软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(2):964−984 [doi: 10.13328/j.cnki.jos.006693]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

视频群体行为识别综述*

吴建超, 王利民, 武港山

(计算机软件新技术国家重点实验室(南京大学), 江苏 南京 210023)

通信作者: 王利民, E-mail: lmwang@nju.edu.cn

摘 要: 群体行为识别是指给定一个包含多人场景的视频,模型需要识别出视频中多个人物正在共同完成的群体行为. 群体行为识别是视频理解中的一个重要问题, 可以被应用在运动比赛视频分析、监控视频识别、社交行为理解等现实场景中. 多人场景视频较为复杂, 时间和空间上的信息十分丰富, 对模型提取关键信息的能力要求更高. 模型只有高效地建模场景中的层次化关系, 并为人物群体提取有区分性的时空特征, 才能准确地识别出群体行为. 由于其广泛的应用需求, 群体行为识别问题受到了研究人员的广泛关注. 对近几年来群体行为识别问题上的大量研究工作进行了深入分析, 总结出了群体行为识别研究所面临的主要挑战, 系统地归纳出了 6 种类型的群体行为识别方法, 包含传统非深度学习识别方法以及基于深度学习技术的识别方法, 并对未来研究的可能方向进行了展望.

关键词: 群体行为识别; 计算机视觉; 视频理解; 行为识别

中图法分类号: TP391

中文引用格式: 吴建超, 王利民, 武港山. 视频群体行为识别综述. 软件学报, 2023, 34(2): 964–984. http://www.jos.org.cn/1000-9825/6693.htm

英文引用格式: Wu JC, Wang LM, Wu GS. Group Activity Recognition in Videos: A Survey. Ruan Jian Xue Bao/Journal of Software, 2023, 34(2): 964–984 (in Chinese). http://www.jos.org.cn/1000-9825/6693.htm

Group Activity Recognition in Videos: A Survey

WU Jian-Chao, WANG Li-Min, WU Gang-Shan

(State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

Abstract: Given a video containing a multi-person scene, group activity recognition model needs to recognize the group activity that multiple people in video are completing together. Group activity recognition is an important problem in video understanding and can be applied to sports videos analysis, surveillance video recognition, social behavior understanding, and other real scenarios. Multi-person scene video is complicated, and the spatial-temporal information is rich, which requires the model to extract key information. To accurately recognize group activity, the model should efficiently model the hierarchical relationships in the scene and extract distinguishing spatial-temporal features for people. Due to its wide range of application requirements, the problem of group activity recognition has received extensive attention from researchers. This study has conducted an in-depth analysis of a large number of research work on group activity recognition in recent years, and summarized the main challenges of group activity recognition research, systematically summarized six types of group activity recognition methods, including traditional non-deep learning recognition methods and recognition methods based on deep learning technology, and proposed the possible directions of future research.

Key words: group activity recognition; computer vision; video understanding; activity recognition

1 简 介

1.1 研究背景

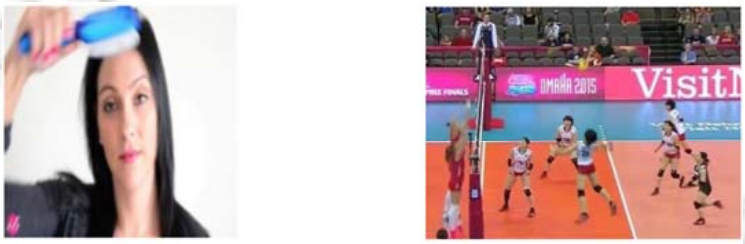
随着计算机技术的快速发展, 各式各样的电子设备, 包括监控摄像头、手机、电脑等, 每天都会产生海量

* 基金项目: 国家自然科学基金(62076119, 61921006)
收稿时间: 2021-05-31; 修改时间: 2021-09-28, 2022-02-17; 采用时间: 2022-04-20

的图像数据. 这些海量图像数据中包含着丰富的有价值的信息, 然而仅仅依靠人力难以完成对所有图像数据的分析和处理. 作为人工智能研究中的一个重要分支, 计算机视觉的主要研究目标就是让计算机模型能够从原始图像数据中提取出有用的、结构化的高层语义信息[1−4]. 在一个完整的人工智能系统中, 计算机视觉模型主要充当了感知器的决策, 可以帮助人工智能系统获取图像中的视觉语义信息, 对当前环境进行感知, 辅助下游的决策模型进行决策.

早期的计算机视觉研究集中在图片理解问题上, 即模型的输入是单张图片. 然而, 由于单张图片所能容纳的信息有限, 人们更加倾向于选择视频来作为记录的信息载体. 随着越来越多视频数据的产生, 现实生活中也出现了许多需要视频理解技术的应用场景. 比如: 在安防监控场景下, 需要视频理解技术来自动识别监控视频画面中的异常行为, 实时发出预警; 在互联网视频场景下, 需要视频理解技术自动地识别海量视频大数据的内容, 提取结构化信息, 提供给安全审查、内容检索、视频推荐等下游模型使用. 近些年来, 部分计算机视觉研究工作也开始关注以视频数据作为输入的视觉理解问题, 视频理解技术也逐渐成为研究热点和发展方向[5−10]. 从技术角度看, 相比于图片数据, 视频数据增加了一个时序维度, 需要模型具有时序建模能力, 能够融合多帧信息, 建模时序动态变化. 此外, 视频中存在许多无关冗余信息, 存在运动模糊问题, 视频处理的计算开销也更大, 处理并且理解视频数据更具挑战性.

本文主要关注视频中的群体行为识别问题. 行为识别问题是指给定一个视频片段, 需要模型识别出视频中人物正在进行的行为动作, 比如吃饭、读书、跑步等. 早期的视频行为识别研究局限在单人场景下[4,12−17], 即视频中仅包含一个人物. 如图 1(a)所示: 单人行为识别模型接受一个视频片段作为输入, 输出一个动作类别, 表示视频主体任务正在进行的动作. 然而在实际应用场景中, 视频中可能存在多个人物, 需要模型具有建模多人复杂场景的能力. 随着相关研究的不断深入, 针对多人场景理解的群体行为识别任务[11,18]被提出并被加以研究. 群体行为识别任务是指输入视频中包含多个人物, 需要模型识别出多个人物正在共同完成的群体行为, 比如一起行走、一起排队、一起交谈. 图 1(b)展示了一个群体行为的样例, 视频画面中的多个排球运动员正在相互配合, 共同完成“右方进攻”这一群体动作.



(a) 梳头 (b) 右方进攻

图 1 单人行为识别[4]与群体行为识别[11]任务示意图

多人场景下的群体行为识别技术, 让视频理解模型能够应对场景较为复杂的多人视频, 应用场景非常广泛. 在多人场景的安防监控视频中, 自动地识别监控画面中人群的行为, 在发现异常行为时进行预警. 此外, 在体育比赛视频中, 可以对运动员们的动作进行自动识别和分析. 由于其广泛的应用需求, 群体行为识别研究也受到了广泛关注. 本文对近几年来群体行为识别问题上的大量研究工作进行了深入分析, 总结出了群体行为识别研究的主要难点和挑战, 系统地归纳出了 6 种类型的群体行为识别方法, 包含传统非深度学习识别方法以及基于深度学习技术的识别方法, 并且对未来研究的可能方向进行了展望.

1.2 面临的挑战

相比于简单的单人场景视频, 多人场景视频所包含的视觉信息更加丰富, 对模型提取关键信息的能力要求更高. 通过对大量相关研究工作进行分析, 我们总结了群体行为识别问题的主要难点, 这些难点也是之前大多数研究工作的研究切入点.

1) 如何提取具有区分性的群体特征?

好的特征是准确分类的前提, 高效的特征提取方法也一直是计算机视觉领域研究的核心问题. 传统非深度学习方法往往采用手工设计的特征描述子来提取输入图像中的特征. 而深度学习方法使用卷积神经网络自动学习卷积核来提取特征. 在群体行为识别问题上, 研究人员挖掘出了许多对识别准确率提升有帮助的特征, 包括视觉特征、空间位置结构特征、关系信息特征、场景上下文特征、人物姿态特征、人物运动特征等. 如何准确地提取有利于群体行为识别的特征是一大研究挑战.

2) 如何建模场景中的多层次关系?

群体行为中包含着多层次的关系信息, 比如人与人的交互关系、人与群体的关系、子群体与子群体的关系. 建模并利用这些关系, 可以帮助模型更好地区分不同的群体行为. 然而这些关系都比较抽象, 也缺少标记数据让模型来学习. 因此, 如何让模型关注场景中的关系信息是一个研究难点. 大量研究工作围绕这一问题展开了研究, 并提出了许多针对群体行为识别的关系建模模型.

3) 如何建模视频中的时序动态性?

由于人物行为动作发生在视频中的一个时间段, 仅依靠单张图片往往难以区分, 因此需要模型具有时序建模的能力, 能够提取出时序动态特征来表示人物动作和群体行为的时序动态性. 如何高效并准确地对视频数据进行时序建模, 被许多研究工作所关注.

4) 如何避免无关人物或噪声信息对识别结果造成影响?

群体行为所发生的多人场景往往非常复杂, 包含许多噪声信息. 例如, 可能存在不参与群体行为的无关人物, 背景环境中可能存在与群体行为无关的区域. 这些噪声信息容易给群体行为识别模型造成误导, 给相关研究带来挑战. 因此, 模型需要有注意力机制, 能够关注场景中的关键信息, 并且抑制噪声信息.

1.3 主要技术方法

我们分析了大量的群体行为识别研究工作, 并且根据它们所使用的技术方法和研究动机将这些工作分成了 6 类. 表 1 展示了这 6 种类别的群体行为识别方法, 并且列出了每类方法的典型工作及其年份, 可以看到整个群体行为识别研究的大致时间脉络.

表 1 群体行为识别方法归类

非深度学习方法	基于空间位置结构信息的方法	DCIO[19](2003); NUS-HGA[20](2009); GPE[21](2010)
	基于多元特征描述子的方法	AC[22](2010); RSTV[23](2011); VICAR[24](2012)
	基于非深度学习关系模型的方法	MIR[25](2015); CK[26](2015); DCM[27](2017)
深度学习方法	基于层次化循环神经网络的方法	HDTM[11](2016); CERN[28](2017); SSU[29](2017)
	基于深度学习关系模型的方法	stagNet[30](2018); HRN[31](2018); ARG[32](2019)
	基于注意力机制的方法	SPA[33](2018); AT[34](2020); GAIM[35](2020)

下面我们简单介绍每类方法的基本动机和做法, 在第 2 节和第 3 节详细介绍每种方法的具体内容以及相应的研究进展.

1) 基于空间位置结构信息的方法

最早期的群体行为识别研究工作主要关注人群的空间结构, 希望根据人物的位置信息、移动轨迹来识别出聚集、分开、一起行走等简单的群体行为. 该类方法的主要做法是: 将视频画面看作是一个二维平面, 每个人物看作二维平面上移动的点, 然后提取人物移动轨迹的特征作为分类模型的输入. 这类方法的缺陷是, 仅仅依靠位置信息难以区分一些复杂的、与空间位置无关的群体行为.

2) 基于多元特征描述子的方法

仅利用空间位置信息的群体特征的表征能力较差, 因此研究人员开始尝试挖掘视频中对识别群体行为有帮助的各种特征, 包括视觉特征、人物姿态特征、场景上下文特征、时序动态特征等. 这些基于多元特征描述子的方法通过设计新的特征描述子, 向模型中引入新的人群特征, 增强模型对于不同群体行为的区分能力. 此外, 研究人员也不断对现有特征描述子进行改进, 使得提取的特征更加准确, 提升特征对于视角变化、尺度

变化、背景变化等噪声的鲁棒性.

3) 基于非深度学习关系模型的方法

多人场景中的关系信息对于理解群体行为至关重要, 在各种特征描述子所提取的人物特征和人群特征的基础上, 基于非深度学习关系模型的方法使用隐马尔可夫模型、贝叶斯网络、图模型、条件随机场等关系模型来建模群体行为中的层次化关系结构, 包括人与人的交互关系、人与场景上下文的关系、人和群体的关系、群体和群体的关系.

4) 基于层次化循环神经网络的方法

受到深度学习技术在多个计算机视觉任务上获得成功的启发, 许多研究工作开始尝试使用深度学习模型来识别群体行为识别. 多人场景视频中存在许多序列关系, 比如某个人物在不同帧上的姿态可以看作一个序列, 同一场景下的不同人物的特征可以是一个序列. 基于层次化循环神经网络的方法先使用卷积神经网络提取局部人物特征, 再使用层次化循环神经网络逐步对不同层次的特征序列进行融合, 最终得到整个群体的全局特征用于群体行为识别.

5) 基于深度学习关系模型的方法

高效地建模场景中的关系信息, 可以有效提升模型的群体行为识别准确率. 基于深度学习关系模型的方法尝试将图卷积网络、关系网络、循环神经网络等深度网络应用到群体行为识别模型中, 用于提取关系信息, 提升模型对于场景中的关系信息的感知能力.

6) 基于注意力机制的方法

在复杂的多人场景中, 存在着许多与群体行为无关的噪声信息, 容易给群体行为识别造成误导. 模型需要对噪声信息进行过滤, 增强场景中的关键信息, 比如群体中的关键人物的特征. 而基于注意力机制的方法正是使用深度网络来学习注意力权重, 让模型关注输入中的关键信息, 抛弃噪声和无关信息, 进而提升群体行为识别模型的准确率.

1.4 本文组织

本文第 1 节介绍群体行为识别研究的背景, 说明相关研究的必要性以及可以应用的场景, 并且总结群体行为识别研究所面临的主要挑战以及当前主要的技术方法归类. 然后, 详细介绍每类方法近几年的研究进展状况. 其中, 第 2 节介绍 3 类传统的群体行为识别方法类别, 第 3 节介绍 3 类基于深度学习技术的群体行为识别方法. 第 4 节介绍几个常用的群体行为识别数据集, 并对比主流先进群体行为识别方法的准确率. 在分析完相关研究工作后, 第 5 节针对群体行为识别问题上未来可能的研究方向进行展望. 最后, 第 6 节对全文进行总结.

2 传统非深度学习群体行为识别方法

2.1 基于空间位置结构信息的群体行为识别方法

在群体行为识别研究的早期阶段, 细粒度的视觉理解技术还不够成熟. 在低分辨率的监控摄像头画面中, 个体人物不够清晰, 难以提取人物的姿态、肢体运动等视觉特征. 一些研究人员尝试首先使用目标跟踪算法或者其他传感器来获取场景中每个人物的位置, 基于空间位置结构来建模群体行为. 在这种情况下, 视频可以看作一个二维平面, 每个人物可以看作在二维平面上移动的点, 可以通过对人们的空间位置关系进行分析, 识别出场景中的群体行为. 如图 2 所示: 根据人们移动轨迹的相互关系, 可以判断出他们正在进行聚集、跑步等群体行为.

文献[19]提出将每一时刻所有人物的位置建模成多边形, 基于多边形的形状以及随时间的变形来识别群体行为, 判断当前视频中是否发生异常行为. 基于形状建模可以将整个人群看作一个整体, 而不是对每个人物单独建模. 作者使用 kendall 形状理论来描述每一帧的多边形形状, 并且估计了正常行为的形状概率分布. 为了检测异常行为, 可以估计测试样本的形状概率分布, 基于它和正常行为的概率分布的距离, 基于 KL 散度

指标来判断是否发生异常行为.

图 2 基于空间位置结构信息的群体行为识别方法展示图[20]

文献[36]同样基于人群的空间结构来对多个人物进行整体建模. 作者首先将人物队形表示成三维的多边形, 其中每个顶点是人群中的个体人物, 人物的运动追踪轨迹作为每个顶点的特征, 这样可以得到人群形状追踪矩阵. 然后对矩阵进行分解并且估计矩阵的秩, 最后对人群的群体行为进行分类.

人物移动轨迹的动态关系对于群体行为识别非常重要. 例如: 对于“聚集”这个群体行为, 人们的移动轨迹是从不同的出发位置相互靠近, 而对于“一起行走”, 人们的运动轨迹是朝着相同的方向进行移动. 文献[20]提出使用 3 个层次的局部运动关系来编码群体行为, 包括自身关系、成对关系以及群体关系. 具体而言, 自身关系表示每个人物移动轨迹的内部联系, 每个时刻相比于上一时刻是如何变化的, 刻画了个体人物的行为. 成对关系是指两个人物之间的相互位置变化关系, 比如远离、靠近、追逐. 而群体关系表示某个人物与其他所有人物的行为关系. 给定一个视频, 作者首先使用目标跟踪算法获取每个人物的运动轨迹. 然后, 基于手工设计的滤波器, 从所有运动轨迹中提取 3 种运动关系的特征. 最后, 作者使用 *K*-means 算法来对所有运动关系特征进行聚类, 构建特征字典, 再使用特征直方图来表示每个视频, 用于群体行为分类. 除了提出基于局部运动关系的群体行为识别方法以外 这篇文章还收集了一个包含 476 个视频片段、6 个群体行为类别的监控场景下的群体行为识别数据集 NUS-HGA.

文献[21]同样使用了基于运动轨迹分析的方法来分类群体行为. 不同于之前工作的做法, 作者引入了高斯过程方法来表示运动轨迹, 从概率的角度来处理运动的随机性. 基于高斯过程回归和运动分析, 作者设计了 3 种特征描述子来提取运动轨迹的特征: 个体特征、成对特征以及整体特征. 最后, 基于特征词袋方法, 将数量不固定的人群表示成固定长度的特征向量, 再使用 SVM 对群体行为进行分类.

文献[37]设计了更加细致的运动轨迹特征提取方法. 给定一个视频以及目标跟踪结果, 首先对所有人物进行分组. 作者没有进行硬分组来划分人群, 而是采用概率分组的方法构建概率连接图, 每条边的权重表示两个人物属于同一群体的概率. 这种基于概率的软分组方法可以让模型提取更加鲁棒的群体上下文特征. 为了让提取出的群体特征能够高效地捕捉群体中的结构、运动、动态变化信息, 作者设计了 4 种群体特征提取方法, 分别提取群体连接性特征、连接变化特征、运动方向特征以及运动速度特征. 最后, 同样采用特征词袋方法和 SVM 分类器来识别群体行为.

在一些人数较多的复杂场景中可能会存在多个人物群体, 分别正在进行不同的群体动作. 在这种情况下, 需要模型具有人群划分的能力, 定位出每个小群体的分布, 再结合多个小群体之间的相互关系, 分别对每个小群体的群体行为进行分析. 文献[38]首先使用了一个鲁棒的多目标跟踪算法来跟踪视频画面中每个人物的移动轨迹, 然后计算人物之间的距离, 再使用最小生成树算法来对人物进行聚类, 获得每个群体的分布. 基于划分好的群体以及他们的运动轨迹, 作者设计了基于社交网络分析的特征描述子集合, 包括群体中心、运动直方图、距离直方图、集中性直方图, 为每个群体提取固定长度的特征向量. 这些特征向量描述了每个群体的全局结构以及局部的运动模式. 最后, 作者为每个群体行为类别训练高斯过程动态模型, 计算群体特征向量属于某个群体行为的条件概率, 用于群体行为分类.

文献[39]对人物群体中的层次性进行了分析, 提出一个群体行为由 3 个层次构成: 个体人物动作、子群体行为以及整个群体行为. 这 3 个层次中的关系信息对于识别复杂的群体行为十分重要, 而之前的工作都忽视了子群体之间的关系信息. 基于这些分析, 作者在群体行为识别框架中引用了子群体之间的关系, 不仅考虑

了个体人物之间的相互关系, 还提取了多个子群体之间的交互信息. 作者采用了局部约束线性编码分别为 4 种类型的关系信息提取特征向量, 包括个体关系、成对关系、组间关系以及行为关系.

文献[40]对在单人行为识别任务上表现较好的行为描述子向量进行了扩展, 提出一个新的基于人物运动轨迹的群体行为描述子向量来识别群体行为. 具体而言, 该描述子从 3 个方面来提取群体行为特征: 群体移动轨迹、个体移动轨迹的相关性以及不同子群体的相互移动关系. 基于构建好的群体行为特征向量, 作者在多种自组织神经网络上进行了实验, 都取得了比较好的分类性能.

为了解决在复杂多人场景中存在大量噪声信息、群体行为难以识别的问题, 文献[41]提出了群体交互空间的概念. 对于一个群体行为, 场景中只有部分人物在参与. 而群体交互空间指的就是场景中实际参与群体行为的人物集合. 需要首先对群体交互空间进行检测, 在识别群体行为时排除无关人物, 才能抑制场景中的噪声信息. 作者提出了基于空间关系学的群体交互空间检测方法. 在检测出的群体交互空间上, 作者提出了两种新的特征描述子——群体交互能量特征和引力斥力特征来描述群体行为. 这些特征考虑了人物之间的相互位置关系: 靠近、远离、保持距离. 最后, 使用特征词袋方法和线性 SVM 来进行群体行为分类.

2.2 基于多元特征描述子的群体行为识别方法

除了空间位置信息, 视觉外观信息对于识别群体行为也至关重要. 早期非深度学习算法主要利用手工设计的特征描述子, 比如 HOG, 来提取图像中的视觉特征. 视觉特征可以帮助模型结合人物的姿态外观信息来判断群体行为类别. 此外, 也有工作尝试设计不同的特征描述子, 以提取多人场景视频中人物运动、时序变化、空间结构、上下文环境等多维度信息. 相比于仅使用空间位置结构信息的方法, 基于多元特征描述子的方法充分利用了视频中多种类型的信息, 可以更加准确地对群体行为进行分类.

为了实现鲁棒的群体行为识别, 文献[18]同时利用人群的空间分布、人物的姿态以及人物的运动来识别场景中的群体行为. 提出的解决方案首先采用基于 HOG 描述子的算法来检测视频帧中的人物位置, 并且对他们的姿态进行估计; 然后, 采用扩展的 Kalman 滤波器来追踪场景中所有人物的运动轨迹. 为了让获取的人物运动轨迹更加鲁棒, 作者还估计了摄像头参数和场景水平线, 以减少摄像头移动、视角变化、尺度变化、背景移动等问题所带来的影响. 最后, 作者设计了多种时空局部描述子, 基于直方图统计获取了人物周围其他人物的时空位置、姿态、运动的分布特征, 用于群体行为的分类.

在文献[18]的基础上, 文献[23]提出了基于随机森林的算法来让模型自动学习群体的上下文特征, 用于群体行为识别. 在之前其他工作设计的特征描述子中, 一般采取固定划分区间的直方图来对群体的时空特征进行统计, 而这一工作提出使用自适应的特征空间划分来表示场景上下文. 该方法基于随机森林模型, 首先随机地从时空区域中进行采样, 挑选出最具有区分性的时空区域来计算群体行为特征, 并且生成最优的划分超平面来进行群体行为分类. 与之前工作手工设计的特征描述子不同, 基于随机森林的特征提取方法能够自动发现最优的特征空间划分, 可以更加有效地建模上下文信息. 此外, 作者采用时空马尔可夫随机场模型同时在场景中识别和定位群体行为.

绝大多数的群体行为研究工作都在关注行为识别问题, 即模型对给定视频进行分类. 而文献[22]提出了行为检索的任务: 给定一个行为标签, 需要模型从整个视频中找到与该行为标签最相关的人物. 对于一些发生频率很低的动作, 比如摔倒、打架等异常行为, 正负样本差距较大, 难以训练相应的分类器来识别该行为. 如果基于检索的思想, 可以让模型计算出视频中所有人物与目标行为的相关性, 并且对所有人物进行排序, 则更加容易发现目标行为. 因此在这个工作中, 作者使用了排序 SVM 模型, 根据每个视频片段和目标行为的相关性, 对所有视频片段进行排序. 此外, 作者还设计了人物描述子以及动作上下文描述子, 以提取每个人物自身及其周围环境的特征信息.

动作上下文描述子[22]对于视角变化十分敏感, 不够鲁棒. 文献[24]提出了新的改进特征描述子: 相对动作上下文描述子, 来编码群体中的相对关系, 这一表示具有视角不变性的优点. 该文还提出了另外两个方法以提升模型性能: 首先, 为了让提取的上下文特征对各种情况都比较鲁棒, 作者设计了两种后处理操作——阈值处理操作让群体行为表示来从充满噪声的上下文中获取显著特征属性, 高斯处理操作来减少特征值量化

过程中的误差. 此外, 为了减少局部分类错误, 作者在分类过程中使用了全连接条件随机场模型, 以概率的形式来假设场景中的所有人物都相互关联.

文献[42]提出了一种新的基于局部视觉线索的视频特征表示, 该特征表示旨在从充满噪声的多人场景中发现真正参与目标群体行为的人物. 在提取视频帧的特征时, 不使用所有检测到的人物信息来统计特征, 而是仅仅使用那些被估计正在参与群体行为的人物信息. 基于这一新的特征表示, 作者设计了一种生成式链式模型, 并且使用最大后验概率算法来识别群体行为. 链式模型在推理时会确定哪些人物正在参与群体行为, 并且从全部特征中提取相应的特征子集, 构建一组时序推理链条. 所使用的最大后验概率算法包含两个主要步骤: 对于给定的特征子集, 优化模型使得后验概率最大; 重新确定输入的特征子集, 提取最优的群体行为特征. 两个步骤迭代进行, 最终不仅可以识别出群体行为, 还可以判断出群体行为的实际参与者.

文献[43]提出了一种基于姿态语义的时空特征描述子, 可以在视频中同时捕捉多个人物之间的多尺度行为交互信息. 人物行为可以被看作是一个特定的人物姿态变化序列, 因此可以在视频帧中对人物的重要部位进行检测, 比如头、躯体、腿, 基于人物的姿态信息来判断人物行为. 作者设计了时序姿态描述子, 首先使用姿态估计算法, 在每个视频帧上检测人物重要部位的位置, 作为姿态语义信息. 再统计每个时空位置上的所有人物部位的激活值大小分布, 得到特征向量.

2.3 基于非深度学习关系模型的群体行为识别方法

群体行为中往往包含多种类型的关系信息, 比如人与人的交互关系、人与场景上下文的关系、人与群体的关系、群体与群体的关系. 建模场景中各个元素之间的层次化结构, 利用它们之间的依赖关系, 可以让模型更好地理解多人复杂场景.

文献[44]将群体行为识别问题分解成个体和群体两个层次, 提出了一个两层隐马尔可夫模型框架来识别视频序列中的群体行为. 其中, 第 1 层模型基于原始的音频和视频特征建模了个体人物的行为动作, 第 2 层模型建模了个体人物之间的交互. 通过将原始问题拆解成两个层次, 让模型更加简单、易实现, 并且具有较好的可解释性. 此外, 这也将整体模型进行了模块化, 每一层可以针对不同的子问题使用不同的隐马尔可夫模型, 使得整个模型易于扩展. 实验结果也表明, 两层隐马尔可夫模型的识别准确率要远远高于单层隐马尔可夫模型. 在这项工作中, 作者还使用了多模态的特征, 包括音频和视频两部分的特征, 让模型能够识别出讨论、独白、演讲等需要同时结合声音和画面来判断的群体交互动作.

文献[45]提出了一个基于事件的动态上下文模型来解决群体交互识别任务中的上下文感知问题. 作者设计了一个事件驱动的多层次动态贝叶斯网络来检测场景中的多层次事件. 其中, 低层次的事件包括视频和音频信号, 中层次的事件包括每个个体人物的动作, 而高层次的事件包括整个场景的群体行为. 在事件驱动的多层次动态贝叶斯网络中, 自底向上和自顶向下的推理过程相辅相成, 低层次的事件可以作为判断高层次事件的线索, 而高层次事件也可能作为上下文信息来辅助判断低层次事件. 基于概率图模型, 图中的每个状态节点对应不同抽象层次的事件, 而图边表示事件之间的依赖关系. 在模型进行推理时, 每一时刻都可以计算得到每个状态节点的信念值, 继而可以检测出不同层次的事件.

文献[46]使用隐变量模型框架来建模场景中的上下文信息, 包括人物与群体的交互信息以及人物与人物的交互信息. 之前的研究工作一般使用结构固定的隐变量模型来建模场景, 而该文采用结构自适应的隐变量模型. 作者将模型隐层的结构也当作模型的一个隐变量, 并且在算法运行过程中, 隐式地推理出模型隐层结构. 具体而言, 作者在所有人物的个体动作状态变量之间建立了隐式连接, 表明人物之间的关联性. 通过自适应参数来决定两个人物之间是否产生连接. 自适应的模型结构能够过滤掉输入中的噪声信息, 为每个人物提供与其相关的上下文信息.

文献[47]对群体行为识别中的时空一致性问题进行了研究. 为了解决这一问题, 作者提出了一种基于全连接条件随机场的模型来进行群体行为识别. 该模型假设所有人物之间存在联系. 与之前其他工作使用启发式算法来确定人物交互范围的做法不同. 这一工作在模型中从空间位置、大小、运动、时间位置这 4 个方面建立了人物之间多尺度的关系连接. 这一做法可以让模型能够应对不同类型、大小、形状的多人群体. 具体

而言, 给定一个视频片段, 作者首先检测所有人物的位置框, 然后使用 HOG、光流等特征表述子提取每个人物框的特征. 然后进行两部分的概率计算: 个体概率, 即每个人物的行为概率分布; 以及成对的概率, 即两个人物关联的概率. 最后, 在条件随机场进行最大后验概率优化, 得到群体行为类别.

文献[25]同样关注如何在群体行为识别问题中建模人物与人物之间的关系. 作者没有使用手工设计的特征描述子来表示人物交互, 而是提出了一种基于学习的方法, 可以计算两个人物之间、指定行为类别的交互关系. 具体而言, 模型通过计算两个原子行为特征向量的点积相似度来表示两个人物之间的交互关系. 作者构建了一个交互矩阵来表示原子行为之间的关联性. 最后, 作者还使用低秩矩阵分解对模型进行了优化.

文献[48]提出了一个 3 层 AND-OR 图模型, 可以对多个场景进行层次化建模, 同时表示场景中参与个体对象、个体动作以及群体行为. 基于探索利用策略, AND-OR 图模型可以进行理论形式高效、计算开销低的推理. 该模型推理包含 3 个主要步骤: 首先是基于视频中提取到的特征来对每个图节点进行行为检测; 然后是在图上, 基于子节点检测到的局部行为进行自底向上的推理; 此外还有基于父节点检测到的行为上下文进行自顶向下的推理. 作者还设计了基于探索利用策略的模型推理算法, 让模型在准确率和运行开销之间取得平衡. 此外, 这一工作还收集了一个高分辨率的群体行为识别视频数据集, 高分辨率视频中包含的场景更加复杂, 群体行为的层次化也更加明显.

文献[49]同样使用 AND-OR 图模型来建模群体行为中的层次化时空关系. 在复杂场景的长视频中, 画面中人物数量较多, 需要进行繁重的人物检测和跟踪计算, 并且人物检测结果充满许多噪声信息, 会导致时空 AND-OR 图模型上的推理算法计算耗时严重. 为了进一步提升模型的推理效率, 作者在这个工作中提出了基于蒙特卡洛树搜索的时空 AND-OR 图模型推理算法. 在视频中检测行为时, 使用蒙特卡洛树搜索算法来选择最优的时空区域来进行人物检测和跟踪, 并且在可接受的时间开销下规划合理的推理方案. 实验结果表明: 在保持准确率不变的情况下, 基于蒙特卡洛树搜索的时空 AND-OR 图模型推理算法相比于之前工作的算法[48]有两个数量级的速度提升.

为了捕捉视频特征之间长范围的高阶时空依赖关系, 文献[50]提出了一个层次化的随机场模型. 通过逐层的特征融合, 逐步将背景噪声信息过滤, 保留有价值的群体行为特征. 相比于其他使用条件随机场来进行群体行为识别的工作不同, 作者对模型中的隐变量连接进行了调整, 使用了两个隐层, 分别建模单帧的空间关联和多帧的时序关联, 将时空特征融合分开. 为了让模型能够更好地建模视频中的时序结构, 作者还让隐变量之间的时序连接具有局部性. 基于线性规划操作, 作者设计了自底向上和自顶向下的模型变量推理算法. 此外, 作者使用最大边界框架来学习层次化随机场的模型参数.

文献[51]提出了群体行为中社交角色的概念. 例如: 在体育比赛中的某个时刻, 每个球员可能正在充当进攻者、防守者等角色. 随着时间的变化, 每个人物的角色也在发生变化. 社交角色的概念为人物之间的交互关系提供了语义信息. 作者收集了一个曲棍球比赛数据集, 标注了多层次的行为标签, 包括 11 个个体动作标签、5 个社交角色标签以及 3 个场景事件标签. 此外, 作者还提出了一个结构化模型来建模多人场景中多层次行为之间的依赖关系, 包含低层次的个体动作、中层次的社交角色以及高层次的场景事件. 最后, 作者设计了基于最大边界框架的模型参数训练算法.

大多数群体行为识别方法都假设场景中只会出现一种群体行为, 场景中所有人物共享一个群体行为上下文. 文献[27]提出: 在真实应用场景中, 会同时出现多个群体行为, 它们相互提供了多个上下文线索. 因此, 作者提出了一个整体判别学习框架, 基于多个上下文模型来识别场景中的群体动作. 首先, 作者同时考虑了人群中类内和类间的行为交互关系; 此外, 行为发生的场景也提供了很多有用的上下文信息, 作者在识别框架中也考虑了场景上下文信息; 最后, 作者使用了一个最大边界学习框架来同时建模类内、类间、全局场景等多种上下文线索, 并使用贪心前向搜索算法来推理群体行为标签.

由于受摄像头视角变化、人物外观变化、动作的时序变化等因素的影响, 同一类别的行为动作也会有许多变种, 这给行为识别算法带来了挑战. 为了解决这一问题, 文献[52]提出了一种方法, 通过发现行为类别的图元,即每个行为类别的子类别, 可以有效地建模同一类别行为的不同变种. 作者使用数据驱动的聚类方式来

发现具有区分性的行为子类别, 让模型学习识别每个行为的各种子类别. 具体而言, 作者分别进行视频内和视频间的人物行为聚类, 再训练目标检测器检测各个行为子类别. 基于行为子类别检测结果, 作者使用一个多层次的时空模型框架来建模多人场景, 把行为子类别作为框架的最底层, 可以为上层场景理解提供更加细粒度的线索, 并且能够学习到不同行为子类别以及同一场景中不同人物的动作之间的相互关系.

在某些情况下, 许多视觉识别问题可以被当作计数问题. 例如: 判断一个长视频中是否发生某个事件, 需要计数有多少帧包含该事件; 判断一个人群是否正在进行某个群体动作, 需要计数该人群中的个体动作数量. 在这些问题中, 充分利用子元素之间的基数关系可以减少模型对噪声信息的敏感度, 比如视频中的无关帧以及人群中的无关人物. 基于这些想法, 文献[26]基于多实例学习模型提出了一个灵活的框架, 可以推断隐标签之间的基数关系; 然后, 基于硬基数关系和软基数关系来解决标签识别中多种层次的歧义问题. 作者使用概率结构化核模型来编码实例之间的基数关系, 比如更多、更少、大多数. 所提出的模型在群体行为识别、视频事件检测、视频摘要任务上都取得了不错的性能, 展示了基数关系对于视觉识别问题的重要性.

文献[53]针对群体行为识别任务中的关系建模问题提出了一个生成式模型. 作者设计了一个具有 4 个层次结构的概率图模型, 包括场景事件、群体行为、标准姿态以及可见的姿态. 每个人物在模型中具有 4 个关联变量, 分别对应这 4 个层次. 模型按照从场景事件到可见姿态的顺序、自顶向下依次生成各个变量. 此外, 作者将原有的层次内部的关系转化成层次之间的关系, 这样能够让模型既可以建模多个群体之间的交互关系, 也能够建模群体内部多个人物的交互关系.

文献[54]提出了一个整体框架, 可以同时进行多目标人物跟踪以及识别他们的群体行为. 所提出的模型能够同时跟踪多个人物, 给出每个人物的运动轨迹, 并且能够识别他们的个体原子动作、相互交互动作以及群体行为动作. 之前, 其他工作往往将目标跟踪和群体行为识别分开进行, 这样会丢弃两者之间上下文联系信息. 而高层次的群体行为理解可以帮助获得更加稳定的目标跟踪轨迹, 低层次的个体运动轨迹可以帮助模型更好地理解人物行为. 在这个工作中, 作者没有将目标人物跟踪问题和群体行为识别问题分开独立解决, 而是设计了一个层次化的关系图模型来编码场景中的多尺度关系信息, 同时建模移动轨迹、原子行为、交互行为、群体行为之间的依赖关系.

文献[55]使用两种上下文线索来提升群体行为识别的准确率. 首先是个体人物运动轨迹上下文, 即每个人物的时序变化信息; 然后是单帧场景上下文, 即场景中其他人物的关系信息. 作者设计了一个关系图模型以及全局推理算法, 可以同时进行动作识别、标识维护以及场景上下文建模.

文献[56]使用无向图来表示多人场景, 图中每个节点表示一个人物, 每条边的权重表示两个人物之间的相关程度. 作者提出了一个基于图表示的聚类算法来发现多人场景中的不同交互子群体. 该算法考虑了两种社交信号线索: 社交距离线索以及视觉关注度线索. 社交距离线索是基于空间关系学理论, 定义了两个人物之间不同的社交距离; 而视觉关注度线索表示人物正在看向什么方向. 当两个人物的距离足够近且互相注视时, 他们在图上的边连接权重会很高. 通过对人物进行分组, 可以排除无关人物、较少噪声信息对群体行为识别的影响. 在发现的子群体上, 作者设计了特征描述子来捕捉每个子群体中的运动和交互信息. 最后, 同样使用特征词袋方法来表示群体行为, 并使用 SVM 来作为分类器.

文献[57]设计了一个整体框架来解决时空群体行为定位问题. 这一问题包含两个任务: 首先要识别出视频中发生的群体行为类别, 并且要定位出行为对应群体的时空位置. 由于多人场景十分复杂, 难以准确地追踪多目标, 并且每个人群的人数未知, 群体行为定位问题十分困难. 作者提出了一个隐式关系图模型来同时解决这两个任务, 包括多目标追踪、群体定位以及群体行为识别. 该模型利用了人物之间的上下文关系来对多人复杂场景进行建模, 层次化地编码了人物轨迹间的潜在联系, 并且探索了两种类型的上下文关系: 群体内部的交互信息以及群体间的关联性. 通过在图上迭代地进行信息传递, 让模型能够基于整个图的结构信息来进行群体行为定位与识别.

在某些监控场景, 比如学校、公园、工厂, 会存在多个监控摄像头, 需要同时对多个监控画面中的群体行为进行识别. 在这种多摄像头监控场景中, 人物数量多、场景复杂、存在人物冗余, 给群体行为识别带来了巨

大的技术挑战. 文献[58]针对多摄像头场景提出了基于多摄像头上下文的群体行为检测方法, 能够同时利用摄像头内部以及摄像头之间的上下文信息, 并且不需要知道多个摄像头的拓扑结构. 具体而言, 作者使用了一个包含隐变量的图模型, 其中, 摄像头内部以及摄像头之间的上下文信息通过模型中的隐变量结构, 即图中的边连接来表示. 不同于隐马尔可夫模型、隐条件随机场等结构固定的图模型, 提出的模型没有固定隐变量的结构. 通过对图结构进行自动优化, 可以高效地获取多摄像头的上下文信息. 此外, 作者还设计了新的时空特征描述子, 基于区域内运动的数量和外形来提取复杂场景中的群体行为特征.

文献[59]同样针对多摄像头场景设计了异常行为检测算法. 作者首先搭建了基于多摄像头的多目标跟踪系统; 然后, 基于追踪结果提取每个人物的时空特征. 作者设计了自底向上和自顶向下两种人物聚类方法, 用于确定群体结构. 最后, 基于群体结构特征来对群体行为进行分类.

3 深度学习群体行为识别方法

3.1 基于层次化循环神经网络的群体行为识别方法

近些年来, 基于卷积神经网络的深度学习技术在计算机视觉领域的各个子任务上取得了成功[60−63], 推动了相关模型性能的大幅度提升. 在群体行为识别问题上, 也有许多工作开始选择卷积神经网络来提取视频中的视觉特征. 此外, 由于多人场景视频中存在许多序列关系, 比如某个人物在不同帧上的图像序列、场景中的人群序列, 这些序列可以使用深度循环神经网络来建模, 对特征序列进行融合.

文献[11]提出了一个层次化的深度时序模型, 使用长短期记忆网络来建模整个多人场景的时序动态变化. 该模型包含两个阶段: 第 1 阶段的长短期记忆网络接受个体人物特征序列, 提取个体人物的时序动作变化表征; 第 2 阶段的长短期记忆网络对场景中所有人物的表征进行融合, 得到全局表征用于群体行为理解. 深度卷积神经网络被用于提取人物的视觉特征, 再经过两阶段时序模型, 逐步将低层次特征融合成高层次信息. 这一两阶段的时序模型也被之后许多基于深度学习技术的群体行为识别工作所沿用, 即先提取个体人物表征, 再对所有人物表征进行融合得到全局场景表征, 用于群体行为分类. 作者进行了详尽的消融实验, 验证了层次化时序建模的有效性. 相比于简单的整张图片使用卷积神经网络分类以及单阶段时序建模, 都有很大的性能提升. 此外, 文献[11]还提出了新的群体行为识别数据集 Volleyball, 采集于排球比赛视频, 设置了一传、二传、扣球等排球运动中球员们的技术动作以作为群体行为标签.

文献[64]基于长短期记忆网络提出了一个循环交互上下文建模框架用于群体行为理解. 所提出的群体行为识别框架使用 3 个层次的长短期记忆网络逐层融合人物特征, 分别是个体层次、群体层次和场景层次. 其中, 个体层次长短记忆网络负责提取表征来表示个体人物动作的动态变化, 而群体层次和场景层次的长短期记忆网络分别融合同一群体或整个场景的人物特征. 3 个层次的网络可以分别获得不同层次的上下文特征, 能够进行多层次的多人场景行为理解. 这一框架相比于文献[11]增加了群体层次的融合网络, 可以处理包含多个群体的复杂场景, 并且提供描述能力更强的交互上下文特征. 此外, 作者还使用了光流模态作为模型输入, 并且为每个人物使用双流网络同时提取空间特征和运动特征.

文献[28]同样基于两阶段长短期记忆网络来提取场景特征用于群体行为识别. 由于存在的群体行为识别数据集样本数量都较少, 使得深度循环网络训练不够稳定. 在这个工作中, 作者基于能量最小化、置信度最大化的策略, 对群体行为识别模型的损失函数进行了优化. 首先, 之前的群体行为识别模型往往采用常见的 softmax 层来预测每个类别的概率, 而作者进行了改进, 提出了一个新的能量层对每个类别预测的能量进行预测, 使用能量方程来捕捉各个长短期记忆网络预测结果之间的依赖关系; 然后, 由于输入扰动会带来数值不稳定的原因, 作者没有简单地选择能量最小的类别作为预测结果, 而是额外计算每个预测结果的 p 值, 在能量最小的基础上, 结合置信度最大来选择最终的预测结果, 这让模型预测更加稳定.

文献[65]提出了一个基于语义的群体行为识别框架. 受到循环神经网络在自然语言处理、文本生成等领域获得成功的启发, 作者尝试使用长短期记忆网络来为输入视频生成文本, 得到可解释的语义文本, 用于基于语义的群体行为识别. 具体而言, 所提出的基于语义的群体行为识别框架包含两个阶段: 第 1 阶段, 作者使用

长短期记忆网络来为每个视频帧生成一段说明文字; 第 2 阶段, 作者再将所有生成的说明文字输入到另一个长短期记忆网络中, 预测视频的群体行为类别.

在群体行为识别过程中, 一些相似的局部运动可能会给区分不同的群体行为带来混淆. 例如: 在“聚集”和“分开”两个群体行为类别中, 个体人物的动作都是“行走”, 这时, 只有充分利用上下文信息和人物关系才能准确识别群体行为. 文献[66]提出了一种具有区分性的群体上下文特征来增强模型的识别能力. 作者首先对场景中的子人群进行识别, 然后重点关注子人群之前的关系, 突出不同群体行为之前的差异. 此外, 由于深度神经网络参数多, 在数据量小的情况下很容易发生过拟合问题. 针对群体行为识别任务, 作者提出了一种新的数据增强方法, 以增强跟踪到的人物轨迹. 在这一工作中, 作者使用了门控循环单元网络来学习输入序列中的时序动态变化, 相比于长短期记忆网络, 参数更少、计算更快.

群体行为识别模型一般基于监督学习方式来进行训练, 需要大量标记数据. 文献[67]提出了一个可以进行半监督学习的多层次序列生成对抗网络来进行群体行为识别. 作者使用对抗生成网络来让模型为每个视频自动学习有利于群体行为识别的中间表征. 整个模型由两个子网络构成: 生成器接受长短期记忆网络生成的视频动态特征作为输入, 并生成动作编码作为视频的中间表示; 而判别器接受动作编码和视频特征作为输入, 输出群体行为类别, 并且判断输入样本是否为真样本. 整个网络可以进行半监督学习, 除了进行有监督的群体行为识别任务来训练网络, 还可以使用无监督的对抗样本识别任务来进一步训练网络. 此外, 这一工作还验证了: 在人物视觉特征的基础上, 增加整个场景的视觉特征可以提升最终的识别性能.

文献[68]提出了一个基于全局运动模式的事件识别算法. 为了消除视频中复杂背景的噪声干扰, 作者利用光流来提取视频帧序列中的全局运动模式, 并且同时使用全局运动模式的空间和时间特征来识别视频中的事件. 在这一工作中, 作者仅采用光流作为模型输入, 而没有使用原始 RGB 图像. 这样让模型可以重点关注每个人物的运动信息, 消除人物外观等无关信息的干扰. 在提取好的基于光流的全局运动模式上, 作者首先使用卷积神经网络提取光流图像的空间结构特征, 再使用 LSTM 提取时序动态特征.

在文献[68]的基础上, 文献[69]优化了视频运动模式的提取方式, 不光提取两组球员的运动光流, 还会提取相机运动的光流. 在得到全局群体运动模式后, 作者同样首先使用卷积神经网络提取光流的空间信息, 再使用长短期记忆网络来建模时序动态变化. 在这一工作中, 作者对篮球比赛中群体行为识别进行了深入分析, 并且提出篮球比赛中的一个语义事件往往包含 3 个阶段: 事件前阶段、事件发生阶段、事件后阶段. 这 3 个阶段对场景事件的区分起到了不同的作用: 事件前、后事件发生阶段主要可以区分事件的类别, 比如是三分球还是二分球; 而事件后阶段主要对区分事件的结果有帮助, 比如是否进球. 基于这些分析, 作者设计了两阶段的篮球比赛事件分类方式: 第 1 阶段利用事件前和事件发生时的特征来判断事件类型, 第 2 阶段利用事件发生后的特征来判断事件发生的结果. 最后, 融合两阶段结果得到最终的预测结果.

文献[70]提出: 在多人场景中, 群体行为和个体行为具有互相依赖的关系. 从局部视角来看, 个体人物的行为决定了群体的行为; 而从全局视角来看, 群体行为中包含特定的个体人物行为. 例如: 由于场景中的个体人物都在“行走”, 决定了群体行为是“人群行走”; 另一方面, 由于场景的群体行为是“人群行走”, 决定了场景中多数人物的个体行为是“行走”. 常用的两阶段长短期记忆网络融合个体人物特征得到全局场景特征, 仅仅建模了群体行为依赖于个体行为, 没有建模个体行为对群体行为的依赖关系. 基于这些分析, 作者提出了一种图与长短期记忆网络结合的模型, 同时建模个体行为和群体行为互相依赖的关系. 在局部视角, 人物级别的长短期记忆网络基于人物之间的交互来提取人物动作表征; 在全局视角, 场景级别的长短期记忆网络基于图结构建模群体行为. 此外, 作者还设计了残差长短期记忆网络来提取人物的时序特征.

为了识别视频中的群体行为, 往往首先需要使用目标检测器定位每个视频帧中所有人物的位置, 再进一步提取人物特征. 大多数工作都采用离线的人物检测器, 预先提取好人物候选框, 再进行群体行为识别. 而文献[29]尝试将人物检测网络和群体行为识别网络整合在一起, 端到端一起训练, 形成一个完整的多人场景理解模型. 该模型的优势是让人物检测和群体行为识别共享一个特征提取主干网络, 大大减少了整个模型的运行时间和参数量. 此外, 由于整个模型端到端一起训练, 可以从多任务学习中获取收益. 比如: 通过群体行为

分类信号, 人物检测器可以学习到哪些人物没有参与群体行为, 在人物检测阶段就将无关人员排除. 为了提升人物检测的准确率, 作者使用马尔可夫随机场来从所有候选框中筛选可信候选框, 取代了传统的非极大值抑制做法. 作者采用了基于门控循环单元的两阶段模型来融合时空特征, 并且基于特征相似度来获取一个人物在不同视频帧上的时序特征序列, 因此不需要额外的目标跟踪模型.

文献[71]同样选择将人物检测器和群体行为分类器端到端训练, 通过这种方式来共享主干网络计算, 提升模型的运行效率, 并且让人物检测器能够检测到关键人物, 提升群体行为识别的准确率. 此外, 文献[71]还提出了基于弱监督学习的交互关系识别方法, 通过隐向量学习来挖掘人物和群体之间的交互关系, 并且不需要交互关系和个体行为的标注数据.

人物检测和目标跟踪是很多群体行为识别模型中必不可少的步骤, 然而这两个步骤的计算非常耗时, 并且如果人物检测或目标跟踪发生错误, 会对后续的群体行为识别造成干扰. 为了解决这一问题, 文献[72]提出了不需要人物检测和跟踪的群体行为识别模型. 该模型接受视频作为输入, 直接使用卷积神经网络提取每个视频帧的全局视觉特征, 再使用长短期记忆网络来融合多帧特征, 得到视频表征用于群体行为分类.

3.2 基于深度学习关系模型的群体行为识别方法

由于关系建模对于理解多人场景以及群体行为至关重要, 进入深度学习时代, 相关研究人员尝试结合深度学习技术来对多人场景中的人与人、人与群体的交互关系进行建模. 图模型、关系网络等技术被应用到群体行为识别模型中, 用于提取关系信息.

文献[73]提出了一个基于深度神经网络的层次化图模型来识别多人场景中的个体行为和群体行为. 在第1阶段, 作者使用深度神经网络来识别场景中每个人物的个体行为以及场景的群体行为; 在第2阶段, 作者使用一个基于神经网络的层次化图模型, 通过考虑不同行为类别之间的依赖关系来改进第 1 阶段的预测结果. 具体而言, 第 1 阶段模型使用卷积神经网络为每个人物输出个体行为预测分数, 以及场景的群体行为预测分数; 然后将这些预测分数作为因子图的变量节点, 然后使用信念传播, 在图中的不同区域进行消息传递, 更新每个节点的状态. 在文献[73]中, 作者使用神经网络来模拟因子图上的消息传递过程, 可以更加灵活地建模不同行为类别之间的相互依赖关系.

文献[74]同样尝试将图模型引入到深度网络框架中, 利用场景中丰富的关系信息来进行群体行为识别. 作者首先使用卷积神经网络得到每个人物以及整个场景的行为预测分数. 在初步预测分数的基础上, 再构建关系图, 图上的节点表示个体人物行为或者整个场景的群体行为, 图上的边表示各个行为之间的关系. 通过在图节点之间进行消息传递, 可以对每个节点的行为预测分数进行改进. 作者设计了一种基于深度网络的结构化推理框架, 通过迭代推理的方式确定场景中哪些人物正在交互、哪些人物正在参与群体行为, 并且相应地动态调整关系图中边连接结构. 所提出的结构化推理框架使用循环神经网络来计算节点间传递的消息, 每个节点融合周围节点传递的信息来更新自己的状态.

文献[30]提出了一个注意力语义循环神经网络来建模场景中的时空上下文信息, 用于识别视频中的群体行为. 给定一个视频, 作者显式地构建语义图来描述整个场景, 再通过结构化的循环神经网络来融合时空上下文信息. 语义图中, 每个节点表示一个人物, 而每条边表示人物之间的关系, 边连接根据空间距离和时序连接来确定. 作者设计了节点循环神经网络和边循环神经网络这两个单元, 用来在语义图上进行消息传递. 节点循环神经网络储存了节点的状态特征, 接受连接节点的边的特征作为输入, 更新自己的状态; 边循环神经网络储存了边的状态特征, 接受相邻节点的特征作为输入, 更新自己的状态. 通过消息传递机制, 这两种循环神经网络可以捕捉人物之间的关系, 并且提取具有区分性的时空特征. 此外, 作者还采用时空注意力模型来发现视频中的关键人物和关键帧, 进一步提升了模型的识别准确率.

文献[31]提出了一个层次化的关系网络模型来计算人群中的关系表征. 给定一个描述人物间潜在连接关系的图结构以及节点特征, 关系网络层为每个节点计算它与相邻节点的关系表征, 并且融合得到该节点的新表征. 通过叠加使用多个关系网络层[75], 并且在每一层使用不同的图连接结构, 可以逐步将个体人物特征融合得到全局场景特征, 用于群体行为识别. 作者根据人物的空间位置关系, 手工设计了每一层的关系图连接

结构. 此外, 作者还提出了一个关系自动编码器模型, 在自动编码器模型中插入多个关系网络层, 可以无监督地学习人物和场景表征. 学习到的表征使用 K 近邻搜索算法就能够应用于行为和场景检索任务, 实验结果表明, 不需要标注数据就可以取得比较好的检索效果.

文献[76]设计了一个卷积关系模型, 利用个体人物之间的空间关系信息来识别群体行为. 该模型首先为输入视频图像生成行为激活图. 行为激活图是一个具有空间结构的行为表征, 根据所有人物的位置框, 在每个空间区域上赋值不同行为类别发生的可能性, 包含个体行为和群体行为. 行为激活图编码了行为发生的可能性以及空间位置分布, 能够让下游网络提取人物行为之间的空间关系. 为了提升识别的准确率, 作者还使用一个多阶段的改进模块逐步降低行为激活图中的错误预测. 最后, 一个全局预测模块使用改进后的行为激活图来识别群体行为.

群体行为视频中往往包含许多人物以及交互关系信息, 然后只有少数关键帧中的少数人物决定了群体行为的类别. 因此, 如果要准确地识别群体行为, 需要高效地建模群组中的重要关系, 并且抑制无关的个体动作和交互信息. 文献[77]提出一个基于深度强化学习的模型以逐步改进群体行为中的低层次特征以及高层次关系. 作者首先构建了一个关系图来显式建模视频中的语义关系. 基于关系图, 作者设计了两个强化学习智能体来逐步改进关系图上的低层次时空特征以及高层次语义关系. 具体而言, 在特征层次上, 一个特征蒸馏智能体基于强化学习策略来蒸馏最重要的低层次时空特征. 在关系层次上, 一个关系门控智能体进一步改进图上的关系连接, 让语义图关注与群体行为相关的关系信息.

文献[32]同样使用人物关系图来建模多人场景中的关系信息, 图上的节点表示场景中每个人物的特征, 边表示两个人物之间的关系. 作者设计了灵活、高效的人物关系图构建方法, 让人物关系图可以同时捕捉人物之间的外形关系和位置关系. 其中, 外形关系由一个可学习子网络计算得到, 输入两个人物的视觉特征, 输出他们之间的外形关系权重; 而位置关系根据人物的空间距离决定. 作者设计并对比了多种外形关系和空间关系计算的方法. 为了让模型能够捕捉到复杂场景中多种多样的关系信息, 作者提出了多张关系图集成方法, 即对于一个输入视频, 会使用多个参数不同的子网络计算多张人物关系图. 在构建好的人物关系图上, 作者使用图卷积网络[78]在图上进行特征传递, 每个节点会根据与其他节点的连接权重获取关系信息. 此外, 为了提升模型效率, 作者通过稀疏时序采样来减少模型输入帧数.

文献[79]提出了一个基于多模态关系表征的群体行为识别模型. 作者首先设计了一个对象关系模块, 能够同时利用场景中所有对象的视觉外形特征以及空间位置信息, 建模他们之间的交互关系. 为了提取人物的运动特征, 作者在模型中引入光流提取网络, 并且在模型训练时使用行为分类损失作为监督信号, 对光流提取网络进行微调. 然后, 作者设计了两种门控循环单元来提取每一帧的特征表示: 光流门控循环单元和关系门控循环单元. 其中, 前者提取了对象运动信息和视觉线索之间的关系, 而后者提取了对象位置信息和视觉线索之间的关系. 最后, 作者使用基于时空注意力机制的时序融合层给每一帧赋予对应权重, 将不同帧的特征融合得到视频表征, 用于最终的群体行为预测.

3.3 基于注意力机制的群体行为识别方法

对于人类的视觉感知系统, 输入图像中的某些部分相比于其他部分更加重要, 对于最终决策可以起到更加关键的作用. 因此, 我们的大脑也倾向于首先发现图像中的关键区域, 再对关键区域提供更多的注意力. 近些年来, 研究人员也在尝试在神经网络中应用注意力机制, 让模型能够关注输入中的关键信息, 抛弃噪声和无关信息, 进而提升模型识别的准确率. 如图 3 所示: 使用注意力机制能够定位到篮球赛场上的关键球员, 进而识别出正在进行的三分球、二分球等事件.

在发生群体行为的多人场景视频中, 尽管存在多个人物, 但是其中只有一小部分人物的动作决定了群体行为的类别. 文献[80]提出了一个基于注意力机制的模型, 可以学习在多人场景中检测决定群体行为的关键人物. 该模型不需要额外的关键人物标注数据, 仅仅根据群体行为的标注数据就可以训练. 具体而言, 作者在每一帧为每个人物计算注意力权重, 并且对所有人物的特征根据注意力权重做加权和, 得到该帧的特征表示. 所计算的注意力权重体现了关键人物的时空位置变化. 这些注意力权重根据人物视觉特征, 由一个可学习子

网络计算得到. 通过可视化, 可以看到, 注意力权重较大的人物往往就是场景中的关键人物, 比如真正投篮的球员. 此外, 文献[80]还提出了一个篮球比赛数据集, 定义了三分球、二分球、扣篮等行为事件. 相比于之前的群体行为数据集, 所提出的数据集样本更多, 数据量更大.

图 3 基于注意力机制的群体行为识别方法展示图[80]

文献[81]同样尝试通过定位多人场景中的关键人物来帮助准确地识别群体行为. 作者提出了一个注意力模型, 在每个时刻判断每个人物的群体行为参与度, 并且逐步融合关键人物的时序动态特征. 在这一工作中, 作者主要通过人物的运动特征来判断人物的群体行为参与度, 并且关注两种类型的关键人物: 第 1 种关键人物是在整个视频中都在稳定地移动, 并移动了较长距离; 第 2 种关键人物是在某个时刻进行了非常快速的运动. 作者认为: 这两种人物在多人场景中会是决定群体行为的关键人物, 应该赋予较大的注意力权重. 整个模型也分为两个部分: 第 1 部分网络提取每个人物的个体时序特征; 第 2 部分网络根据学习到的注意力权重融合多个人物的特征, 用于群体行为分类.

文献[33]针对群体行为识别问题提出了一个语义保留的知识蒸馏模型, 希望让模型通过语义保留的注意力自动发现场景中的关键人物, 并且抛弃无关人物. 作者使用知识蒸馏技术来让模型学习如何关注语义重要的人物. 首先, 作者使用教师网络从语义领域学习如何识别群体行为. 教师网络输入的是场景中所有个体人物动作的语义文本, 输出的是群体行为类别. 在教师网络中, 会为每个人物生成注意力权重, 并对所有人物的语义特征做加权和. 然后, 作者设计了包含注意力机制的学生网络, 从视觉领域识别输入视频的群体行为. 在训练过程中, 让学生网络从两方面模仿教师网络的输出: 首先是两者生成的注意力权重要尽可能地一致; 然后, 两者预测的群体行为分数也要尽可能地一致. 由于教师网络根据语义来确定注意力权重, 通过知识蒸馏技术, 学习到的学生网络也能保留一定的语义.

在文献[33]的基础上, 文献[82]在教师网络和学生网络中都增加了图卷积模块. 图中节点表示人物特征, 边表示人物之间的关系, 通过图结构来建模人物之间的依赖关系. 作者根据人物的空间位置来确定图的邻接矩阵, 并使用图卷积网络在人物之间传递信息. 此外, 作者还在未剪辑的长视频中使用模型的中间特征进行群体行为时序检测, 定位群体行为发生的起始时间点.

文献[83]提出了一个两阶段的注意力交互模型来建模群体行为中的层次化交互关系. 在个体层次, 作者在每个时刻计算场景中所有人物两两之间的注意力权重, 每个人物根据与其他人物的注意力权重获取上下文信息, 并且更新自己的特征表示. 作者基于两个人物姿态的相似度来计算他们之间的注意力权重. 在场景层次, 作者使用多层感知机计算每个人物的注意力权重, 表示个体动作与群体行为的关系, 并对人物特征做加权和融合, 得到场景特征用于群体行为分类.

文献[35]提出了一个图注意力关系模型, 可同时从个体和群体两个层次来推理场景中的交互关系, 并且从这些交互关系中学习群体行为的时空演变过程. 作者首先针对群体行为识别问题设计了一种时空图, 包含人物节点和群体节点, 分别对应个体人物动作和群体行为. 图结构为群体行为识别提供了空间结构信息和外形语义特征. 然后, 在构建好的时空图上, 作者使用图卷积网络来推理群体行为中两个层次的交互关系. 在个体层次上, 网络学习到如何为每个人物节点确定其与周围其他节点的关系. 在群体层次上, 网络学习到让群体节点赋予每个人物节点不同的关注度.

Transformer[84]是一种基于自注意力机制的网络结构, 最先在自然语言处理的机器翻译任务上被提出, 并取得成功. 近几年来, 许多研究工作尝试将 Transformer 网络应用到计算机视觉领域的各个任务上. 文献[34]在群体行为识别问题上, 使用人物 Transformer 模型来有选择地提取人物之间的关系信息. 作者首先提取视频中所有人物的特征, 输入到 Transformer 网络中, 融合并改进每个人物的特征. 为了让 Transformer 网络的输入能够充分表示人物的外形特征和运动特征, 作者同时使用 2D 姿态网络和 3D 卷积神经网络分别为每个人物提取静态和动态表征, 并且使用了 RGB 和光流两个模态输入. 作者进行了详尽的实验以探究应该如何融合静态和动态表征, 并且展示了这两种特征的互补关系.

文献[85]关注多人场景中的子群体划分问题. 在多人场景中, 可能存在多个社交群体, 分别在进行不同的社交行为. 针对这一情况, 作者设计了一个端到端可学习模型框架, 可以根据人物之间的社交关系将人物划分成不同子群体, 并且预测每个人物的个体行为以及每个子群体的社交行为. 作者使用 I3D 网络来提取每个人物的时空特征, 再使用 Transformer 结构来改进每个人物的特征. 然后, 作者还使用了图注意力模块来建模人物之间的交互关系. 该模型在传统的、假设场景仅有一个群体的群体行为识别数据集上取得了不错的识别性能. 此外, 作者还在现有群体行为识别数据集上进行了子群体划分以及子群体行为标注, 验证了模型对于场景中多个子群体的行为识别能力.

4 主流先进方法准确率对比

4.1 准确率评估数据集

为了准确评估群体行为识别模型的性能, 促进相关研究的发展, 近几年来, 有许多针对群体行为识别问题的数据集被提出. 这些数据集大多采集于运动比赛、安防监控等包含多人场景的视频中, 定义了多种群体行为的标签. 我们在这一小节中将介绍 8 个群体行为识别数据集, 它们的基本情况展示在表 2 中, 表中列出了每个数据集的数据量大小、类别数量、采集自哪种场景的视频以及当前主流先进方法所取得的群体行为分类准确率结果.

表 2 群体行为识别数据集

数据集名称	长视频数量	视频片段数量	群体行为类别	个体行为类别	场景	年份	典型方法准确率(%)
Volleyball[11]	55	4 830	8	9	运动	2016	94.4[34]
Collective[18]	44	≈2500	5	6	监控	2009	95.7[33]
NBA[86]	181	9 172	9	–	运动	2020	47.5[86]
C-Sports[87]	–	2 187	5	–	运动	2020	81.3[87]
NCAA[80]	257	14 548	11	–	运动	2016	58.1[69]
Hockey[51]	5	58	3	11	运动	2012	62.9[51]
Nursing Home[88]	22	2 990	2	5	监控	2012	85.5[74]
UCLA Courtyard[48]	–	106 分钟	6	10	监控	2012	83.7[49]
BEHAVE[89]	4	163	10	–	监控	2010	77.6[39]
NUS-HGA[20]	5	476	6	–	监控	2009	91.7[21]

Volleyball 数据集[11]从 55 场排球比赛中收集了 4 830 个短视频片段, 其中, 3 493 个视频片段作为训练集, 1 337 个视频片段作为测试集. 每个视频片段都标记了群体动作类别, 总共有 8 个群体行为标签: right set、right spike、right pass、right winpoint、left set、left spike、left pass、left winpoint. 每个视频片段中的中间帧被标记了每个人物的人物框以及他们的个体动作. 总共有 9 个个体动作标签: waiting、setting、digging、failing、spiking、blocking、jumping、moving、standing.

Collective 数据集[18]采集自监控摄像头视频, 包含 44 个视频序列, 总共约有 2 500 个关键帧, 每个关键帧作为一个样本. 一般选择约三分之一的视频序列作为测试集, 剩下的作为训练集. 该数据集设置了 5 个群体行为标签: crossing、waiting、queueing、walking、talking, 以及 6 个个体动作标签: NA、crossing、waiting、queueing、walking、talking. 每个关键帧中, 所有人物正在进行的最多的个体动作被定义成群体动作.

NBA 数据集[86]采集自 181 场 NBA 篮球联赛的比赛视频, 包含 9 172 个视频片段. 其中, 7 624 个视频片段

作为训练集, 1 548 个视频片段作为测试集. 该数据集仅标注了视频的群体行为标签, 没有标注人物个体动作标签. 共有 9 个群体行为标签, 包括二分投篮、二分上篮、三分投篮这 3 种动作, 并且根据动作结果再分成成功、失败并且进攻方获得篮板、失败并且防守方获得篮板这 3 种情况.

不同于其他运动场景数据集仅仅包含一种运动, C-Sports 数据集[87]采集了 11 种运动的视频数据, 包括美式橄榄球、篮球、躲避球、足球、手球、曲棍球、冰球、袋棍球、英式橄榄球、排球、水球, 并且定义了 gathering、dismissal、passing、attack、wandering 这 5 种群体行为. 整个数据集共有 2 187 个视频片段, 其中, 1 317 个用作训练集, 435 个用作验证集, 435 个用作测试集.

NCAA 数据集[80]采集自 257 场 NCAA 篮球联赛的比赛视频. 作者以 4 s 为时间窗口, 将原始长视频划分成短视频片段, 再去除掉没有事件发生的片段, 最后得到 14 548 个标记视频片段. 其中, 11 436 个用作训练集, 856 个用作验证集, 2 256 个用作测试集. 作者定义了 11 种群体行为, 包括三分球成功、三分球失败、罚球成功、罚球失败、二分球成功、二分球失败、上篮成功、上篮失败、灌篮成功、灌篮失败、断球.

Hockey 数据集[51]采集自 5 场曲棍球比赛, 包含 58 个视频片段. 该数据集定义了 3 个群体行为标签: attack play、free hit、penalty corner, 以及 11 个个体动作标签: pass、dribble、shot、receive、tackle、prepare、stand、jog、run、walk、save. 与其他群体行为识别数据集不同, 作者还定义了 5 种人物的社交角色标签: attacker、first defenders、defenders defend against person、defenders defend against space、other.

Nursing Home 数据集[88]采集自疗养院的监控视频, 包含 22 个视频片段, 标记了 2 990 帧画面. 作者定义了有人摔倒和无人摔倒两种场景事件, 希望模型能够自动检测病人的异常行为. 此外, 该数据集还定义了站立、行走、坐着、弯腰、摔倒这 5 种人物个体动作标签.

UCLA Courtyard 数据集[48]采集自校园的高分辨率监控摄像头画面, 包含 106 分钟的视频. 作者定义了 6 种群体行为标签: Walking-together、Standing-in-line、Discussing-in-group、Sitting-together、Waiting-in-group、Guided-tour, 还有 10 种人物个体动作标签: Riding-skateboard、Riding-bike、Riding-scooter、Driving-car、Walking、Talking、Waiting、Reading、Eating、Sitting. 此外, 作者还标注了画面中 17 种对象的位置和类别.

BEHAVE 数据集[89]包含监控场景下的 10 种群体行为标签: InGroup、Approach、Walk Together、Meet、Split、Ignore、Chase、Fight、RunTogether、Following. 作者提供了详尽的人物位置框标注以及人物追踪标注, 以方便研究人员利用空间位置信息和人物动态信息来识别群体行为.

NUS-HGA 数据集[20]同样采集自监控场景视频, 包含 6 种群体行为标签: walk-in-group、run-in-group、stand-and-talk、gathering、fighting、ignoring.

4.2 准确率对比分析

Volleyball 数据集和 Collective 数据集是近几年来被使用最为广泛的群体行为识别模型评估数据集, 我们选择这两个数据集作为评估标准, 对主流先进方法的群体行为识别准确率进行对比, 结果展示在表 3 中. 表 3 中列出了各个工作所使用的方法类型, 展示了每种方法类型中不同方法的性能对比以及不同方法类型的性能差异. 表中字体加粗的工作是每种方法类型的典型方法.

由于不同工作所使用的特征提取方法、主干网络模型、实验设置都有所差异, 再加上数据集本身的局限性, 准确率的高低并不能完全反映群体行为识别方法的优劣. 基于方法差异分析和性能对比, 我们认为以下技术方法对提升群体行为识别模型的准确率有较大帮助.

1) 提升模型输入特征的表征能力. 在群体行为识别问题上, 研究人员挖掘出了许多对区分不同群体行为有帮助的特征, 包括视觉特征、空间位置结构特征、关系信息特征、场景上下文特征、人物姿态特征、时序动态特征等. 通过引入更多类型的输入特征以及提升输入特征的准确性, 模型可以更加准确地识别群体行为. 例如: 文献[24]在文献[22]的基础上提出了具有视角不变性的相对动作上下文描述子, 增强了输入特征对于视角变化的鲁棒性, 取得了更好的识别精度. 此外, 基于深度模型的群体行为识别方法往往可以比非深度学习方法取得更高的准确率, 主要是因为深度网络的表征能力更加强大;

2) 对人群进行关系建模. 群体行为中包含着多层次的关系信息: 人与人的交互信息、人与群体的关系、子群体与子群体的关系. 建模并且利用这些关系信息, 可以帮助模型更好地区分不同的群体行为. 例如, 文献[29,32]都基于深度神经网络来提取人物特征, 其中, 文献[32]使用了人物关系图来建模多人场景中的人物关系, 相比没有进行关系建模的文献[29], 提升了模型的群体行为识别准确率;

3) 使用注意力机制提取场景中的关键信息. 群体行为所发生的多人场景非常复杂, 存在许多与行为类别无关的噪声信息, 给模型识别带来挑战. 而注意力机制可以帮助模型关注场景中与群体行为相关的关键信息, 并且抑制噪声信息. 从表 3 中可以看到, 当前基于注意力机制的方法[33,34]在 Collective 和 Volleyball 这两个数据集上都取得了领先的准确率结果.

表 3 主流先进群体行为识别方法准确率对比

方法	是否使用深度学习	方法类型	Collective (%)	Volleyball (%)
AC[22]	否	基于多元特征描述子	68.2	–
RSTV[23]	否	基于多元特征描述子	70.9	–
VICAR[24]	**否**	**基于多元特征描述子**	**73.2**	–
UF[54]	否	基于非深度学习关系模型	79.4	–
SIM[74]	是	基于深度学习关系模型	81.2	–
HDTM[11]	是	基于层次化循环神经网络	81.5	81.9
MIR[25]	否	基于非深度学习关系模型	83.3	–
Cardinality kernel[26]	否	基于非深度学习关系模型	83.4	–
DCM[27]	**否**	**基于非深度学习关系模型**	**85.5**	–
CRM[76]	是	基于深度学习关系模型	85.8	93.0
SBGAR[65]	是	基于层次化循环神经网络	86.1	66.9
CERN[28]	是	基于层次化循环神经网络	87.2	83.3
stagNet[30]	是	基于深度学习关系模型	89.1	89.3
HRN[31]	是	基于深度学习关系模型	–	89.5
SSU[29]	**是**	**基于层次化循环神经网络**	**–**	**90.6**
GAIM[35]	是	基于注意力机制	90.6	91.9
ARG[32]	**是**	**基于深度学习关系模型**	**91.0**	**92.6**
Actor-transformers[34]	是	基于注意力机制	92.8	94.4
SPA[33]	**是**	**基于注意力机制**	**95.7**	**90.7**

5 未来研究方向

群体行为识别技术的应用场景十分广泛, 在监控视频识别、运动比赛分析、社交行为理解等任务上都可以被使用. 然而, 当前的模型算法在识别准确率、鲁棒性、运行速度等方面依然存在不足, 还难以在现实场景中被大规模应用. 下面介绍几个未来可能的研究方向.

1) 提升模型准确率和鲁棒性. 虽然主流先进的群体行为识别模型[33,34]在相关数据集上已经取得了较高的识别准确率, 但是由于存在的群体行为识别数据集的样本量都较少[11,18], 模型可能发生过拟合, 在实际应用中依然不够稳定. 未来工作可以从扩大训练数据规模、增强模型的人群表征能力、优化关系建模方法、抑制噪声信息等方向入手, 进一步提升群体行为识别模型的准确率和鲁棒性;

2) 关注群体行为检测问题. 当前, 大多数工作主要都只关注简单的群体行为识别问题, 即对于一个输入短视频片段, 只需要输出一个群体动作标签. 而现实视频中可能存在多个群体, 分别在完成不同的群体行为. 这时需要模型具有群体行为检测的能力, 能够自动定位每个群体的时空位置、包含的人物、正在进行的群体行为以及行为发生的起止时间. 当前已经存在了一些关于视频中个体动作检测的研究工作[90–94], 但在群体行为研究领域, 依然缺少针对群体行为检测问题的视频数据集以及高效的模型;

3) 将群体行为识别模型与人物检测模型、跟踪模型进行整合. 许多群体行为识别方法依赖人物检测和跟踪结果[11,32–34], 然而大多数工作都采用相互独立的模型来对视频进行人物检测、跟踪和群体行为识别. 未来工作可以尝试研究如何将这几个模型进行整合, 一方面, 可以通过共享底层特征来节约计算资源; 另一方面, 可以通过端到端学习让检测和跟踪模块为群体行为识别提供更优的结果[29];

4) 提升低频群体行为的识别精度. 在现实场景中, 有些群体行为(比如打架)的发生频率很低, 而准确地识别这些低频异常行为非常具有应用价值. 由于低频群体行为的标记样本少, 导致模型在这些类别上的性能较差. 未来的工作需要提升模型在少量样本标记、类别分布不均衡情况下的模型识别准确率.

6 总 结

本文对多人场景视频中群体行为识别的研究进展进行了总结. 我们首先介绍了视频群体行为识别研究的背景、可以应用的场景以及技术上的主要挑战. 通过对现有研究工作进行分析, 我们总结了 6 类群体行为识别方法, 并且分别对每类方法的内容和研究进展情况进行了详细介绍. 然后, 我们对主流先进方法的识别准确率进行了对比. 最后, 我们总结了当前群体行为识别模型的不足, 并且提出了若干个未来可能的研究方向. 希望本文能够帮助研究人员了解群体行为识别技术, 促进相关研究的进一步发展.

References:

[1] Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. In: Proc. of the CVPR. 2009. 248−255.

[2] Everingham M, Eslami SMA, Gool LV, *et al*. The pascal visual object classes challenge: A retrospective. Int'l Journal of Computer Vision, 2015, 111(1): 98−136.

[3] Lin TY, Maire M, Belongie SJ, *et al*. Microsoft COCO: Common objects in context. In: Proc. of the ECCV, Vol.8693. 2014. 740−755.

[4] Kay W, Ja, ,o C, Simonyan K, *et al*. The kinetics human action video dataset. abs/1705.06950, CoRR, 2017.

[5] Wang L, Xiong Y, Wang Z, *et al*. Temporal segment networks: Towards good practices for deep action recognition. In: Proc. of the ECCV, Vol.9912. 2016. 20−36.

[6] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proc. of the CVPR. 2017. 4724−4733.

[7] Feichtenhofer C, Fan H, Malik J, *et al*. SlowFast networks for video recognition. In: Proc. of the ICCV. 2019. 6201−6210.

[8] Ji S, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition. In: Proc. of the ICML. 2010. 495−502.

[9] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proc. of the NIPS. 2014. 568−576.

[10] Xie S, Sun C, Huang J, *et al*. Rethinking spatiotemporal feature learning for video understanding. abs/1712.04851, CoRR, 2017.

[11] Ibrahim MS, Muralidharan S, Deng Z, *et al*. A hierarchical deep temporal model for group activity recognition. In: Proc. of the CVPR. 2016. 1971−1980.

[12] Goyal R, Kahou SE, Michalski V, *et al*. The "something something" video database for learning and evaluating visual common sense. In: Proc. of the ICCV. 2017. 5843−5851.

[13] Kuehne H, Jhuang H, Garrote E, *et al*. HMDB: A large video database for human motion recognition. In: Proc. of the ICCV. 2011. 2556−2563.

[14] Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. abs/1212.0402, CoRR, 2012.

[15] Liu K, Liu W, Gan C, *et al*. T-C3D: Temporal convolutional 3D network for real-time action recognition. In: Proc. of the AAAI. 2018. 7138−7145.

[16] Sharma S, Kiros R, Salakhutdinov R. Action recognition using visual attention. In: Proc. of the NIPS Workshop. 2015.

[17] Tu Z, Xie W, Qin Q, *et al*. Multi-stream CNN: Learning representations based on human-related regions for action recognition. Pattern Recognition, 2018, 79: 32−43.

[18] Choi W, Shahid K, Savarese S. What are they doing? Collective activity classification using spatio-temporal relationship among people. In: Proc. of the ICCV. 2009. 1282−1289.

[19] Vaswani N, Roy C, Amit K, Chellappa R. Activity recognition using the dynamics of the configuration of interacting objects. In: Proc. of the CVPR. 2003. 633−642.

[20] Ni B, Yan S, Kassim AA. Recognizing human group activities with localized causalities. In: Proc. of the CVPR. 2009. 1470−1477.

[21] Cheng Z, Qin L, Huang Q, *et al*. Group activity recognition by Gaussian processes estimation. In: Proc. of the ICPR. 2010. 3228−3231.

[22] Lan T, Wang Y, Mori G, *et al*. Retrieving actions in group contexts. In: Proc. of the ECCV Workshops, Vol.6553. 2010. 181−194.
[23] Choi W, Shahid K, Savarese S. Learning context for collective activity recognition. In: Proc. of the CVPR. 2011. 3273−3280.
[24] Kaneko T, Shimosaka M, Odashima S, *et al*. Viewpoint invariant collective activity recognition with relative action context. In: Proc. of the ECCV, Vol.7585. 2012. 253−262.
[25] Chang X, Zheng WS, Zhang J. Learning person-person interaction in collective activity recognition. IEEE Trans. on Image Processing, 2015, 24(6):1905−1918.
[26] Hajimirsadeghi H, Yan W, Vahdat A, *et al*. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In: Proc. of the CVPR. 2015. 2596−2605.
[27] Zhao C, Wang J, Lu H. Learning discriminative context models for concurrent collective activity recognition. Multimedia Tools and Applications, 2017, 76(5): 7401−7420.
[28] Shu T, Todorovic S, Zhu SC. CERN: Confidence-energy recurrent network for group activity recognition. In: Proc. of the CVPR. 2017. 4255−4263.
[29] Alahi A, Fleuret F, Fua P, *et al*. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: Proc. of the CVPR. 2017. 3425−3434.
[30] Qi M, Qin J, Li A, *et al*. stagNet: An attentive semantic RNN for group activity recognition. In: Proc. of the ECCV, Vol.11214. 2018. 104−120.
[31] Ibrahim MS, Mori G. Hierarchical relational networks for group activity recognition and retrieval. In: Proc. of the ECCV, Vol.11207. 2018. 742−758.
[32] Wu J, Wang L, Wang L, *et al*. Learning actor relation graphs for group activity recognition. In: Proc. of the CVPR. 2019. 9964−9974.
[33] Tang Y, Wang Z, Li P, *et al*. Mining semantics-preserving attention for group activity recognition. In: Proc. of the MM. 2018. 1283−1291.
[34] Gavrilyuk K, Sanford R, Javan M, *et al*. Actor-transformers for group activity recognition. In: Proc. of the CVPR. 2020. 836−845.
[35] Lu L, Lu Y, Yu R, *et al*. GAIM: Graph attention interaction model for collective activity recognition. IEEE Trans. on Multimedia, 2020, 22(2): 524−539.
[36] Khan SM, Shah M. Detecting group activities using rigidity of formation. In: Proc. of the MM. 2005. 403−406.
[37] Zhang Y, Ge W, Chang MC, *et al*. Group context learning for event recognition. In: Proc. of the WACV. 2012. 249−255.
[38] Yin Y, Yang G, Xu J, *et al*. Small group human activity recognition. In: Proc. of the ICIP. 2012. 2709−2712.
[39] Zhang C, Yang X, Lin W, *et al*. Recognizing human group behaviors with multi-group causalities. In: Proc. of the {IEEE/WIC/ACM} Int'l Conf. on Web Intelligence and Intelligent Agent Technology. 2012. 44−48.
[40] Azorin-Lopez J, Saval-Calvo M, Fuster-Guillo A, *et al*. Group activity description and recognition based on trajectory analysis and neural networks. In: Proc. of the IJCNN. 2016. 1585−1592.
[41] Kim YJ, Cho NG, Lee SW. Group activity recognition with group interaction zone. In: Proc. of the ICPR. 2014. 3517−3521.
[42] Amer MR, Todorovic S. A chains model for localizing participants of group activities in videos. In: Proc. of the ICCV. 2011. 786−793.
[43] Nabi M, Bue AD, Murino V. Temporal poselets for collective activity detection and recognition. In: Proc. of the ICCV. 2013. 500−507.
[44] Zhang D, Daniel GP, Bengio S, *et al*. Modeling individual and group actions in meetings: A two-layer HMM framework. In: Proc. of the CVPR Workshops. 2004. 117.
[45] Dai P, Di H, Dong L, *et al*. Group interaction analysis in dynamic context. IEEE Trans. on on Systems Man Cybernetics, 2008, 38(1): 275−282.
[46] Lan T, Wang Y, Yang W, *et al*. Beyond actions: Discriminative models for contextual group activities. In: Proc. of the NIPS. 2010. 1216−1224.
[47] Kaneko T, Shimosaka M, Odashima S, *et al*. Consistent collective activity recognition with fully connected CRFs. In: Proc. of the ICPR. 2012. 2792−2795.
[48] Amer MR, Xie D, Zhao M, *et al*. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In: Proc. of the ECCV, Vol.7575. 2012. 187−200.
[49] Amer MR, Todorovic S, Fern A, *et al*. Monte Carlo tree search for scheduling activity recognition. In: Proc. of the ICCV. 2013. 1353−1360.

[50] Amer MR, Lei P, Todorovic S. HiRF: Hierarchical random field for collective activity recognition in videos. In: Proc. of the ECCV, Vol.8694. 2014. 572−585.

[51] Lan T, Sigal L, Mori G. Social roles in hierarchical models for human activity recognition. In: Proc. of the CVPR. 2012. 1354−1361.

[52] Lan T, Chen L, Deng Z, *et al*. Learning action primitives for multi-level video event understanding. In: Proc. of the ECCV, Vol.8927. 2014. 95−110.

[53] Zhou Z, Li K, He X, *et al*. A generative model for recognizing mixed group activities in still images. In: Proc. of the IJCAI. 2016. 3654−3661.

[54] Choi W, Savarese S. A unified framework for multi-target tracking and collective activity recognition. In: Proc. of the ECCV, Vol.7575. 2012. 215−230.

[55] Khamis S, Morariu VI, Davis LS. Combining per-frame and per-track cues for multi-person action recognition. In: Proc. of the ECCV, Vol.7572. 2012. 116−129.

[56] Tran KN, Gala A, Kakadiaris IA, *et al*. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. Pattern Recognition Letters, 2014, 44: 49−57.

[57] Sun L, Ai H, Lao S. Localizing activity groups in videos. Computer Vision and Image Understanding, 2016, 144: 144−154.

[58] Zha ZJ, Zhang H, Wang M, *et al*. Detecting group activities with multi-camera context. IEEE Trans. on Circuits and Systems for Video Technology, 2013, 23(5): 856−869.

[59] Chang MC, Krahnstoever N, Lim SN, *et al*. Group level activity recognition in crowded environments across multiple cameras. In: Proc. of the AVSS. 2010. 56−63.

[60] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the NIPS. 2012. 1106−1114.

[61] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the ICLR. 2015.

[62] Szegedy C, Liu W, Jia Y, *et al*. Going deeper with convolutions. In: Proc. of the CVPR. 2015. 1−9.

[63] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: Proc. of the CVPR. 2016. 770−778.

[64] Wang M, Ni B, Yang X. Recurrent modeling of interaction context for collective activity recognition. In: Proc. of the CVPR. 2017. 7408−7416.

[65] Li X, Chuah MC. SBGAR: Semantics based group activity recognition. In: Proc. of the ICCV. 2017. 2895−2904.

[66] Kim PS, Lee DG, Lee SW. Discriminative context learning with gated recurrent unit for group activity recognition. Physics Reports, 2018, 76: 149−161.

[67] Gammulle H, Denman S, Sridharan S, *et al*. Multi-level sequence GAN for group activity recognition. In: Proc. of the ACCV, Vol.11361. 2018. 331−346.

[68] Wu L, He J, Jian M, *et al*. Global motion pattern based event recognition in multi-person videos. In: Proc. of the CCCV, Vol.773. 2017. 667−676.

[69] Wu L, Yang Z, He J, *et al*. Ontology-based global and collective motion patterns for event classification in basketball videos. IEEE Trans. on Circuits and Systems for Video Technology, 2020, 30(7): 2178−2190.

[70] Shu X, Zhang L, Sun Y, *et al*. Host-parasite: Graph LSTM-in-LSTM for group activity recognition. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(2): 663−674.

[71] Zhang P, Tang Y, Hu J, *et al*. Fast collective activity recognition under weak supervision. IEEE Trans. on Image Processing, 2020, 29: 29−43.

[72] Zhuang N, Yusufu T, Ye J, *et al*. Group activity recognition with differential recurrent convolutional neural networks. In: Proc. of the IEEE Int'l Conf. on Automatic Face and Gesture Recognition (FG). 2017. 526−531.

[73] Deng Z, Zhai M, Chen L, *et al*. Deep structured models for group activity recognition. In: Proc. of the BMVC. 2015. 179.1−179.12.

[74] Deng Z, Vahdat A, Hu H, *et al*. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: Proc. of the CVPR. 2016. 4772−4781.

[75] Santoro A, Raposo D, Barrett DGT, *et al*. A simple neural network module for relational reasoning. In: Proc. of the NIPS. 2017. 4974−4983.

[76] Azar SM, Atigh MG, Nickabadi A, *et al*. Convolutional relational machine for group activity recognition. In: Proc. of the CVPR. 2019. 7892−7901.

[77] Hu G, Cui B, He Y, *et al*. Progressive relation learning for group activity recognition. In: Proc. of the CVPR. 2020. 977−986.

[78] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. abs/1609.02907, CoRR, 2016.

[79] 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.

[80] Ramanathan V, Huang J, Abu-E-Haija S, *et al*. Detecting events and key actors in multi-person videos. In: Proc. of the CVPR. 2016. 3043−3053.

[81] Yan R, Tang J, Shu X, *et al*. Participation-contributed temporal dynamic model for group activity recognition. In: Proc. of the MM. 2018. 1292−1300.

[82] Tang Y, Lu J, Wang Z, *et al*. Learning semantics-preserving attention and contextual interaction for group activity recognition. IEEE Trans. on Image Processing, 2019, 28(10): 4997−5012.

[83] 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.

[84] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: Proc. of the NIPS. 2017. 5998−6008.

[85] Ehsanpour M, Abedin A, Saleh F, *et al*. Joint learning of social groups, individuals action and sub-group activities in videos. In: Proc. of the ECCV, Vol.12354. 2020. 177−195.

[86] Yan R, Xie L, Tang J, *et al*. Social adaptive module for weakly-supervised group activity recognition. In: Proc. of the ECCV, Vol.12353. 2020. 208−224.

[87] Zalluhoglu C, Nazli IC. Collective sports: A multi-task dataset for collective activity recognition. Image and Vision Computing, 2020, 94: 103870.

[88] Lan T, Wang Y, Yang W, *et al*. Discriminative latent models for recognizing contextual group activities. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012, 34(8): 1549−1562.

[89] Blunsden S, Fisher R. The BEHAVE video dataset: Ground truthed video for multi-person behavior classification. Annals of the BMVA, 2010, 4(1−12): 4.

[90] Xu M, Zhao C, Rojas DS, *et al*. G-TAD: Sub-graph localization for temporal action detection. In: Proc. of the CVPR. 2020. 10156−10165.

[91] Zeng R, Huang W, Tan M, *et al*. Graph convolutional networks for temporal action localization. In: Proc. of the ICCV. 2019. 7094−7103.

[92] Chen P, Gan C, Shen G, *et al*. Relation attention for temporal action localization. IEEE Trans. on Multimedia, 2019, 22(10): 2723−2733.

[93] Gu C, Sun C, Ross DA, *et al*. AVA: A video dataset of spatio-temporally localized atomic visual actions. In: Proc. of the CVPR. 2018. 6047−6056.

[94] Lin T, Zhao X, Su H, *et al*. BSN: Boundary sensitive network for temporal action proposal generation. In: Proc. of the ECCV. 2018. 3−19.

吴建超(1996－), 男, 硕士, 主要研究领域为计算机视觉, 视频理解, 行为识别.

王利民(1988－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机视觉, 深度学习.

武港山(1967－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为计算机视觉, 多媒体技术.