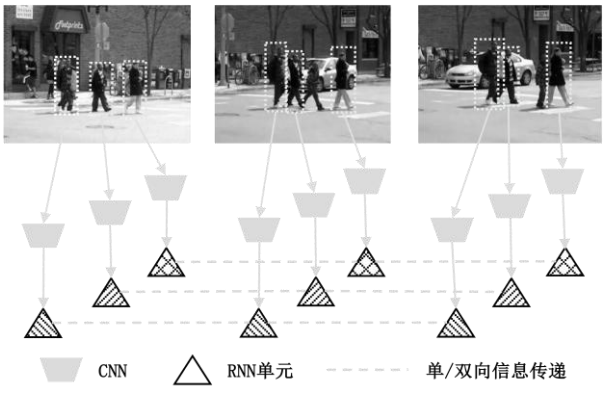

图3 群体动作识别中“卷积-序列”个体特征编码示意

3.3 非局部自注意力特征

由若干非局部自注意力层堆叠构建的Transformer[82]在自然语言处理领域展现出强大非

表 1 常用特征类型对比

特征类型	优点	缺点
卷积特征	擅长局部空间建模；并行度高	无法建模长距离依赖；计算开销大
卷积-序列特征	擅长捕获时序特征；易于和其他结构结合	对并行计算不友好
非局部注意力特征	擅长非局部依赖建模；计算开销相对较低	依赖较大规模的数据样本
多流特征	动态时序建模能力强	额外模态数据的提取成本高

局部建模能力。Dosovitskiy 等人[83]尝试将图像划分为均匀的小块并进行简单编码，接着送入到多层 Transformer 中以构建非局部特征。随后，在视频分析领域也涌现出一系列基于 Transformer 的非局部视频表示方法，例如 TimeSformer[84]、MViT[85]和 MotionFormer[86]等。

然而，有研究[83]表明 Transformer 结构缺乏 CNN 结构中先天的归纳偏置，直接将 Transformer 应用在小规模数据上容易过拟合。因此，鉴于现有群体动作识别数据集规模普遍较小，现有方法并未直接使用 Transformer 从像素级视觉信息中提取特征，而是使用 Transformer 进行高阶的关联建模。例如，Gavrilyuk 等人[68]和 Yuan 等人[87]尝试将基于卷积网络提取的个体动作特征直接送入较深的 Transformer 结构中以构建非局部依赖交互特征，并取得了很好的效果。未来期待更多的工作能够尝试用 Transformer 从底层视觉信息中直接提取群体动作表示，以增强各层级视觉特征的交互表示。

3.4 多流特征

传统 2D 卷积神经网络从视频数据中提取特征时，擅长于挖掘静态空间线索，但却忽略了人物动作的动态时序线索。为此，有学者提出一种双流网络(Two-Stream Network)[1]，采用两个支路分别提取视频数据中的空间线索和时序线索。具体而言，双流网络中的空间支路负责抽取 RGB 模态数据中的外观、场景、物体等静态空间线索，而时序支路则以光流图像序列为输入并挖掘人体动作的动态时序线索；之后，两个支路编码所得特征会被融合以完成识别任务。通过将空间和时序分开处理，双流网络使得动作特征表示的质量获得了极大的提升，并成为了动作特征表示领域的一个里程碑，启发了后续诸多工作。

在群体动作识别任务中，现有基于双流网络的方法[43][61]通常是根据人物运动轨迹先从视频中截取出人物运动的图像序列，然后分别编码 RGB 图像序列及其对应的光流图像序列，从而挖掘个体运动的静态空间线索和动态时序线索。两个支路所得特征被拼接为最终的个体动作特征，以用于后续的群体动作特征构建。双流网络结构多采用现有 2D 卷积网络（如 AlexNet [64]和 GoogleNet[68]）作为各个分支的特征编码器。

此外，现有不少群体动作识别算法[61]将双流结构拓展到多流结构，额外再从姿态[61][62][63]数据中提取运动特征。因此，本文将基于非 RGB 模态数据提取到的特征统称为多流特征（如图 4 所示）。总体而言，多流特征有效地刻画了人物运动线索，但其高昂的光流模态信息提取成本，限制了其在边缘设备或实时场景下的应用。

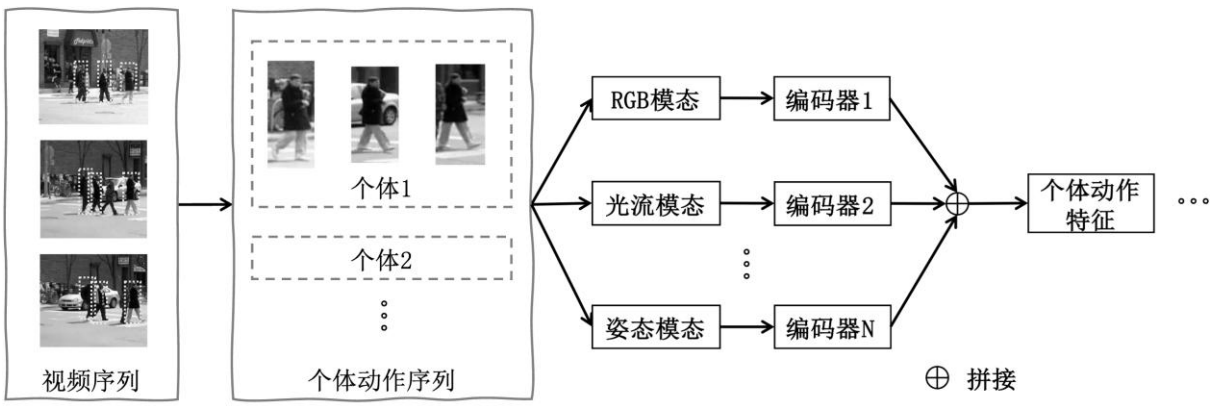

图 4 群体动作识别中多流特征编码示意图

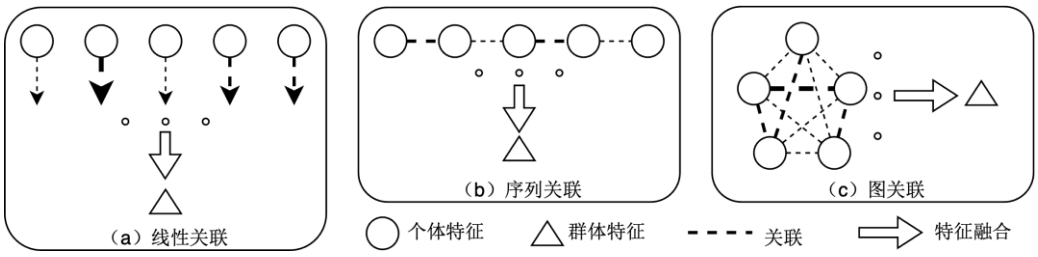

图 5 三种不同的关联建模方式（空心圆圈代表省略，线条的粗细代表不同权重值）

4 常用关联建模方法回顾

本章节首先将现有群体动作识别研究方法中涉及的关联建模方式归纳为三类：线性关联、序列关联和图关联。以此为依据，本节还回顾了基于上述建模方式的研究方法，并讨论了各类方法的优劣。

线性关联：基于若干个可学习的权重，对多个个体动作特征表示进行加权平均。这样的权重一般可以利用注意力机制[42][60][88][89]捕获个体特征与全局上下文之间的关系来获取。该过程往往不涉及个体特征之间的显式交互，故称为线性关联，如图 5-(a)所示。

序列关联：以序列的方式将若干个体动作特征串联起来，它们之间的连接通常是一种信息传递。常用的实现方式是利用递归神经网络[90][78]在输入的个体特征序列上构建单向[43][78]或者双向[42][45]的信息传递。序列关联在个体动作特征间引入了稀疏的交互，但这种交互局限于相邻节点间，如图 5-(b)所示。

图关联：在任意两个个体动作特征之间构建关联，最终呈现出类似图结构的关联，如图 5-(c)所示。这样的图关联可以通过图卷积神经网络[91]、图递归神经网络[53][92]或非局部网络[82][93]来实现。

4.1 线性关联方法

线性关联的核心是为个体动作特征寻求合理的权重以加权融合成群体表示（如图 6 所示）。现有大部分研究都借助注意力机制来实现，其中又可细分为两小类：1）基于空间：单独从每帧的个体动作特征中学得关联权重以聚合为群体表示；2）基于时空：额外考虑个体动作的时序依赖，学习出具有时空上下文的线性关联权重。相关的工作如下所述。

4.1.1 基于空间的线性关联方法

Ibrahim 等人[16]提出一个层级式深度时序模型（hierarchical deep temporal model，HDTM），类似于图 2 中所示结构。HDTM 包含两个阶段，第一阶段：采用卷积时序结构（CNN+LSTM）建模个体动作的时空表示，并在每个时间步使用简单的平均池化将它们聚合为静态群体表示；第二阶段：采用另一个 LSTM 从每帧对应的群体表示中挖掘最终的动态群体表示，以识别群体动作。该方法[16]将个体特征进行平均聚合，无需额外参数，属于最简单的一种线性关联方法。基于此，后续涌现了诸多两阶段的线性关联方法。

Yan 等人[42]观察到群体动作中少部分人物的瞬间运动对群体动作的贡献更大。为此，他们提出一种参与度敏感的方法[42]，在每个时刻直接从个体动作特征本身独立地学习出注意力权重，以挖掘场景中运动速度快但与群体动作高度关联的人物动作特征。同期，Tang 等人[60]提出了一种语义保留教师-学生模型。在语义域，教师网络将个体动作的标签文本作为输入，并利用注意力机制学习出文本语义与群体动作的关系；在视觉域，学生网络利用空间注意力机制学习出视频帧与群体动作的关系。在训练过程中，教师网络的注意力知识用于引导学生网络中对个体动作注意力权重的学习，以有效挖掘关键人物并剔除无关人物。上述两种方法都注意到了群体动作场景中人物的冗余，因此都聚焦于设计一种能够抑制无关个体动作、突出关键个体动作的特征聚合方法。为了完成这一目标，两种方法都引入了自注意力机制，但其中自注意力机制作用的上下文信息有所不同。Yan 等人[42]着重探索视觉信息中提供的关键线索，而 Tang 等人[60]认为视觉-文本间的关键线索应该具有一致性。两种方法都带来了显著增益，启发了后续诸多基于自注意力机制的线性关联/图关联方法。

然而，Yan 等人[42]和 Tang 等人[60]提出的方法仅仅关注了单层级信息融合，即个体动作特征聚合为群体特征的过程。因此，Kong 等人[94]和 Lu 等人[95]进一步将线性关联建模引入到更多层级信息中。具体而言，Kong 等人[94]提出了一种层级式注意力网络（hierarchical attention network，HAN）。HAN

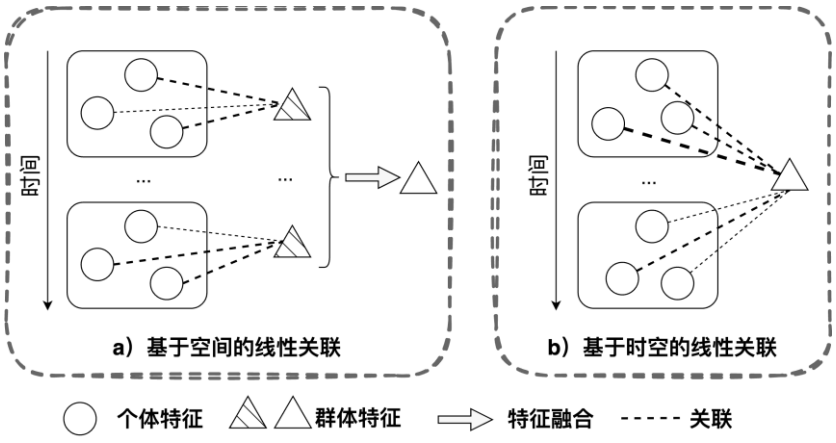

图 6 线性关联建模的不同实现方式（线条的粗细代表不同权重值）

的核心是在肢体动作-个体动作两个层级构建注意力机制，以捕捉与群体动作高度相关的独特局部或个体特征。此外，获得关注后的特征被送入后续的 LSTM 单元以构建完整时序上的个体动作表示，但不同个体间不涉及交互。类似地，Lu 等人[95]提出了一种双层注意力交互模型，分别依赖于“个体”和“场景”两个级别的注意力机制。个人级的注意力由姿势特征（作为上下文信息）来引导，用以建模场景内若干个体间的交互关系。场景级的注意力用以构建个体动作与高层级群体动作之间的交互。Kong 等人[94]和 Lu 等人[95]提出的多层级线性建模方法更深层次、更细粒度地挖掘了人体局部运动信息，从而能够更好地理解群体动作。这也启发了后续诸多基于多层级信息关联建模的工作。

另一方面，Azar 等人[96]将线性关联拓展至多模态数据以提升个体动作特征表示质量。他们分别在 RGB、姿态、光流和扭曲光流四种模态上提取个体动作特征表示，并采用最为简单的最大池化操作将其融合为群体动作表示。其中，每个模态的输出又被细分为个体级和场景级表示，分别用以预测个体和群体动作。最后，各模态预测结果会被融合为最终预测结果。

4.1.2 基于时空的线性关联方法

上一节中阐述的方法都只考虑了某个时刻上，肢体/个体动作特征与群体动作间潜在的线性关联。但事实上，行为动作的发生一定是具有时空性的。不同时刻间的线性权重相互影响，更加符合人类对动作的感知过程。因此，有学者们进一步将简单有效的线性关联思想扩展到时空维度，并提出一系列代表方法。

Qi 等人[97][98]提出了多层级注意力机制（“身体局部级-个体级-帧级”）以挖掘关键时刻的关键人物的关键部位。具体而言，身体局部注意力将身体局部特征加权融合为个体动作表示；个体注意力将每一帧中个体表示聚合为静态群体表示；帧级注意力旨在以不同权重将每帧的静态群体表示聚合为群体动作表示。

然而，上述方法中在完成单帧群体表示的构建后才进行时序建模，忽略了群体行为中更深层次的时空特征，例如个体特征间和身体局部特征间的时空关联。因此，Xu 等人[99]提出了一种基于时空注意力的多模态关系模块，在个体特征和帧级特征层面都进行了时空线性建模。该方法用空间注意力对个体动作特征（包含外观特征、几何位置特征和光流运动特征）之间的交互关系进行建模。然后，提出了一个基于注意力的时序聚合层来挖掘帧之间的语义关系，以不同权重将帧级特征融合成有效的视频级表示。

实际上，随着建模层级增多，时空线性关联的推理过程愈发接近图模型推理。因此，后续学者开始采用更为统一的图模型来将场景中的局部信息（包括肢体或个体）建模为群体动作之间的关联。

4.2 序列关联方法

序列关联的主要特点是只在相邻特征节点间构建有限的联系，优点是计算量小。大部分方法都是采用单向或双向的递归神经网络来实现。

Wang 等人[43]首次以序列形式建模群体动作中个体间交互，提出了一个递归的多层上下文交互框架。该框架由三层上下文交互组成，包括个体级、组级和场景级。其中的组级交互构建在“子动作”组之内，场景级交互构建在“子动作”组之间。这里的“子动作”组是根据图划分算法[100]划分场景

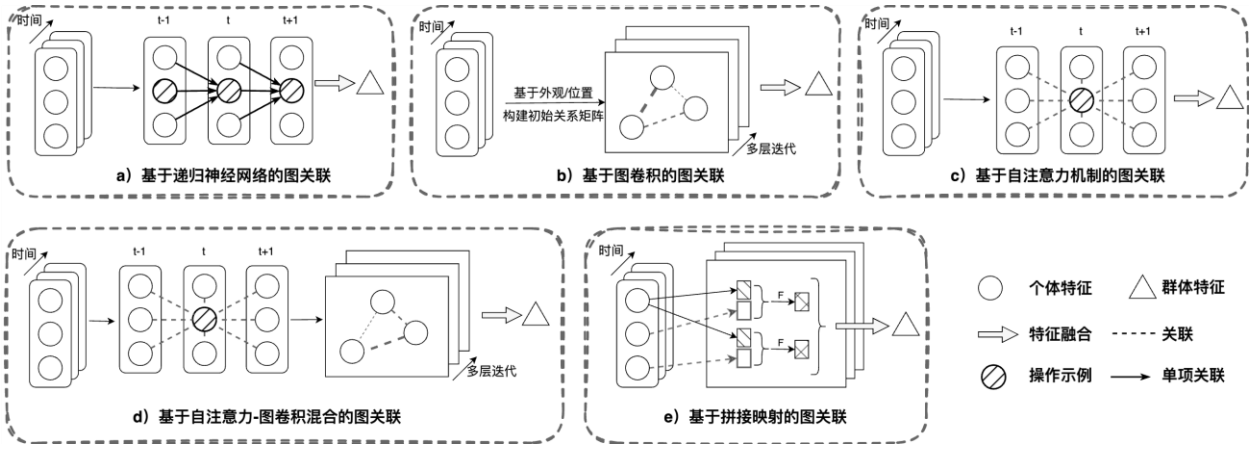

图 7 图关联建模的不同实现方式（线条的粗细代表不同权重值）

内人物得到的。三层级的上下文交互都是通过 LSTM 来进行序列式的信息传递，并将序列末端节点特征输出到下一个层级。其中，三个层级中的 LSTM 独立不共享，且输入特征节点按照位置顺序排列。场景级 LSTM 的末端节点输出作为最终的群体动作表示以分类。

上述方法[43]验证了序列关联在群体动作识别中的有效性，但多层级的序列建模所需计算成本较高。Yan 等人[42][89]和 Tang 等人[45]发现，将这种序列关联仅用于个体动作间的特征交互建模即可显著提升群体表示的质量，同时也能够控制计算成本。因此，Yan 等人[42][89]不仅对场景内个体表示构建线性关联以捕捉重要个体，也使用了 LSTM 来构建个体间的序列交互和特征融合。Tang 等人[45]也将图结构推理后的个体关联特征送入到 LSTM 进行序列关联建模。

此外，Wang 等人[43]提出的方法被嵌入至对抗生成网络（generative adversarial network，GAN）后，序列关联建模可以很自然地生成不同层级的序列关联表示。Gammulle 等人[101]提出了一种基于条件 GAN 的多级顺序生成对抗网络（MLS-GAN），包含生成器和判别器。在生成器框架中，首先使用一个时序 LSTM 将个体动作特征和场景特征序列映射为单一向量表示；然后设计了一种门控融合单元（gated fusion unit，GFU）对所得两层级的时序表示（个体动作和场景）按照位置顺序进行序列关联，并生成群体动作编码。而在判别器框架中，给定群体动作编码和场景视觉表示，用相同的门控融合单元（GFU）进行特征融合以供判别分类。

同样地，Bagautdinov 等人[44]也利用 RNN 结构对个体动作特征进行关联建模。不过该方法将个体人物检测、个体动作识别、群体动作识别任务相结合，以多任务训练方式提升了个体运动轨迹的质量，增强了人物间关联表示，以准确估计群体动作。这样的架构不依赖于外部检测算法，而是经过端到端地训练以生成密集的检测框。

此外，为了进一步增强人物运动表示的判别性，有工作将序列关联结构拓展至多模态数据。例如，Li 等人[58]将原始 RGB 视频帧和对应的光流图像各自输入到两个不同的 CNN 模型之中，并输出四类表示作为时空 LSTM 模型的输入。此时，空间 LSTM 结构用于探索空间内不同模态的全局表示间的联系。之后，将空间 LSTM 生成的多帧表示送入时序 LSTM 结构中将其进一步聚合，以预测群体动作类别。基于此，Li 等人[57]又进一步利用文本模态帮助理解群体动作的语义。他们将前述空间 LSTM 替换成一个基于 LSTM 的文本生成器，从每一帧中生成一段文本描述送入时序 LSTM 以聚合，用于预测群体动作。

4.3 图关联方法

图关联将场景中所有个体动作视为节点，按照一定规则在这些节点间进行信息传递，以构建出潜在的图结构联系。前文所述线性关联和序列关联可以视作图关联的特例，但它们确实也在一段时期内深深影响并启发了一类方法的设计与思考，故在前面几节作详细分析。而本节则重点回顾真正意义上的图关联方法，其中的图关联过程主要借助以下几种常用手段来实现（如图 7 所示），包括：1）基于具有时空约束的递归；2）基于标准图神经网络；3）基于自注意力机制。

4.3.1 基于递归神经网络的图关联方法

标准递归神经网络无法被直接用于处理群体

动作识别问题中的时空序列数据。如图 7-（a）所示，现有方法通常需要将递归单元更改为能够接受多个前序输入的结构[45][47]，或者叠加两个递归神经网络分别处理时序和空间维度数据[98][102]。

具体而言，Biswas 等人[102]提出结构化递归神经网络（structural recurrent neural network，SRNN[103]）以构建出两种不同类型的层级式模型，从而挖掘场景内人物间的时空上下文。SRNN 包含 nodeRNN 和 edgeRNN 两部分，nodeRNN 用来对场景内每个人物在时序上建模个体动作表示，edgeRNN 用来构建不同人物间的关联。整体来看，该方法可以在人物间构建出一个具有时空图结构的上下文关联。此外，Biswas 等人[102]还提出了两种变体：1），先 edgeRNN 后 nodeRNN（SRNN-MaxNode）和 2），先 nodeRNN 后 edgeRNN（SRNN-MaxEdge）。类似地，Qi 等人[98]利用结构化递归神经网络从视觉和文本（个体动作预测标签）两个模态中构建语义关系图，并送入层级式注意力模型中挖掘重要个体特征；之后，将这些个体特征聚合为群体表示。但是上述两种方法中的结构化递归神经网络对时空数据的处理依旧是不同步的，即时空关系被割裂，这不利于对群体动作语义的深度理解。

因此有学者开始关注如何保持群体动作场景中个体动作间的“时空共现”。其主要实现方式是将递归神经网络单元修正为可以接受多个前序输入的结构，以支持在某个时刻处理场景内多个人物特征。Tang等人[45]提出了一个带有“时-空”置信约束的图结构LSTM单元来建模个体人物在时空域中的动态关联演变过程。其中，时序置信约束意在保持某个各期的当前特征和时序特征的一致性；相似地，空间置信约束控制了某个时刻内该个体特征与其空间上下文表示的一致性。这里的空间上下文来自于前一个时刻所有其余个体特征的汇聚。Shu等人[47]提出了Graph LSTM-in-LSTM（GLIL）网络，不仅保持了图结构的LSTM单元，还额外引入了具有残差连接的残差LSTM来学习人物级别的残差特征。GLIL是一种“宿主-寄生”架构，它在局部视图中用来建模个体交互的个体LSTM(寄生虫架构)，或者全局视图中建模群体级别交互的图LSTM（宿主架构）。

无论是构建结构化的递归神经网络[98][103]还是带有时空约束的图型递归神经网络[47]，在建模群体场景中的时空图关联时，都面临着严重的效率问题。因为递归神经网络难以并行化，当其拓展为时空结构后，计算效率会更低。因此，基于递归神经网络的图关联方法并未能成为主流，学者们纷纷转向利用图卷积结构实现图关联，从而更高效地建模群体场景内多人间的交互。

4.3.2 基于图卷积的图关联方法

如图7-（b）所示，现有方法基于人物外观特征或位置信息构建初始人物关系矩阵，然后叠加多层图卷积网络对场景内个体人物进行关联建模。现有基于图卷积网络（graph convolution network，GCN）的方法主要分为两个发展阶段：早期简单应用阶段和后期改进阶段。早期方案多是将图卷积网络简单应用于个体人物间的关联建模，相比于线性或序列关联而言带来明显增益。后期方法开始考虑如何定制化地构建场景中个体间的关联建模，衍生出两类趋势：动态修正和多图互补。

图卷积最初是由Wu等人[104]引入到群体动作识别任务中的。基于外观和位置，他们提出利用非局部相似计算构建出人物关系图（Actor Relation Graph，ARG）。然后将此关系图作为标准图卷积网络中邻接关系矩阵的初始值，并构建多层图卷积以实现个体动作间的关联推理。为了满足对实时性的需求，他们进一步提出了空间局部ARG和时间随机ARG来实现人物关系图的稀疏化。

ARG方法假设人物在时空演变过程中时刻保持着关联，每一个空间关系图也是相互独立的。这自然是不符合群体动作发生规律的。因此，后续有不少对此方法[105][106][107]的改进，主要包括动态地修正人物关系图以寻得最优关联结构或关键个体动作特征，并通过聚类对人群进行分组等。

不同于 ARG[104]对所有个体人物都进行建模，Mao 等人[105]在不同尺度下将场景内人物以聚类的形式划分为若干组，并利用 GCN 对多个分组进行关联建模。此外，Yuan 等人[106]基于 ARG[104]提出了一个由 dynamic relation（DR）模块和 dynamic walk（DW）模块组成的动态推理网络（dynamic inference network，DIN）来实现个人动作间的时空推理。给定一个局部交互域，使用 DR 预测关系矩阵并用 DW 预测动态偏移。通过不断更新，模型能够构建出具有全局视野的特征关联。类似地，基于 ARG[104]，Duan[107]等人利用自注意力机制设计了一种提取只含有关键个体子图的方法。该方法以个体动作的重要程度为依据，按照固定比例保留原始关系图中的重要程度较高的节点。个体动作的重要程

度是通过额外的编码器自动学得的自注意力权重，他们尝试了两种不同实现方式（图卷积编码和全连接编码）。

上述两种方法[106][107]中的人物关系图修正过程主要依赖于多批次样本的迭代优化，但在每批样本内仍然停留在单步修正阶段。因此，有学者进一步利用强化学习机制渐进式地在空间内多步修正人物关系图。具体而言，Hu 等人[49]提出了一种基于强化学习的渐进关系学习策略以充分修正群体内人物间的高层级语义关系。该方法利用图网络[108]构建了一个基础的语义关系图（semantic relation graph，SRG）来显式建模人与人之间的关系。为了消除关系图中的群体无关交互，他们设计了两个遵循马尔可夫决策过程的代理，即特征蒸馏（feature distilling，FD）代理和关系门控（relation-gating，RG）代理，以逐步细化 SRG。给定个体动作特征，首先用 FD 代理筛选出信息丰富的视频帧对应的个体表示；然后将它们送入语义图中并用 RG 代理进一步捕获与群体动作语义相关的个体间交互。SRG、FD 和 RG 三者交替优化，相互提高性能。

对初始人物图进行精细化的修正有利于挖掘细粒度的人物关系 [49][105][106][107]，但是这一过程需要精心设计且带来的增益有限。因此，有学者开始考虑是否可以从不同层级数据[109][110]或堆叠多层级的动态图卷积[111]来构建多张人物关系图，以互补理解群体动作。具体而言，Lu 等人[109]基于图卷积网络构建了多层级交互（multi-level interaction relation，MIR）模型。具体而言，他们设计了一种基于关键人物的图池化层（KeyPool）来选择群体动作中的关键人物以构建出粗粒度关系图。然后提出了一种基于关键人物的反池化层（KeyUnPool）来重构出细粒度关系图。通过构建多粒度群体关系图并执行图卷积以捕获多级交互。

Lu 等人[109]旨在从不同粒度的数据中构建多层级人物关系图。与之不同的是，Pei 等人[110]尝试从不同类型数据中构建多种人物关系图，并提出了一个双流关系网络[110]。具体包含两个子网络，即位置分布网络（position distribution network，PDN）和外观关系网络（appearance relation network，ARN）。PDN 和 ARN 分别以个体位置信息和外观信息作为输入，利用图卷积输出两种不同类型的图关系表示，接着拼接二者作为最终用于识别的群体表示。

不同于文献[110]将不同信息独立建模，Feng 等人[111]尝试将多模态信息合并后，在此基础上构建更深层级的图卷积，且每一层输出的人物关系图会通过聚类不断被修正。这一过程也能够获取不同的人物关系图以互补地理解群体动作。具体而言，Feng 等人[111]先将个体动作外观特征和位置特征都划分为若干更小的切片特征，并将两种切片特征拼接为建模单元。然后利用 DeepGCN[112]动态构建关系边，包括使用卷积层更新结点特征、使用最大池化更新结点之间的关系。接着，使用 KNN 动态聚集每个结点的邻居节点，并应用残差思想构建深层图卷积结构来探索个体之间的高层级语义关系，从而识别群体动作。

上述方法在对个体人物进行图关联建模时，忽略了场景内多个人物间的相对位置关系。为此，Azar 等人[59]提出了一种端到端的卷积关系机（convolutional relational machine，CRM），隐式地将人物空间位置关系嵌入到关联建模中。他们利用二维高斯概率密度函数对个人位置信息生成动作图（activity map）。然后利用多个卷积层在若干阶段中不断细化和纠正该动作图。最后通过卷积层将初始特征图和细化后的动作图融合以预测群体动作。

4.3.3 基于自注意力机制的图关联方法

自注意力机制可以理解为图卷积网络的一种泛化形式，主要区别在于邻接关系矩阵的学习方式。图卷积中的关系矩阵需要人为初始化，而自注意力机制中的关系矩阵是自适应学习而得的。因此，自注意力机制具有更强的表示能力且建模方式更加灵活，在近期的群体动作识别方法中备受青睐（如图 7-（c）所示）。自注意力机制的具体实现又呈现出两种形式，即非局部建模[93]和 Transformer[82]。非局部建模是一种单层自注意力操作，而 Transformer 内封装了多层自注意力操作。

具体而言，Wang 等人[93]利用注意力机制在卷积特征图之上构建不同特征点之间的非局部关系，使得图像和视频的表示质量获得明显改善。此后，学者们很自然地将这种非局部建模方法应用到了群体动作识别中。Yan 等人[55]设计了一个层级式的交叉推理网络来构建肢体动作、单人动作和群体动作三个层次的运动特征。具体而言，为了捕获肢体动作间和单人动作间的潜在时空依赖，作者们提出了一个稀疏高效的交叉推理模块（cross inference block，CIB）来替代标准的全连通推理。CIB 的核心思想是将标准非局部网络[93]中的稠密 2D 时空交互，简化为十字交叉的稀疏交互（1D 空间+1D 时

序）。该方法在群体场景中构建起了多层级的稀疏非局部图推理，为群体动作表示带来了质的提升。

上述方法只利用非局部操作在个体/肢体动作特征间构建了浅层图关联表示。同期，Transformer结构在计算机视觉领域的兴起，启发了后续一系列工作[61][113]在个体动作特征间构建深层图关联表示的热潮。

Pramono 等人[113]提出了一种将自注意力增强的条件随机场（self-attention augmented conditional random fields，SA-CRF）。用时序注意力机制学习场景中个体在时序上的演变，用空间注意力捕获个体间的空间依赖。然后使用双向通用 Transformer 编码器将时空关系上下文与场景信息聚合以进行群体动作识别。

不同于 SA-CRF 只将 Transformer 构建于个体动作的单一粒度视觉特征之上，Zhu 等人[114]将Transformer 拓展于个体动作特征的多个粒度，以充分建模群体动作场景内的多粒度行为语义。除此之外，Gavrilyuk 等人[61]尝试将其应用于多种模态的个体动作特征之上以挖掘不同模态间的深层互补信息。具体而言，Gavrilyuk 等人首先从视频输入中提取出多个模态的个体动作表示，如 RGB、光流和姿态。然后基于其上用 Transformer 构建了个体动作间的图关联表示。此外，他们将人物框中心点信息用作标准 Transformer 结构中的位置编码。相比于RNN 或者 CNN，Transformer 中的自注意力机制更适合对群体内非局部特征的构建和融合。

类似于Yan等人[55]对标准非局部操作的稀疏改进，Li 等人[50]使用聚类操作将 Transformer 中的全连通自注意力改进为聚类自注意力，从而增强模型感知少量关键个体动作特征的能力。具体而言，Li 等人[50]提出了一种基于聚类的时空 Transformer（clustered spatial-temporal，CSTT）来同时建模空间和时序上下文以增强个体和群体表示。CSTT 包含编码器和解码器，编码器负责从个体动作特征中提取空间和时序依赖；解码器用于交换空间上下文和时序上下文信息。此外，为了避免大量无关信息的卷入，作者们将标准全连通注意力改为聚类注意力（cluster attention，CA），用于将个体动作特征表示聚成若干组，其中聚类注意力包括组内注意力和组间注意力。

上述工作[50] [55]都是以单一视角将时空域中所有个体动作线索同步建模。然而，Han 等人[115]认为不同群体动作语义依赖于多样的交互线索，并提出了一种独特的双路人物交互（dual-path actor interaction, Dual-AI）框架。该框架将空间和时序上的自注意力图关联灵活地安排在两种不同的互补顺序上，通过整合不同时空路径的优势特征来增强个体间交互关系的表达。类似地，Du 等人[116]基于自注意力机制设计了空间和时序注意力分支，从同一个视频的不同视图特征中独立提取出空间和时序表示，并用对比损失进行约束。对于空间特征，将其帧内“特征对”视为正样本，帧间“特征对”视为负样本，以构建出空间依赖对比损失；对于时序特征，则以相反方式构建出帧间时序依赖对比损失。使用基于上述自监督训练后主干网络提取视频特征，并根据人物框获取对应个体动作特征，用以最终的群体动作分类。相比于以前的方法，这种基于自监督的方法将多人物间关联建模提前至自监督损失构造阶段。但是，当数据量不足时，这类方法的性能将难以得到保证。

总的来看，自注意力机制着重于构建特征间的全局依赖关联，为群体动作建模提供了最大的感受野，但忽略了不同局部信息间的差异化关联。因此，针对该问题，有大量文献[17][50]对标准自注意力机制进行改进。

4.3.4 基于自注意力-图卷积混合的图关联方法

标准图卷积网络结构可以轻松构建出具有局部性质的图结构关联，而标准自注意力机制则作用于全局所有特征。因此，将自注意力机制与图卷积结构结合，能够优雅地结合两类图模型的各自优势（如图 7-（d）所示）。

例如，Yuan 等人[87]利用自注意力计算个体动作特征间的图关联，以替代图卷积结构中预先定义的关系图。具体而言，他们利用 Transformer 从个体动作特征中编码出上下文特征。继而再根据人物空间位置和上下文特征间的相似性构建时空关系图，并用图卷积网络在上下文特征中进行信息传递。其中，上下文编码器首先将全局特征图与个体特征进行对齐，然后使用注意力机制进行特征聚合。增强后的个体动作表示就不局限于某个标注框内信息，而是可以关联到周围更多的判别性信息。该方法分步运用了自注意力机制和图卷积，以优化图卷积结构中的初始关系图。但这并未彻底改进图卷积中局部关系的构建劣势。

因此，有学者受到图注意力网络（graph attention network）的启发，用自注意力机制替换图卷积结构中每层推理中的平均聚合操作，以构建个

体人物在多层级间的不均衡交互。Lu 等人[95]提出了一种图注意力模块（graph attention blocks，GAB）并嵌入到图卷积网络中以分析群体动作。GAB 被应用在了个体层和群体层以挖掘群体动作场景内不均衡交互关系。在个体层中，GAB 在群体表示的指导下，从相邻的个体动作表示中学得不同层次的交互。在群体层中，GAB 估计个体动作对群体表示的不同程度交互。

总体而言，基于自注意力机制与图卷积结合的图关联方法，旨在将非局部和局部特征聚合操作进行一种有机的结合，以扩充模型关联建模的容量。

4.3.5 基于特征拼接映射的图关联方法

不同于标准的图神经网络，有些工作尝试通过将场景内个体动作特征拼接并简单编码（比如全连接）以构建图关联表示（如图 7-（e）所示）。

例如，Ibrahim 等人[48]提出了一个可用于群体动作识别与检索的层级关系网络（hierarchical relational networks，HRN）。给定描述潜在交互关系的图结构，每一个关系层（relational layer）负责将有联系的个体动作表示串联并投影到一个新的关联表示中。值得注意的是，其中输入的关联图结构并非是可学习的，而是直接经验性地将相邻的“个体”或“子群体”定义为有关联。通过堆叠多级关系层，模型建立了层次化关联表示，其中包含个体间和不同规模“子群体”间关联。

不同于 HRN 中的固定关联图结构，Zhang 等人[56]设计了一种隐藏编码从而构建出了可学习的关联图结构。具体而言，Zhang 等人[56]提出一种迭代隐藏编码（iterative latent embedding）策略来构建场景内所有人物间的图关联。具体而言，在给定从场景特征图中获取的个体动作特征，该方法拼接三种特征（场景特征，某个个体特征及其邻近所有个体特征的均值）来构建上下文信息，从而更新某个个体的隐藏变量编码。

上述类型的方法通过特征拼接构建图关联的方法在人物数量不多的情况下是完全可行的，但在密集人群应用中则会遇到计算瓶颈。

4.3.6 基于语义结构的图关联方法

除了基于个体动作特征直接构建关系图之外，有学者尝试探索个体动作类别语义与群体类别语义之间存在的显式或隐式结构化关系。显式类别结构化推理方法是根据模型对个体动作类别预测结果进行进一步的结构化推理。例如，Deng 等人[117]使用卷积神经网络预测场景、个体动作和个体姿态类别，并基于三种预测结果构建图结构消息传递[118]以估计出最终的群体动作标签。基于此，Deng 等人[119]进一步提出了一种结构推理机（structure inference machine），利用递归神经网络构建场景中个体人物间的时空消息传递，并且设计了门控函数以抑制与群体动作无关的个体人物信息并增强相关信息的流通。类似地，Shu 等人[46]设计了一种置信能量层（confidence-energy）从各层级表示对应的群体动作类别的预测分布中，推理出最终的预测结果。此外，隐式类别结构化推理方法[120]是以可学习字典的方式构建个体动作类别原型表示，并用原型表示替换原始视觉信号，然后在此基础上进行推理。

4.3.7 面向弱标注设置的自适应图关联方法

上述大部分方法都是根据准确的人物运动轨迹来提取个体动作特征，从而用于后续关联建模和特征聚合。但在视频数据中标注人物运动轨迹成本极高，使得上述方法难以被应用于实时场景。为此，Yan 等人[17]首次提出了弱标注群体动作识别任务，该任务要求算法在无任何人物细粒度监督的情况下，完成场景内的群体动作识别。该工作启发了后续一系列面向弱标注设置的群体动作识别研究。弱标注群体动作识别的核心挑战在于如何自适应感知群体内关键人物动作并对其进行关联建模。无任何细粒度人物监督信息的情况下，现有解决方案只能从全局视觉数据出发，对所有局部视觉信息都进行图关联建模，基于此再进一步实施筛选、聚合或优化操作，以达到自适应感知的目的。因此，本文将这类解决方案归纳为自适应图关联，并在此进行详细回顾与总结。

Yan 等人[17]利用现成的目标检测和跟踪算法估计出场景内多人运动轨迹，并提出了一个社交适应模块（social adaptive module，SAM）从若干不确定人物动作特征中推理并筛选出有效的群体动作表示。他们首先计算个体动作特征间的非局部密集关系图，然后仅保留 Top-K 个高度关联的节点以获取稀疏关系图。该方法主要是为了解决在弱标注情况下（无细粒度个体人物标注）的群体动作识别。

为了进一步摆脱对检测框的依赖，Wu 等人[121]利用自注意力机制构建可学习的掩码，以定位场景内人物活跃位置，并用其建模所有区域块特征间的依赖关系。它只在训练阶段需要人物运动轨迹框，并从整个场景中定位到活跃位置后自动消除背景特征。该方法可以根据不同尺度的区域块特征自然地推断出层次关系，从而提升群体动作识别性能。

但是，该方法依然需要在训练阶段使用到人物运动轨迹标注，因此依然不够灵活。

为此，Kim 等人[122]尝试用一组少量的可学习向量从整个场景中聚集关键的局部视觉线索。具体而言，作者们基于 Transformer 模型，利用注意力机制对群体动作局部上下文信息进行定位和编码，并将视频片段表示为一组局部上下文特征向量。然后将这些特征向量聚合成一个单独的群体表示，以反映整个群体动作的上下文，同时捕获每个局部上下文特征向量的时序演变过程。

此外，Chappa 等人[123]尝试直接使用自监督方式针对全局视觉信息进行密集的图关联建模，然后自适应地筛选出更具判别性的时空运动表示，以避免对细粒度人物标注的依赖。具体而言，他们采用不同空间块尺寸或帧率从同一个视频样本中构建局部和全局的时空视图特征，来进行自监督对比约束。

4.4 讨论

本文将基于深度学习的动作识别研究发展历程总结在图 8，并将三种关联模型的优缺点归纳在表 2。具体而言，Ibrahim 等人提出的分层时序模型[16]将群体动作识别分解成个体动作分析和群体动作分析两个阶段，启发了后续一系列研究。然而，他们忽略了实际场景中群体动作通常是由少数关键人物所决定的。

基于线性建模的方法最初是为了挖掘场景中与群体动作高度关联的少数个体。该类方法的核心目标是以可学习的方式直接给不同人物分配权重，以突出关键行为线索。现有研究通常利用注意力机制为每个个体动作特征学习一个与群体标签关联程度得分，用来作为特征融合的依据。此类方法只需一个浅层的线性编码即可学到个体动作特征与群体动作类别之间的相关性，因此计算开销比较小。然而，这种建模方式忽略了群体动作场景下广泛存在的人物间交互行为，导致模型对群体动作缺乏深层次理解。因此，在一些人物交互复杂的场景下，这类方法表现不佳。为了解决该问题，现有学者引入序列或图模型捕捉人物间的关联，以及使用更复杂的模型来增强模型在人物关联上的表达能力。

为了更高效地建模群体动作中的上下文信息，一些研究者开始探索基于序列关联的方法。这类方法充分利用递归神经网络（RNN）的序列建模能力来建模个体动作间的关联。通过 RNN 的门控机制，这些方法可以高效地将个体特征稀疏地关联起来，通常只需要 1-2 层建模即可达到预期的效果。这种方法可以有效地捕捉人物间的交互关系，从而更准确地预测群体动作。然而，RNN 的缺点在于无法并行计算，因此这类方法难以被拓展至更长的序列或更深的层次中，这限制了该类方法在复杂场景中的应用。总而言之，基于序列关联的方法是一种高效建模群体动作内上下文信息的方法，但是序列结构的计算效率限制了它在更长序列或更深层次情况下的应用。

为了解决该问题，一些研究者尝试使用其他的神经网络结构来代替线性关联方法中的RNN结构，例如图卷积神经网络和自注意力机制等。这些模型可以并行计算，并且能够处理更长的序列和更深的层次，从而提高了建模效率和准确性。由此演化出了一类基于图关联的方法，这类方法旨在捕捉场景内人物间的潜在复杂交互关系，且能够挖掘不同子群体内和子群体间的模式。这类方法在两个流行的数据基准上不断刷新最高性能。该类方法的缺点是

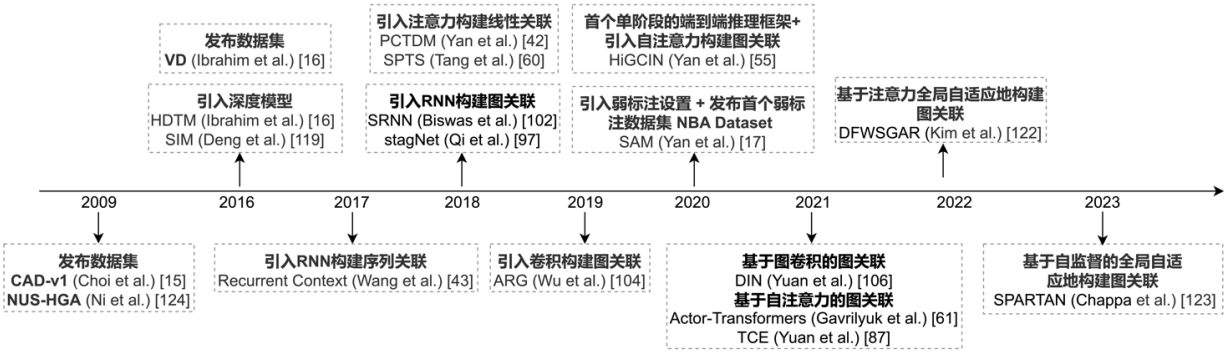

图 8 基于深度学习的群体动作识别研究的发展历程

表 3 群体动作识别的数据集汇总

数据集	发布时间	场景类型	群体动作类别	样本数	类别数
NUS-HGA[124]	2009	监控	群体行走、群体奔跑、站立说话、聚集、打架、人员独立行走	476	6
BEHAVE[125]	2009	日常	聚集但不移动，相互靠近，一起走,相互打招呼,彼此分开,人员独立行走,追逐,打架,一起跑、跟随	174	10
CAD-v1[15]	2009	日常	过马路、等待、排队、步行、交谈	约 2500	5
CAD-v2[126]	2011	日常	过马路、等待、排队、步行、舞蹈、交谈、慢跑	约 3300	6
CAD-v3[127]	2012	日常	聚会、谈话、解散、一起散步、追逐、排队	约 2000	6
UCLA Courtyard[128]	2012	日常	一起散步、排队、小组讨论、坐在一起、小组等待、导览	约 120	6
Nursing Home[129]	2012	监控	跌倒、无跌倒	2990	2
Broadcast Field Hockey[130]	2012	体育	进攻球、任意球和罚球	58	3
Volleyball[16]	2016	体育	左二传、右二传、左扣球、右扣球、左传球、右传球等	4830	8
RIT-18[131]	2016	体育	左发球、左一传、左二传、左扣球、左截击、左一击、左吊球、左拦网、左得分、右发球、右一传等	1530	18
C-Sports[132]	2020	体育	聚集、解散、传球、攻击、游走	2183	5
NBA[17]	2020	体育	三分球-成功、三分球-失败-防守篮板、三分球-失败-进攻篮板、两分球-上篮-成功、两分球-上篮-失败-防守篮板等	9172	9

计算复杂度随着场景内人物数量的增加而呈平方倍的飙升，难以应用于人群密集的实际场景中。现有基于图关联的方法一味追求使用更加精巧但复杂的图结构，以获取更高的性能表现，这可能会导致模型在有限训练数据上过拟合。

综上所述，三类关联方法对群体动作识别任务而言各具优劣。一方面，线性关联和序列关联方法虽然高效简洁，但对群体动作场景内多人间交互建模能力不足。另一方面，图关联方法具有很强的多人间关系表达能力，但其模型计算开销大且容易在有限数据上产生过拟合现象。未来的研究可能需要进一步结合各类关联方法的优势，将各类关联方法部署于算法不同阶段以同步协作解决复杂的群体动作识别任务。

5 相关数据集回顾

一个优质的数据集是计算机视觉研究得以开展的前提，同时也是验证算法有效性的必需品。因此，在研究群体动作识别时，构建高质量数据集是非常重要的一步。与单人动作相比，群体动作涉及到的场景、人物以及运动范围都更加复杂，这使得高质量数据集制作困难。现有群体动作数据集大多采集于监控或者体育比赛视频。本节列举了十二种可用于群体动作识别的数据集，并在表 3 中给出了相关信息的汇总和在图 9 中展示了一些示例样本。值得注意的是，在下述数据集中仅有少数常用于现有研究中，大多数数据集存在采集场景有限或者样本量不足等问题，并未得到广泛使用。

5.1 NUS-HGA[124]

该数据集由新加坡国立大学通过监控室外场景（大学停车场）收集得到的视频组成，共包含五期帧率为 25 FPS 的视频。每期视频都包含 6 种群体动作，每个群体动作又包含 4-8 个人。整个数据集共包含 476 个带标签的视频片段样本，每个类别平均约有 80 个视频片段，每个片段约 8-15 秒。

5.2 BEHAVE[125]

该数据集也来自于监控视频，包含 163 个样本，共计 76800 帧。数据集中包括 10 种群体动作。每个视频中通常有 2-5 个人作为一个小组，组内或者组间存在人物的互动。每个参与互动的人都会被标记上边框，共有 125 个人物，83545 个边界框。该数据集提供了群体动作中的人员轨迹信息以及交互群体的行为描述。该数据集样本量较少，因此在深度学习时代难以被广泛采用。

5.3 Collective Activity[15](CAD-v1)

该数据集是由消费者手持摄像机录制而成，共 44 个视频序列。其中包含 5 个单人动作类别和 8 个人物位置方向类别。该数据集每 10 帧标注出场景内人物位置和动作信息，整个场景的群体动作类别

图 9 不同数据集（CAD-v1[15]、CAD-v2[126]、CAD-v3[127]、Volleyball Dataset[16]、Collective Sports[132]和 NBA[17]）中所含群体动作样本示例

取决于个体动作类别最多的类。该数据集是群体动作识别任务中广泛使用的基准之一。

5.4 Collective Activity Extended[126](CAD-v2)

该数据集是在 CAD-v1 的基础上扩增而得，共计 75 个视频。鉴于 CAD-v1 中“步行”的定义模糊（它更像是个单人动作而非群体动作），CAD-v2 删除了“步行”并添加了“舞蹈”和“慢跑”。其他标注形式和内容与 CAD-v1 一致。

5.5 New Collective Activity[127](CAD-v3)

该数据集由 32 个视频片段组成，包含 6 种群体动作（如表 3 所列），9 种人物位置关系（接近、相反方向行走、面对面、站成一排和并排走等），3 种单人动作（行走、静止和跑步）。同时，该数据集与 CAD-v1 拥有相同的 8 个位置标签。相比于前两个版本，CAD-v3 提供了更为丰富的标注内容。

5.6 UCLA Courtyard[128]

该数据集是一个拍摄于加利福尼亚大学洛杉矶分校校园的长达 106 分钟的高分辨率视频，其包含多个共同发生的群体动作和单人动作。该数据集标注了 6 种群体动作（如表 3 所列）、10 种单人动作及 17 种物体。其中，单人动作包括：滑滑板、骑自行车和骑踏板车等。物体包含：食物、书、汽车和滑板车等。该数据集被以 1:1 的比例划分为训练集和测试集。

5.7 Nursing Home[129]

该数据集采集于疗养院餐厅的监控摄像。选取了 1 个无人跌倒的视频与 13 个有人跌倒的视频。这些视频被划分为 22 个片段，共计 2990 帧。其中

14 个片段用于训练，8 个片段用于测试。该数据集提供了包括行走、站立、坐着、弯腰和摔倒在内的 5 个单人动作注释。同时还将每帧分为有跌倒和无跌倒两个类别，即群体动作类别。该数据集仅包含两个群体动作且采集于特定场景，因此可用性较低。

5.8 Broadcast Field Hockey[130]

该数据集来自于 5 场真实曲棍球比赛，共 58 个视频片段，标注了 11 种原子动作（传球、运球、射门和扑救等）和 3 种群体动作（如表 3 所列）。此外，该数据集还定义了曲棍球比赛中 5 种专业人物角色。该数据集是首个采集于体育运动场景下的群体动作数据集，但由于可用样本量和类别数量较小，所以并未被广泛使用。

5.9 Volleyball[16]

该数据集是从 55 个排球比赛视频中收集而来，共计 4830 个视频段。提供者为每个视频提供了一个群体动作类别，在视频中间帧标注了单人动作类别及其位置。共包括 8 种群体动作（如表 3 所列）和 7 种单人动作（等待、二传、救球、坠落、扣球、阻挡和其它）。

5.10 RIT-18[131]

该数据集是对 Volleyball 数据集[16]的扩充，收集了 YouTube 上的 51 场排球比赛视频，提取了从发球开始经过多次传球直到得分为止的 1530 个视频片段，共 12035 帧。该数据集为每个视频片段注释了群体动作类别和时间边界，包括 18 个群体动作类别（如表 3 所列）。该数据集进一步丰富了 Volleyball 数据集，扩充了更多类别定义和样本数量，但依旧局限于某个特定运动场景中。

5.11 Collective Sports[132]

该数据集是从网络上收集体育视频集，涵盖了篮球、足球、手球、冰球、长曲棍球、橄榄球、排球和水球等 11 种运动，共计 2187 个视频，167935 帧。训练集/验证集/测试集分别含有 1317、435、435 个视频。其中训练和测试分别针对不同的运动类别。每个视频包含至少两个群体动作且被手工裁剪为 5 到 10 秒。视频标记了 5 种群体动作（如表 3 所列）。每个视频都被标注了体育运动类别和群体动作类别。该数据集混合了多种体育运动，为评估多场景的群体行为识别算法提供了保障。

5.12 NBA[17]

该数据集由 181 场篮球比赛视频组成，共 9172 个视频片段，其中 7624 个用于训练，1548 个用于测试。该数据集对每个视频片段标记了共计 9 类群体动作（如表 3 所列）。该数据集的作者们提倡弱标注，即只提供视频级的标注，不提供以往数据集中含有的帧级别细粒度标注（如单人动作、及其位置信息和交互信息）。这给群体动作识别任务带来了新的挑战。

6 现有算法性能对比与分析

本章节首先介绍了群体动作识别中常用的两种评价指标：准确率和平均每类准确率。上一节中介绍了十二种可用于群体动作识别任务的数据集，但其中最为常用的为 VD、CAD-v1 和 NBA，其余数据集并未得到普及。因此，本节只在三个最为常用的数据基准 Volleyball Dataset[16]（VD）、Collective Activity Dataset[15]（CAD-v1）和 NBA[17]上，对现有方法的性能表现从特征和模型两个维度进行对比分析。

6.1 评价指标

群体动作识别是一个标准的分类任务，现有研究中通常使用两种指标[16]衡量模型性能，即准确率（Acc.）和平均每类准确率（Mean Acc.）。

准确率（Acc.）定义为模型预测概率最大的类别与实际类别相同的样本占总样本的百分比。平均每类准确率（$MeanAcc.$）则表示每个类别的分类准确率的平均值，假设数据集中包含 N 个样本，共有 C 个类别，其中第 i 个样本的实际类别为 G_i，预测类别为 P_i，则：

$$Acc. = \frac{\sum_{i=1}^{N} G_i = P_i}{N} \tag{3}$$

$$MeanAcc. = \frac{\sum_{C=1}^{C} Acc._c}{C} \tag{4}$$

$$Acc._c = \frac{\sum_{i=1}^{N} G_i = P_i = c}{\sum_{i=1}^{C} P_i = c} \tag{5}$$

每个类别的准确率 $Acc._c$ 表示正确分为该类的样本数占该类别实际样本数的比例。上述指标 $Acc.$ 和 $MeanAcc.$ 被广泛使用于群体动作识别任务中以评估模型性能表现。

6.2 对比与分析

本文总结了近年来基于深度学习的群体动作识别方法在常用数据集（即 Volleyball、和 CAD-v1）

上的性能表现。此外需特别说明的是，群体动作识别方法通常是系统化的，在没有代码的情况下难以复现。因此，表中有部分结果缺失，但这并不影响对这些方法的对比与分析。具体地，本文从关联建模（表 4）、动作特征编码（表 5）、数据模态（表 6）和弱标注设置（表 7），四个方面进行分析与对比。

6.2.1 不同的关联建模

从表 4 中“关联建模”一栏可以看到不同的个体动作关联建模方式在各个时期都有所涉及。早期受限于算力，基于线性或者序列的关联方法被常用

表 4 现有方法在两个常用基准数据集 VD[16]和 CAD-v1[15]上仅使用 RGB 模态数据的表现汇总（IN 代表 Inception，RN 代表 ResNet，①代表卷积特征、②代表卷积-序列特征）

方法	年份	特征类型	关联建模	性能			
				VD[16]		CAD-v1[15]	
				Acc.	Mean Acc.	Acc.	Mean Acc
Deng 等人[119]	2016	①(AlexNet)	图	-	-	81.2	-
Ibrahim 等人[16]	2016	②(AlexNet+LSTM)	线性	81.9	-	81.5	-
Shu 等人[46]	2017	②(VGG-16+LSTM)	图	83.3	83.6	87.2	88.3
Bagautdinov 等人[44]	2017	①(IN-v3)	序列	87.1	-	-	-
Wang 等人[43]	2017	②(IN-v3 +LSTM)	序列	-	-	-	90.8
Gammulle 等人[101]	2018	②(RN-50+LSTM)	序列	93.0	92.4	91.7	91.2
Azar 等人[96]	2018	①(IN-v3)	线性	85.4	-	-	-
Tang 等人[60]	2018	②(VGG+LSTM)	线性	89.3	89.0	-	92.5
Yan 等人[42]	2018	②(AlexNet+LSTM)	线性+序列	86.2	86.1	-	91.2
Kong 等人[94]	2018	②(IN-v3+LSTM)	线性	85.1	-	84.3	-
Lu 等人[95]	2018	①(IN-v3)	线性	90.6	-	-	-
Li 等人[57]	2018	②(VGG16)	序列	38.7	-	83.7	-
Zhang 人[56]	2019	①(ZF-net)	图	86.0	-	83.8	-
Qi 等人[98]	2019	②(VGG16)	线性	89.3	-	89.1	-
Wu 等人[104]	2019	①(IN-v3)	图	92.6	-	91.0	-
Azar 等人[59]	2019	①(IN-v3+I3D)	图	92.1	-	83.4	-
Tang 等人[45]	2020	②(AlexNet+LSTM)	图＋序列	-	89.3	-	93.0
Yan 等人[17]	2020	①(RN18)	图	94.0	-	-	-
Shu 等人[47]	2020	②(VGG-16+LSTM)	图	-	93.0	-	94.9
Hu 等人[49]	2020	②(VGG-16+RNN)	图	91.4	91.8	-	93.8
Yan 等人[55]	2020	①(RN18)	图	91.4	92.0	93.4	93.0
Gavrilyuk 等人[61]	2020	①(I3D)	图	93.0	-	92.8	-
Lu 等人[109]	2020	①(IN-v3)	图	91.9	-	90.6	-
Xu 等人[99]	2020	①(IN-v3)	线性	92.8	-	-	-
Yuan 等人[87]	2021	①(VGG16)	图	94.1	94.4	-	95.4
Li 等人[50]	2021	①(I3D)	图	94.9	-	94.7	-
Pramono 等人[113]	2021	①(I3D)	图	94.8	-	94.4	-
Yuan 等人[106]	2021	①(VGG-16)	图	93.6	93.8	-	95.9
Han 等人[115]	2022	①(IN-v3)	图	94.4	-	-	96.5
Li 等人[120]	2022	①(IN-v3)	图	93.5	93.9	96.5	95.3
Zhu 等人[114]	2022	①(IN-v3)	图	94.5	-	96.8	
Liu 等人[133]	2022	①(I3D)	图	91.9	-	90.5	-
Mao 等人[105]	2023	①(IN-v3)	图	93.1	-	92.5	-

于挖掘群体中浅层但有效的人物交互信息。但这类方法所建模的人物交互关系较为简单，在基准数据集上的性能很快遇到了瓶颈。因此，基于图关联的方法自 2019 年开始在群体动作识别的研究中兴起。该类方法能够深度挖掘群体动作场景内在的结构化信息，在三个不同数据集上都能够取得最佳表现，备受学者们的青睐。此外，从表中结果来看，图关联方式在强标注数据集（即 VD 和 CAD-v1）上已经达到饱和状态，但在弱标注数据集（即 NBA）上还有一定的研究空间。由此可见，在不准确的个体动作特征输入的情况下，如何有效挖掘关键视觉线索并进行图关联建模依然有待进一步深入的研究。

综上所述，基于线性或序列的关联方法高效简洁但仅能建模场景内多个人物间的浅层交互依赖关系，而基于图关联的方法能够深度挖掘多人物间的交互依赖关系但计算成本较高。因此，本文认为基于图的关联建模依然会是未来几年群体动作识别研究的首选方式，但需要进一步提升该类方法的计算效率。

6.2.2 不同动作特征编码

从表 4 和中“特征类型”一栏中可以观察到不同方法所采用的深度视觉特征类型各不相同。大部分方法仅使用单个深度网络从视频中提取视觉特征，有些方法[43][50][58]会用到多种网络结构提取不同模态（如光流和姿态等）的特征表示。在同一年份内，使用较深的卷积网络（例如 VGG，Inception 等）所提取的特征优于使用一些较浅的卷积网络（例如 AlexNet）所得到的特征。随着计算力的提升，目前最新的方法均采用较深的 2D 甚至 3D 网络结构。

为了探究同一个方案下不同深度网络结构对方法性能的影响，本文在表 5 统计汇总了现有文献中采用不同网络结构的实验结果。从表 5 中可以观察到，对于同一个方法而言， 特征编码网络结构越深越复杂，群体动作识别准确率越高。例如，Incpetion 比 ResNet 更优，但 VGG 和 Incpetion 旗鼓相当。因为，较深或较复杂的网络结构能够从视频数据中提取到更高质量的视觉特征，但随着关联方法设计的精细化和复杂化，视觉特征质量给群体动作识别带来的增益达到了瓶颈。例如 Yuan 等人[106]使用简单的 ResNet-18 网络也能在三个基准数据集

表 5 不同动作特征编码器对主流方法的性能影响（仅使用 RGB 模态数据，IN 代表 Inception，RN 代表 ResNet）

方法	年份	特征类型	性能			
			VD[16]		CAD-v1[15]	
			Acc.	Mean Acc.	Acc.	Mean Acc.
Wu 等人[104][122]	2019	VGG-16	91.9	-	90.1	-
		VGG-19	**92.6**	-	-	-
		IN-v3	92.5	-	**91.0**	-
Yan 等人[17]	2020	RN-18	93.1	-	-	-
		IN-v3	**94.0**	-	-	-
Shu 等人[47]	2020	RN-50	-	92.0	-	93.7
		VGG-16	-	91.1	-	92.5
		IN-v3	-	**93.0**	-	**94.9**
Yan 等人[55]	2020	AlexNet	88.6	89.4	92.5	92.3
		RN-18	**91.4**	**92.0**	**93.4**	**93.0**
Yuan 等人[87]	2021	VGG-16	**94.1**	**94.4**	-	**95.4**
		IN-v3	93.3	93.4	-	95.1
Yuan 等人[106]	2021	RN-18	93.1	93.3	-	95.3
		VGG-16	**93.6**	**93.8**	-	**95.9**
Han 等人[115]	2022	RN-18	-	-	-	96.0
		IN-v3	-	-	-	**96.5**
Li 等人[120]	2022	RN-18	93.2	93.7	95.7	**95.3**
		IN-v3	**93.5**	**93.9**	**96.5**	**95.3**

表 6 不同数据模态对主流方法的性能影响

方法	年份	数据模态	性能					
			VD[16]		CAD-v1[15]		NBA[17]	
			Acc.	Mean Acc.	Acc.	Mean Acc.	Acc.	Mean Acc.
Azar 等人[96]	2018	RGB	84.7	-	-	-	-	-
		+ Optical Flow	88.6	-	-	-	-	-
		+ Warped Optical Flow	88.9	-	-	-	-	-
		+ Pose	**90.4**	-	-	-	-	-
Tang 等人[60]	2018	RGB	89.3	89.0	-	92.5	-	-
		RGB + Optical Flow	**90.7**	**90.0**	-	**95.7**	-	-
Yan 等人[42]	2018	RGB	86.2	86.1	-	91.2	-	-
		RGB + Optical Flow	**87.7**	**88.1**	-	**92.2**	-	-
Lu 等人[95]	2018	RGB	90.6	-	-	-	-	-
		RGB + Pose	**91.2**	-	-	-	-	-
Li 等人[57]	2018	RGB	38.7	-	83.7	-	-	-
		Optical Flow	54.3	-	70.1	-	-	-
		RGB + Optical Flow	**66.9**	-	**86.1**	-	-	-
Azar 等人[59]	2019	RGB	92.1	-	83.4	-	-	-
		Optical Flow	91.5	-	85.4	-	-	-
		RGB + Optical Flow	**93.0**	-	**85.8**	-	-	-
Gavrilyuk 等人[61]	2020	RGB + Optical Flow	93.0	-	**92.8**	-	-	-
		RGB + Pose	93.4	-	91.0	-	-	-
		Pose + Optical Flow	**94.4**	-	91.2	-	-	-
Xu 等人[99]	2020	RGB	92.8	-	-	-	-	-
		Optical Flow	91.9	-	-	-	-	-
		RGB + Optical Flow	**93.5**	-	-	-	-	-
Yuan 等人[87]	2021	RGB	94.1	94.4	-	95.4	-	-
		Pose + Optical Flow	92.9	93.2	-	94.9	-	-
		RGB+ Pose + Optical Flow	**94.7**	**95.0**	-	**96.4**	-	-
Li 等人[50]	2021	RGB	94.9	-	94.7	-	-	-
		RGB + Pose	**95.7**	-	**96.3**	-	-	-
Pramono 等人[113]	2021	RGB	94.8	-	94.4	-	-	-
		RGB + Optical Flow	95.3	-	95.1	-	-	-
		RGB + Optical Flow + Pose	**96.4**	-	**96.0**	-	-	-
Han 等人[115]	2022	RGB	94.4	-	-	-	51.5	44.8
		Optical Flow	-	-	-	-	56.8	49.1
		RGB + Optical Flow	**95.4**	-	-	-	**58.1**	**50.2**
Liu 等人[133]	2022	RGB	91.9	-	90.5	-	-	-
		Optical Flow	90.4	-	89.7	-	-	-
		RGB + Optical Flow	93.3	-	92.3	-	-	-
		RGB + Pose + Optical Flow	**94.4**	-	**93.2**	-	-	-

上取得不错的结果。这启发学者们未来或许可以统一视觉特征提取方式，聚焦群体动作任务本身，设

计更高质量的群体动作分析建模方案。

综上所述，更深更复杂的网络结构和更多样的模态信息，能够为群体动作理解带来明显增益。但在未来如何统一且高效地从视频数据中提取高质量个体动作特征仍亟待解决。

6.2.3 不同数据模态

从视频数据中构建 RGB、光流和人物姿态等信息用以群体动作的识别，有助于从多个视角互补地理解群体动作。本文在表 6，汇总了现有文献所涉及的所有基于多模态特征，在三个主流数据集上的对比结果，以探究不同数据模态对群体动作识别的影响。需要说明的是，Azar 等人[96]在文中所汇报的结果是逐行模态特征叠加使用，而非单独增加。此外，NBA[17]数据集较新，大部分文献并未在该数据集上进行多模态数据的对比分析。

光流（Optical Flow）数据能够有效刻画出相邻视频帧中的物体运动趋势，已经成为大多数多模态解决方案的首选。从表 6 中可以发现，不少方法仅使用光流数据模态依然能够保持不错的性能。在使用相同的特征编码器的情况下，如果将其与 RGB 数据同时使用，能够带来 1%~4%明显的增益。Azar 等人[96]额外使用了一种增强的光流数据（Warped Optical Flow），但与使用普通光流数据相比，在群体动作识别结果上并无明显差异。

姿态（Pose）数据能够刻画出运动人物肢体结构的时空变化，也被不少群体动作识别方法用于进一步增强个体动作特征的判别性。Lu 等人[95]和 Li 等人[50]都验证了姿态数据可以作为 RGB 数据的补充，其可以显著地提升群体动作识别的准确率。此外，Gavrilyuk 等人[68]所得的实验结果表明，姿态和光流特征所带来的增益旗鼓相当，但是姿态数据维度低且特征编码成本低。此外，从 Gavrilyuk 等人[68]所得的实验结果中发现，使用姿态和光流特征能够获得最佳性能，启发未来研究可以尝试丢弃高度冗余的 RGB 视觉数据，转而使用更为高效的姿态数据。Yuan 等人[87]，Pramono 等人[113]和 Liu 等人[133]获得的实验结果表明，同时使用 RGB、光流和姿态数据能够获得最佳的性能。

总体而言，额外的模态信息能够在一定程度上缓解深度模型在有限数据样本上的过拟合问题。该现象在数据样本规模较小的 CAD-v1 数据集上呈现的更为明显。各类模态信息都给群体动作识别带来一定的增益，但其获取成本较高，因此对实际应用不友好。此外，目前对额外模态信息的使用和建模没有一个较为统一的准则，导致各种方法的实验对比不够公平。

6.2.4 弱标注设置

弱标注设置下的群体动作识别根据符合实际应用需求，近年来相关解决方案层出不穷。为此，本文在表 7 统计了所有弱标注解决方案在 NBA 和 VD 数据集上的实验结果。其中部分方法[104] [61] [62] [106] [115]的结果由 Kim 等人[122]基于 RGB 数据复现，且依然使用了由 Yan 等人[17]提供的人物检测框作为监督信号且不进行任何筛选。因此，其“自适应筛

表 7 弱标注设置下现有方法在基准数据集 NBA[17]和 VD[16]上仅使用 RGB 模态数据的的表现汇总（RN 代表 ResNet，①代表卷积特征、②代表卷积-序列特征、③代表自注意力特征）

方法	年份	特征类型	关联建模	自适应筛选机制	（弱标注）性能		
					NBA		VD
					Acc.	Mean Acc.	Acc.
Wu 等人[104]	2019	①(RN-18)	图	-	59.0	56.8	87.4
Yan 等人[17]	2020	①(RN-18)	图	基于特征相似度选择	49.1	47.5	86.3
Gavrilyuk 等人[61]	2020	①(RN-18)	图	-	47.1	41.5	84.3
Pramono 等人[62]	2020	①(RN-18)	图	-	56.3	52.8	83.3
Yuan 等人[106]	2021	①(RN-18)	图	-	61.6	56.0	86.5
Han 等人[115]	2022	①(IN-v3)	图	-	51.5	44.8	-
Kim 等人[122]	2022	①(RN-18)	图	基于注意力的原型学习	75.8	71.2	90.5
Wu 等人[121]	2023	①(IN-v3)	图	自注意力掩码	49.4	-	90.2
Chappa 等人[123]	2023	③(ViT-16)	图	自监督训练	**82.1**	**72.8**	**92.9**

选机制”字段为空。

相比于 Yan 等人[17]提供的基准结果，得益于较深的图关联推理结构，部分传统方法[104] [62] [106]在不进行人物框筛选的情况下，依然能获得不错的性能。此外，Wu 等人[121]直接构建可学习的掩码从全图选取关键信息，但效果不佳。这表明基于空间的信息选择对弱标注群体动作识别而言并非必要，甚至可能会适得其反。

Kim 等人[122]通过构建少量“原型（Prototype）”表示自适应地汇聚关键时空线索，在两个弱标注数据集上都取得了显著的性能提升。此外，Chappa 等人[123]等人通过引入自监督训练策略，增强模型运动感知能力，也获得了明显的性能增益。值得注意的是，这两种方法都使用到了较深的自注意力网络（Transformer）对场景中所有视觉信息进行编码或关联建模。这表明通过强大的 Transformer 从全图信息中进行隐式的自适应选择关键时空线索，能够基本摆脱群体行为识别算法对细粒度人物标注的依赖。其在一定程度上确立了未来几年的群体动作识别研究的主流方向。

7 未来展望

随着计算机视觉和深度学习技术的飞速发展，群体动作分析技术已经成为备受关注的研究领域。然而，该技术的落地应用仍面临许多限制，例如可支持场景单一、监督成本高、模型推理和特征表示效率低等。因此，未来的研究可集中于解决上述挑战，以便推进群体动作识别技术的落地。本文深刻地探讨了上述挑战，并对未来研究趋势进行了分析和展望。

7.1 多场景通用的群体动作识别

目前，已经有许多数据集可用于群体动作分析算法的训练和测试。然而，现有数据集通常是从某个特定场景（如某个体育运动场景或某个室内场景）中采集而得，这极大程度地限制了群体动作的多样性和复杂性，从而给现有算法在落地应用方面带来诸多困难与挑战。因此，我们需要设计一个能够应用在多种场景下的识别模型。该模型应该具有足够的泛化能力，能够识别和理解不同场景中的群体动作。然而，实现该目标具有一定难度，因为不同场景中的群体动作具有很大的差异性，甚至同样的动作在不同场景中具有不同的含义。

为了解决该问题，重新定义群体动作的概念迫在眉睫。目前，群体动作通常被定义为某个特定场景下的一组动作或行为，例如篮球比赛中的投篮、传球和防守等。然而，这种定义方式限制了群体动作的适用范围和泛化能力。从更实际的角度出发，我们应该设计一个更加通用和抽象的群体动作定义。例如，将传球、接球和扣球等球类比赛中的动作抽象为“传递”这一通用动作。这样，我们就可以将群体动作定义为来自不同场景下的一组通用动作或行为，而不是针对特定场景的动作。这种通用的群体动作定义不仅可以提高算法的泛化能力，还可以减少数据集采集和标注的成本，从而推动群体动作分析技术的发展。

综上所述，设计一个能够应用于多种场景的群体动作识别模型是一个具有挑战性的任务。要实现该目标，我们需要重新定义群体动作的概念，并通过多领域的合作来推动技术的发展。

7.2 弱标注群体动作识别

群体动作识别涉及到对一个场景中多个人物的协作行为进行理解和识别，而且个体动作之间的联系非常复杂。因此，现有方法通常需要使用细粒度的标注信息，如人物位置和人物动作等，来帮助模型理解最终的群体动作。然而，该学习范式所用到的细粒度标注需要大量的人力和物力，不利于构建更大规模的数据集。

幸运的是，近年来有一些工作开始探索弱标注群体动作识别。弱标注是指仅给出一个视频级别的群体动作类别标签，而丢弃所有其他细粒度的标注信息。该标注方式有利于轻松构建更大规模数据集，但同时也带来了新的挑战。弱标注使得模型在个体动作间进行关联建模时难以聚焦到有效信息。因此，如何自适应地挖掘有效视觉信息以确保群体动作识别的准确性成为一个难点。

近年来，研究者们提出了许多方法来解决上述问题。其中一种方法是利用现成的目标检测和跟踪算法从视频中获取人物运动轨迹信息，然后基于此提取个体人物运动特征并进行建模以识别群体动作。但这类方法严重依赖目标检测和跟踪算法的质量且不够灵活。另一种方法是基于自注意力机制直接以可学习的方式加权汇聚所有视觉区域块的信息。但这种方法忽略了视觉场景内存在的高度时空冗余，在无效视觉特征编码上浪费了大量的算力。

综上所述，弱标注群体动作识别是一个更具应

用价值且更具挑战的研究方向。随着深度学习技术的不断发展和数据集的不断壮大，该问题将会有更好的解决方法，这个研究方向也将有更广阔的应用前景。

7.3 面向模型或表示轻量化的群体动作识别

现有群体动作识别研究中所用到的视觉模型结构庞大且特征表示维度高，不利于算法的落地应用。因此，未来研究可以聚焦于模型轻量化和表示轻量化两个方面。

在模型轻量化方面，可以考虑通过降低模型的复杂度和计算量来实现。基于对现有研究的分析，我们发现最近提出的一些深层视觉特征提取器并不能给群体动作识别带来明显增益。因此，可以尝试采用一些浅层模型，或者可以考虑采用一些模型轻量优化方法，如剪枝、量化和蒸馏等，以进一步减小模型在对场景内若干人物进行建模时的计算量。

在表示轻量化方面，可以利用其他更为高效的信息（例如人物姿态信息、几何位置信息、深度信息、红外信息、声音信息等）替代 RGB 视觉信号或作为补充，以降低特征提取成本并提高模型的鲁棒性。这些模态的信息相比于原始视觉信号更为稀疏，但也能带来不少增益。另外，在特征提取过程中还可以采用一些优化策略，如快速卷积和通道注意力机制，以加快特征提取速度并减小特征维度。

总之，未来研究中的模型轻量化和表示轻量化是群体动作分析技术落地应用的关键。通过采用简单而有效的模型和轻量级的优化方法，以及利用多模态信息和轻量级的特征提取方法，可以实现更加高效且实用的群体动作分析技术。

7.4 小节

针对上述三个方面的研究局限性，本文认为未来研究应该关注以下方向：一是研究适用于各种类型场景的通用群体动作识别算法；二是研究如何提升算法在无细粒度监督信息的情况下的准确性；三是研究如何设计更加高效的模型推理结构和提取更加紧凑的特征表示，以适应实时应用的需求。上述方向的研究将有助于提高群体动作识别技术在实际应用中的可用性，为相关领域的长远发展做出更大贡献。

8 总结

本文介绍了群体动作识别的任务定义，并调研了其近十年的研究进展。首先，本文认为群体动作识别基本遵循：1）个体动作特征提取，2）个体动作特征关联建模，3）个体特征聚合，4）群体动作预测。其中，本文将最为核心的关联建模主要分为线性关联，序列关联和图关联。并以此为依据，对现有群体动作识别的研究文献进行归纳、分析和总结。此外，本文还总结了群体动作识别研究中常用的三类深度特征表示和十二种可用数据集。最后，本文在三个常用数据集上对比分析了现有方法的性能表现，并对群体动作识别的未来研究方向给出了研判。希望本综述能够帮助读者快速了解群体动作研究的基本概况、核心思想和未来研究趋势。

参 考 文 献

[1] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos//Proceedings of the Annual Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 568-576

[2] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, United States, 2015: 2625-2634

[3] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding// Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea, 2019: 7083-7093

[4] Zhou B, Andonian A, Oliva A, et al. Temporal relational reasoning in videos// Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 803-818

[5] Ma C Y, Chen M H, Kira Z, et al. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. Signal Proceedings: Image Communication, 2019, 71: 76-87

[6] Li Y, Chen L, He R, et al. Multisports: A multi-person video dataset of spatio-temporally localized sports actions//Proceedings of the IEEE International Conference on Computer Vision. Montreal, Canada, 2021: 13536-13545

[7] Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 2556-2563

[8] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012

[9] Goyal R, Ebrahimi Kahou S, Michalski V, et al. The" something something" video database for learning and evaluating visual common sense//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 5842-5850

[10] Sigurdsson G A, Varol G, Wang X, et al. Hollywood in homes:

Crowdsourcing data collection for activity understanding//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 510-526

[11] Dollár P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features//Proceedings of the IEEE International Conference on Computer Vision International Workshop. Beijing, China, 2005: 65-72

[12] Amer M R, Todorovic S, Fern A, et al. Monte carlo tree search for scheduling activity recognition//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 1353-1360

[13] Wu Yun-Peng, Zhao Chen-Yang, Shi Zeng-Lin, et al. A flow density based algorithm for detecting coherent motion with multiple interaction. Chinese Journal of Computers, 2017, 40 (11): 2519-2532 (in Chinese)
(吴云鹏, 赵晨阳, 时增林,等. 基于流密度的多重交互集体行为识别算法. 计算机学报, 2017, 40 (11): 2519-2532)

[14] 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, 2019, 43(3): 1110-1118

[15] Choi W, Shahid K, Savarese S. What are they doing?: Collective activity classification using spatio-temporal relationship among people//Proceedings of the IEEE International Conference on Computer Vision Workshops. Kyoto, Japan, 2009: 1282-1289

[16] Ibrahim M S, Muralidharan S, Deng Z, et al. A hierarchical deep temporal model for group activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, United States, 2016: 1971-1980

[17] 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, United Kingdom, 2020: 208-224

[18] Tsai T Y, Lin Y Y, Jeng S K, et al. End-to-end key-player-based group activity recognition network applied to basketball offensive tactic identification in limited data scenarios. IEEE Access, 2021, 9: 104395-104404

[19] Nakatani C, Sendo K, Ukita N. Group activity recognition using joint learning of individual action recognition and people grouping//Proceedings of the International Conference on Machine Vision and Applications. Nagoya, Japan, 2021: 1-5

[20] Wu L, Wang Q, Li Z, et al. Relation-guided actor attention for group activity recognition//Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision. Beijing, China, 2021: 129-141

[21] Thilakarathne H, Nibali A, He Z, et al. Pose is all you need: The pose only group activity recognition system (POGARS). arXiv preprint arXiv:2108.04186, 2021

[22] 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

[23] Wu L, He J, Jian M, et al. Global motion pattern based event recognition in multi-person videos//Proceedings of the CCF Chinese Conference on Computer Vision. Tianjin, China, 2017: 667-676

[24] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, United States, 2018: 6450-6459

[25] Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, United States, 2015: 4305-4314

[26] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 20-36

[27] Amer M R, Lei P, Todorovic S. Hirf: Hierarchical random field for collective activity recognition in videos//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 572-585

[28] Xu Q, See J, Lin W. Localization guided fight action detection in surveillance videos//Proceedings of the IEEE International Conference on Multimedia and Expo. Shanghai, China, 2019: 568-573

[29] Zhao Yao, Tian Yonghong, Dang Jianwu, et al. Frontiers of transportation video structural analysis in the smart city. Journal of Image and Graphics. 2021, 26(06):1227-1253 (in Chinese)
(赵耀, 田永鸿, 党建武, 等.面向智慧城市的交通视频结构化分析前沿进展.中国图象图形学报, 2021, 26 (06):1227-1253)

[30] Yu Chen, Zhang Lijuan, Jin Hai. Research progress and trend of big data-driven intelligent transportation system. Chinese Journal on Internet of Things, 2018, 2 (1): 56-63 (in Chinese)
(余辰,张丽娟,金海.大数据驱动的智能交通系统研究进展与趋势. 物联网学报, 2018, 2 (1): 56-63)

[31] Yu H, Cheng S, Ni B, et al. Fine-grained video captioning for sports narrative//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, United States, 2018: 6006-6015

[32] Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(4): 509-522

[33] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, United States, 2016: 770-778

[34] Song SJ, Liu JY, Li YH, et al. Modality compensation based action recognition. Journal of Software, 2018, 29 (2): 1-15 (in Chinese)
(宋思捷, 刘家瑛, 厉扬豪, 等. 关联模态补偿的视频动作识别算法.软件学报, 2018, 29 (2): 1-15)

[35] Xie Zhao, Zhou Yi, Wu Ke-Wei, et al. Activity recognition based on spatial-temporal attention LSTM. Chinese Journal of Computers, 2021, 44 (2):261-274. (in Chinese)
(谢昭, 周义, 吴克伟,等. 基于时空关注度 LSTM 的行为识别. 计算机学报, 2021, 44 (2): 261-274)

[36] Ding Chong-Yang, Liu Kai, Li Guang, et al. Spatio-temporal weighted posture motion features for human skeleton action recognition research. Chinese Journal of Computers, 2022, 43 (1): 29-40 (in Chinese)
(丁重阳, 刘凯, 李光,等. 基于时空权重姿态运动特征的人体骨架行为识别研究. 计算机学报, 2020, 43 (1): 29-40)

[37] Wu, Li-Fang and Wang, Qi and Jian, Meng and Qiao, Yu and Zhao, Bo-Xuan. A comprehensive review of group activity recognition in videos. International Journal of Automation and Computing, 2021, 18(3), 334-350

[38] Pei Lishen, Zhao Xuezhuan. A survey of collective activity analysis and recognition based on deep learning. Journal of Frontiers of Computer

Science and Technology, 2021: 1-17 (in Chinese)
(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, 2021:1-17)
[39] Borja-Borja L F, Azorin-Lopez J, Saval-Calvo M, et al. Deep learning architecture for group activity recognition using description of local motions//Proceedings of the International Joint Conference on Neural Networks. Glasgow, United Kingdom, 2020: 1-8
[40] Perez M, Liu J, Kot A C. Skeleton-based relational reasoning for group activity analysis. Pattern Recognition, 2022, 122: 108360
[41] Danelljan M, Häger G, Khan F, et al. Accurate scale estimation for robust visual tracking//Proceedings of the British Machine Vision Conference. Nottingham, United Kingdom, 2014:1-11
[42] 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, 2018: 1292-1300
[43] Wang M, Ni B, Yang X. Recurrent modeling of interaction context for collective activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, United States, 2017: 3048-3056
[44] 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 Conference on Computer Vision and Pattern Recognition. Honolulu, United States, 2017: 4315-4324
[45] 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(02): 636-647
[46] Shu T, Todorovic S, Zhu S C. CERN: confidence-energy recurrent network for group activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, United States, 2017: 5523-5531
[47] 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, 2020, 32(2): 663-674
[48] Ibrahim M S, Mori G. Hierarchical relational networks for group activity recognition and retrieval//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 721-736
[49] Hu G, Cui B, He Y, et al. Progressive relation learning for group activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, United States, 2020: 980-989
[50] 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, Canada, 2021: 13668-13677
[51] Girshick R. Fast r-cnn//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1440-1448
[52] He K, Gkioxari G, Dollár P, et al. Mask r-cnn//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2961-2969
[53] Li Y, Tarlow D, Brockschmidt M, et al. Gated graph sequence neural networks//Proceedings of the International Conference on Learning Representations, San Diego, United States, 2015: 1-20
[54] Xu Kelvin, Ba Jimmy, Kiros Ryan et al. Show, attend and tell: Neural image caption generation with visual attention//Proceedings of the International Conference on Machine Learning, Lille, France, 2015: 2048-2057
[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, 2023, 45(6): 6955-6968
[56] Zhang P, Tang Y, Hu J F, et al. Fast collective activity recognition under weak supervision. IEEE Transactions on Image Processing, 2019, 29: 29-43
[57] Li X, Choo Chuah M. Sbgar: Semantics based group activity recognition//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2876-2885
[58] Li X, Chuah M C. Rehar: Robust and efficient human activity recognition//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, United States, 2018: 362-371
[59] Azar S M, Atigh M G, Nickabadi A, et al. Convolutional relational machine for group activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, United States, 2019: 7892-7901
[60] Tang Y, Wang Z, Li P, et al. Mining semantics-preserving attention for group activity recognition//Proceedings of the ACM International Conference on Multimedia. Seoul, South Korea, 2018: 1283-1291
[61] Gavrilyuk K, Sanford R, Javan M, et al. Actor-transformers for group activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, United States, 2020: 839-848
[62] 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, United Kingdom, 2020: 71-90
[63] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, United States, 2019: 5693-5703
[64] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90
[65] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324
[66] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495
[67] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014
[68] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, United States, 2015: 1-9.
[69] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231
[70] Feichtenhofer C. X3d: Expanding architectures for efficient video recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, United States, 2020: 203-213
[71] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features

with 3d convolutional networks//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4489-4497

[72] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, United States, 2017: 6299-6308

[73] Xie S, Sun C, Huang J, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification//Proceedings of the European Conference on Computer vision. Munich, Germany, 2018: 305-321

[74] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks//Proceedings of the Annual Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 3104-3112

[75] Jozefowicz R, Vinyals O, Schuster M, et al. Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410, 2016

[76] Zia T, Zahid U. Long short-term memory recurrent neural network architectures for Urdu acoustic modeling. International Journal of Speech Technology, 2019, 22(1): 21-30

[77] Li X, Wu X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Brisbane, Australia, 2015: 4520-4524

[78] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[79] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014

[80] Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM//Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc, Czech Republic, 2013: 273-278

[81] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 2005, 18(5-6): 602-610

[82] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Annual Conference on Neural Information Processing Systems, Long Beach, United States, 2017: 5998-6008

[83] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale//Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 2020: 1-21

[84] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding//Proceedings of the International Conference on Machine Learning, Online, 2021: 813-824

[85] Fan H, Xiong B, Mangalam K, et al. Multiscale vision transformers //Proceedings of the IEEE International Conference on Computer Vision. Montreal, Canada, 2021: 6824-6835

[86] Patrick M, Campbell D, Asano Y, et al. Keeping your eye on the ball: Trajectory attention in video transformers//Proceedings of the Annual Conference on Neural Information Processing Systems. Granada, Spain, 2021: 12493-12506

[87] Yuan H, Ni D. Learning visual context for group activity recognition //Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021, 35(4): 3261-3269

[88] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015:1412-1421

[89] 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, 2022, 33(12): 7574-7588

[90] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. Nature, 1986, 323(6088): 533-536

[91] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016

[92] Gao J, Zhang T, Xu C. Graph convolutional tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, United States, 2019: 4649-4659

[93] Wang X, Girshick R, Gupta A, et al. Non-local neural networks //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, United States, 2018: 7794-7803

[94] Kong L, Qin J, Huang D, et al. Hierarchical attention and context modeling for group activity recognition//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, Canada, 2018: 1328-1332

[95] 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

[96] Azar S M, Atigh M G, Nickabadi A. A multi-stream convolutional neural network framework for group activity recognition. arXiv preprint arXiv:1812.10328, 2018

[97] 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, 2018: 101-117

[98] 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, 2019, 30(2): 549-565

[99] Xu D, Fu H, Wu L, et al. Group activity recognition by using effective multiple modality relation representation with temporal-spatial attention. IEEE Access, 2020, 8: 65689-65698

[100] Raptis M, Kokkinos I, Soatto S. Discovering discriminative action parts from mid-level video representations//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, United States, 2012: 1242-1249

[101] Gammulle H, Denman S, Sridharan S, et al. Multi-level sequence GAN for group activity recognition//Proceedings of the Asian Conference on Computer Vision. Springer, Cham, Perth, Australia, 2018: 331-346

[102] Biswas S, Gall J. Structural recurrent neural network (srnn) for group activity analysis//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, United States, 2018: 1625-1632

[103] Jain A, Zamir A R, Savarese S, et al. Structural-rnn: deep learning on spatio-temporal graphs//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, United States, 2016: 5308-5317

[104] Wu J, Wang L, Wang L, et al. Learning actor relation graphs for group activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, United States, 2019: 9964-9974

[105] Mao K, Jin P, Ping Y, et al. Modeling multi-scale sub-group context for group activity recognition. Applied Intelligence, 2023, 53(1): 1149-1161
[106] 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, Canada, 2021: 7476-7485
[107] Duan Y, Wang J. Learning key actors and their interactions for group activity recognition//Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision. Zhuhai, China, 2021: 53-65
[108] Sanchez-Gonzalez A, Heess N, Springenberg J T, et al. Graph networks as learnable physics engines for inference and control//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 4470-4479
[109] Lu L, Lu Y, Wang S. Learning multi-level interaction relations and feature representations for group activity recognition//Proceedings of the International Conference on Multimedia Modeling. Prague, Czech Republic, 2021: 617-628
[110] Pei D, Li A, Wang Y. Group activity recognition by exploiting position distribution and appearance relation//Proceedings of the International Conference on Multimedia Modeling. Prague, Czech Republic, 2021: 123-135
[111] Feng Y, Shan S, Liu Y, et al. Drgcn: deep relation gcn for group activity recognition//Proceedings of the International Conference on Neural Information Processing. Vancouver, Canada, 2020: 361-368
[112] Li G, Muller M, Thabet A, et al. Deepgcns: can gcns go as deep as cnns//Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea, 2019: 9267-9276
[113] 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
[114] Zhu X, Zhou Y, Wang D, et al. Mlst-former: multi-level spatial-temporal transformer for group activity recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(7): 3383-3397
[115] Han M, Zhang D J, Wang Y, et al. Dual-ai: dual-path actor interaction learning for group activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, United States, 2022: 2990-2999
[116] Du Z, Wang X, Wang Q. Self-supervised global spatio-temporal interaction pre-training for group activity recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9): 5076-5088
[117] Deng Z, Zhai M, Chen L, et al. Deep structured models for group activity recognition//Proceedings of the British Machine Vision Conference. Swansea, United Kingdom, 2015: 1971-1980
[118] Lan T, Wang Y, Yang W, et al. Beyond actions: discriminative models for contextual group activities//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2010: 1216-1224
[119] 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 Conference on Computer Vision and Pattern Recognition. Las Vegas, United States, 2016: 4772-4781
[120] Li W, Yang T, Wu X, et al. Learning action-guided spatio-temporal transformer for group activity recognition//Proceedings of the ACM International Conference on Multimedia. Lisbon, Portugal, 2022: 2051-2060
[121] Wu L, Lang X, Xiang Y, et al. Active spatial positions based hierarchical relation inference for group activity recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(6): 2839-2851
[122] Kim D, Lee J, Cho M, et al. Detector-free weakly supervised group activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, United States, 2022: 20083-20093
[123] Chappa N, Nguyen P, Nelson A, et al. Spartan: self-supervised spatiotemporal transformers approach to group activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Workshop). Vancouver, Canada, 2023: 5157-5167
[124] Ni B, Yan S, Kassim A. Recognizing human group activities with localized causalities//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, United States, 2009: 1470-1477
[125] Blunsden S, Fisher R B. The BEHAVE video dataset: ground truthed video for multi-person behavior classification//Proceedings of the Annals of the British Machine Vision Association. Aberystwyth, United Kingdom, 2010, 4: 1-11
[126] Choi W, Shahid K, Savarese S. Learning context for collective activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, United States, 2011: 3273-3280
[127] 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, 2012: 215-230
[128] 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, 2012: 187-200
[129] Lan T, Wang Y, Yang W, et al. Discriminative latent models for recognizing contextual group activities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(8): 1549-1562
[130] Lan T, Sigal L, Mori G. Social roles in hierarchical models for human activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, United States, 2012: 1354-1361
[131] Chen J, Hao H, Hong H, et al. Rit-18: A novel dataset for compositional group activity understanding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Seattle, United States, 2020: 362-363
[132] Zalluhoglu C, Ikizler-Cinbis N. Collective Sports: A multi-task dataset for collective activity recognition. Image and Vision Computing, 2020, 94: 103870
[133] 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

Yan Rui, Ph.D. His main research interests include computer vision and multimedia analysis.

Ge Xiaojing, Master. Her main research interests include computer vision and machine learning.

Huang Peng, Ph.D. His main research interests include computer vision and multimedia analysis.

Shu Xiangbo,Ph.D.，professor. His main research interests include computer vision and machine learning.

Tang Jinhui, Ph.D., professor. His main research interests include multimedia analysis and computer vision.

Background

Group activity recognition aims to understand the complex activity composed of multiple persons and their interactions in the scene. Different from traditional simple action recognition (performed by a single person), group activity involves more complex spatio-temporal dependency among persons in the scene and more noisy visual information. Recently, the widespread application of group activity recognition in public safety monitoring, sports video analysis, and social role understanding has attracted the attention of researchers. However, there are few Chinese literatures that can help researchers quickly understand the research overview, and the basis for induction and analysis is vague.

To this end, this paper aims to review the research progress of group activity recognition based on deep learning in recent years. We divide the whole complex recognition system into three parts including person-level feature extraction, feature interaction, and feature fusion. We mainly review and summarize person-level feature extraction and feature interaction. After that, we also introduce more than ten benchmarks that can be used for group activity recognition, but only compare the existing methods on Volleyball Dataset, Collective Activity Dataset and NBA Dataset, because most benchmarks are not used in recent works. Finally, this paper anticipates several practical and more challenging future research directions. We hope this article can help readers understand the research overview of group activity recognition, its core research ideas, and future research trends.

This work is supported by the Postdoctoral Fellowship Program of CPSF (No. GZB20230302), the Jiangsu Funding Program for Excellent Postdoctoral Talent (No. 2023ZB256), the National Natural Science Foundation of China (No. 62302208, No. 61925204, No. 62222207, and No. 62072245), and the Natural Science Foundation of Jiangsu Province (No. BK20211520).