深度学习框架下群组行为识别算法综述

邓海刚[1],王传旭[2],李成伟[1],林晓萌[2]
(1. 哈尔滨工业大学仪器科学与工程学院,黑龙江哈尔滨 150006; 2. 青岛科技大学信息科学技术学院,山东青岛 266061)

摘 要: 群组行为识别目前是计算机视觉领域的一个研究热点,在智能安防监控、社会角色理解和体育运动视频分析等方面具有广泛的应用价值. 本文主要针对基于深度学习框架下的群组行为识别算法进行综述. 首先,依据群组行为识别方法中"是否包含组群成员交互关系建模"这一核心技术环节,将现有算法划分为"无交互关系建模的群组行为识别"和"基于交互关系描述的群组行为识别"两大类. 其次,鉴于"无交互关系建模的群组行为识别方法"主要是聚焦于如何对"群组行为时序过程的整体时空特征的计算和提纯"进行设计的,故本文从"多流时空特征计算融合""个人/群体多层级时空特征计算合并""基于注意力机制的群组行为时空特征提纯"3类典型算法进行概述. 再次,对于"基于交互关系建模的群组行为识别",依据对交互关系描述方法的不同,将其归纳为"基于组群成员全局交互关系建模""基于组群分组下的交互关系建模"和"基于关键人物为主的核心成员间交互关系建模"3种类别分别概述. 然后,对群组行为识别相关的数据集进行介绍,并对不同识别方法在各个数据集的测试性能进行了对比和总结. 最后,分别从群组行为类别定义的二元性、交互关系建模的难点与不足、群组行为数据集弱监督标注和自学习、视角变化以及场景信息综合利用等方面概述了几个具有挑战性的问题和未来研究的方向.

关键词: 群组行为识别; 分组交互关系; 全局交互关系; 关键人物建模; 多流层级网络
中图分类号: TP301.6 **文献标识码:** A **文章编号:** 0372-2112(2022)08-2018-19
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20211359

Summarization of Group Activity Recognition Algorithms Based on Deep Learning Frame

DENG Hai-gang[1], WANG Chuan-xu[2], LI Cheng-wei[1], LIN Xiao-meng[2]
(1. *School of Instrumentation Science and Engineering, Harbin Institute of Technology, Harbin, Heilongjiang* 150006, *China*;
2. *School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, Shandong* 266061, *China*)

Abstract: Group behavior recognition is currently a research hotspot in the field of computer vision, and has a wide range of applications in intelligent security monitoring, social role understanding, and sports video analysis. This article mainly reviews group behavior recognition algorithms based on deep learning framework. Firstly, by judging "whether a method including group member interaction relationship modeling", it can be classified as "group behavior recognition without interaction relationship modeling(GBRWIR)" or "group behavior recognition based on interaction relationship description(GBRBIR)". Secondly, because GBRWIR mainly focuses on how to design "calculation and purification of overall spatiotemporal characteristics of a group behavior sequence", this article summarizes it as the following three typical algorithms, which are "multi-stream spatiotemporal feature calculation fusion", "individual/group multi-level spatiotemporal feature calculation and merging", and "group behavior spatiotemporal feature purification based on attention mechanism" respectively. Thirdly, for GBRWIR algorithms, depending on its different descriptions of interaction relationship, it can be summarized respectively as "based on group member global interaction relationship modeling", "based on group division and subgroup interaction modeling", and "modeling of interactions between core members". Then, the data sets related to group behavior recognition are introduced, and the test performances of different recognition methods in each data set are compared and summarized. Finally, several challenging issues and future research directions are discussed, which respectively are the duality of group behavior category definition, the difficulty of interactive relationship modeling, the weakly supervised labeling and self-learning of group behavior recognition, and the changes of viewpoint and the comprehensive utili-

收稿日期:2021-10-09;修回日期:2022-01-11;责任编辑:王天慧
基金项目:国家自然科学基金(No.61672035)

zation of scene information.

Key words: group behavior recognition; group interaction relation; overall interaction; key person modeling; multi-stream hierarchical network

1 引言

群组行为包括“视频中多个人做相同动作”和“多数人协作完成某一复杂行为”两种情况,而群组行为识别的任务则是通过对视频序列中组群成员运动特征的感知、计算、提纯,并归纳出稳定的、鲜明的模式,进而再通过分类归纳得出代表整个组群典型行为特征的群组行为类别以及每个成员的行为类别. 近年来,它已经成为计算机视觉、人工智能等领域的热点课题,其在体育赛事分析、异常行为检测及预警、实时人群场景的视频分类等方面具有重要价值. 由于群组行为本身具有复杂性和多样性,以及视频据在采集过程中也会受到视角变化、成员彼此遮挡、复杂场景中无关人员干扰等因素的影响,如何设计高效的识别方法成为了该课题的难点.

群组行为识别主要包含两个过程,即群组时空特征描述和行为属性分类,而群组时空特征描述是最关键的一步. 鉴于群组行为是多人协同合作完成的复杂行为,其时空特征的核心应该是成员之间的交互关系,因此,本文依据群组时空特征描述算法中是否包含“组群成员之间交互关系建模”这一核心环节,将群组行为识别方法分为“无交互关系建模的群组行为识别”和“基于交互关系建模的群组行为识别”两大类.

“无交互关系建模的群组行为识别方法”的主要思想是把群组行为过程视为一个时序整体,这类算法主要聚焦于如何对该“视频时序整体的时空特征进行计算和提纯”,本文将从“多流时空特征计算融合”“个人/群体多层级时空特征计算合并”“基于注意力机制的群组行为时空特征提纯”3类典型算法特点进行归纳和概述. 另外,对于“基于交互关系建模的群组行为识别”算法,依据交互关系建模方法的不同,将现有文献归纳为“基于组群成员交互关系的全局化建模”“基于组群分组下的交互关系建模”和“基于关键人物为主的核心成员间交互关系建模”3种类别分别概述. 简明起见,上述这些群组行为识别方法的归纳分类用图1展示.

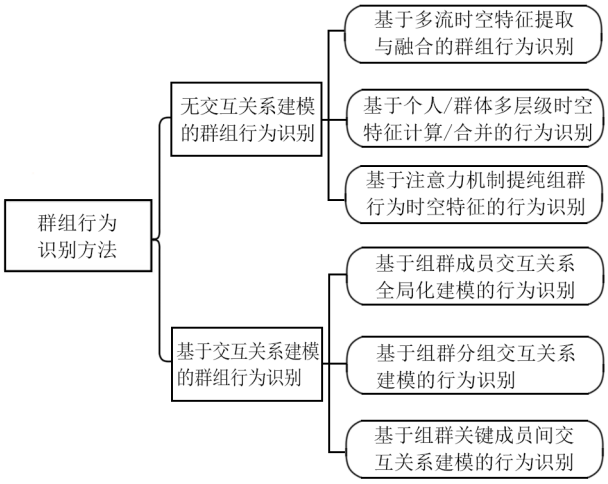


图1　群组行为识别算法的总体分类

2 无交互关系建模的群组行为识别

“无交互关系建模的群组行为识别”实际上是一种相对“粗放的方法”,表现在其缺少了对“群体成员之间细腻的彼此互动”这一环节的描述,而仅仅是对整体场景的时空特征变化进行了刻画. 具体地,主要是对输入图像序列的场景外观、组群成员的姿态、成员运动光流、帧间时间序列依存关系等信息进行提取,并通过分类器对获得的整个群组行为的时空特征进行分类和识别. 在深度学习框架下,这种群组行为识别方法主要是通过CNN,LSTM以及其变形体设计出不同的算法框架,旨在解决“整体组群的时空特征的计算和提纯”. 故本文将现有的对应算法概括为“多流时空特征计算融合”“个人/群体多层级时空特征计算合并”“基于注意力机制的群组行为时空特征提纯”3个类别,现分述如下.

2.1 基于多流时空特征提取与融合的群组行为识别

组群场景信息是多样的,有些信息是相互补充的,因而,利用多种时空特征信息的组合可以达到全面对群组行为特征建模的效果. 这种思路主要是应用在早期的群组行为算法中,典型的就是多流架构特征计算与聚合的识别方法.

为充分利用组群场景的外观信息和运动信息,Simonyan等[1]提出了一种双流网络,其包含空间流支路和运动流支路,如图2所示,其中,空间流支路对RGB图像信息进行处理,主要提取外观特征,运动流对光流数据进行学习和训练,从而提取到运动信息,再将得到的两支路的信息进行融合处理,由于其两条支路提取的不同信息能够互为补充,从而起到丰富组群时空特征的作用,进而达到群组行为有效识别的目的.

由于双流网络模型简单,便于训练,许多学者使用

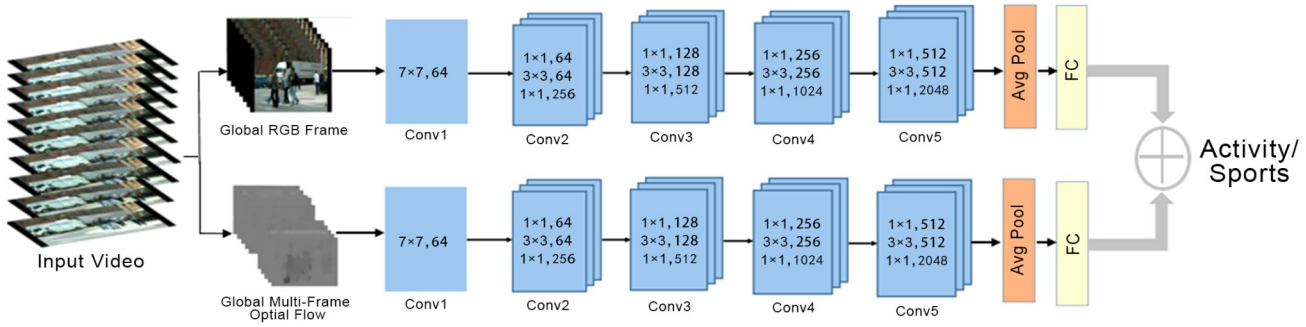


图2 基于双流网络框架的群组行为识别[1]

并发展了它,实现了更多不同特征类别的融合,并应用于群组行为识别中. Borja-Borja等[2]通过一种行为描述向量(Activity Description Vector)得到LRF(Left Right Frequency)图像和UDF(Up Down Frequency)图像的数据,并分别输入到ResNet网络中进行深度特征的提取,最后,将两种特征融合后通过分类器实现群组行为分类. Zalluhoglu等[3]提出了一种利用多个区域提取信息的多流卷积神经网络体系结构,首先将视频序列分为带有背景信息的RGB图像、只包含特定人员的RGB图像信息(Region)和光流序列3种数据;其次,将带有背景信息的RGB图像输入到空间CNN(Convolutional Neural Network)网络,将特定人员(Region)区域图像输入到空间局部CNN网络(Spatial Region Stream CNN)中;同时将光流信息输入到时间局部CNN网络(Temporal Region CNN)和时间CNN网络(Temporal CNN)中;再将Region数据分别输入到空间域CNN网络和时间域CNN网络中的池化层中,从而进一步确定对应的成员和帧;最后提取视频序列的空间信息和时间信息,并将4种CNN网络所提取的时空信息进行融合后,由分类器得到群组行为类别.

鉴于组群外观和光流运动信息易受到光照变化、相机运动等背景因素的影响,组群成员的姿态关节数据逐渐成为被推崇的鲁棒特征选项. 这主要是因为人体姿态可以利用关节点的位置进行刻画,并通过坐标的变化表示姿态的不同,不易受拍摄角度、特征尺度等外界因素的影响,显示其鲁棒性更强而被开发利用. Azar等[4]利用多流卷积网络(Multi-Stream Convolutional Network)对姿态、RGB空间、光流特征进行融合,如图3所示,首先利用CNN1,CNN2和CNN3分别对场景图、光流图和姿态图提取3种特征,其次借助RoiAlign对个体的外观信息、运动信息和姿态信息进行提取,同时,利用全连接层提取整体场景语境表征、运动语境和姿态语境表征,最后对个体特征、整体语义特征分别进行池化操作,实现了不同模态的多流特征融合的群组行为识别.

此外,为了更好地获取群组行为的帧间时序依存信息,王传旭等[5]提出了一种基于多流架构与长短时记忆网络的模型,将全局RGB数据和全局光流数据通过全局LSTM(Long Short-Term Memory)提取全局时空信息,将局部RGB数据和局部光流数据通过局部LSTM提取局部时空特征,并将两种时空特征融合从而得到更加全面的群组特征.

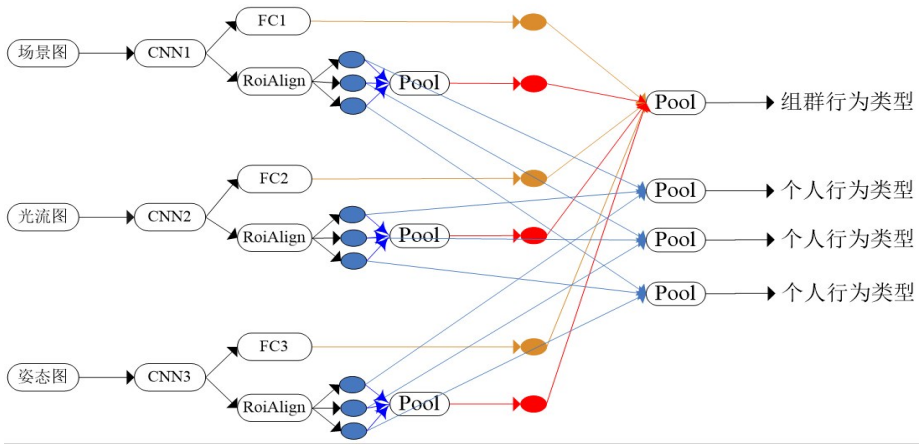


图3 基于多流网络的群组行为识别[4]

概而言之,多流时空特征融合实现群组行为识别的算法,优点是每个支路网络简单,并且在内容上能互为补充,可以全面地描述组群场景的时空特征. 但每一支路常常要预先分开训练,这样会造成整体网络架构训练时间耗费过长;此外,这种多支路网络的训练对数据集规模有一定的要求,如果数据集有限,往往难以收敛或者造成过拟合,故这种多流架构模型的泛化性较差. 因此,为了既能提取不同的时空特征,又能方便网络训练,研究者们提出了能提取个人/群体多层级特征的网络结构,不仅可以获得多语义群组信息,还可以提升模型的泛化能力.

2.2　基于个人/群体多层级时空特征计算/合并的群组行为识别

这类算法的设计原理可以概括为如图4所示的逻辑结构图,鉴于群组行为是由多个成员个体协同完成的,于是,先将每个成员的信息输入到个人级网络中得到个人级特征,再将每帧中个人级特征聚合到组群级网络中得到群组时空特征,最后利用分类器识别群组行为.

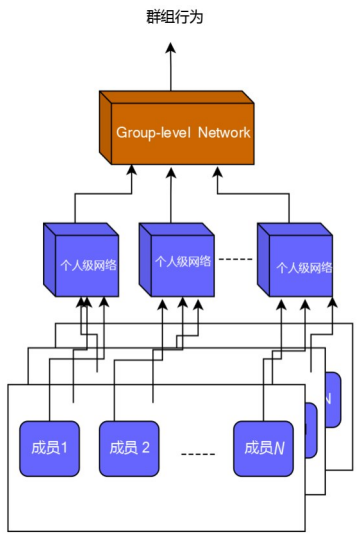


图4　基于层级网络结构的群组行为识别

典型地,Ibrahim 等[6]通过两阶段分层深度时间模型(Hierarchial Deep Temporal Model)实现群组行为识别,首先在第一阶段通过个人级 LSTM 模拟每个个体的轨迹和动作;然后在第二阶段中通过小组级 LSTM 将个体特征进行结合,构成群组特征,建立了人-人、人-群组两种层次的模型,最后针对高层组群特征实现行为识别. Tsunoda 等[7]将分层 LSTM 模型用于对足球运动群组行为的识别中,该模型由 CNN 层和两层 LSTM(即 LSTM1 和 LSTM2)组成,其中 CNN 层提取单人特征,包括成员外观特征和每个人位置信息以及足球位置信息的级联,LSTM1 层提取"球-人之间距离"以及"人-人之间距离",LSTM2 负责集成场景中成员的时序特征;最后由分类器实现了五人制足球群组行为属性的识别. 鉴于 LSTM 可以很好地捕捉序列的长时间依赖,Kim 等[8]提出了一种基于显著子事件的判别组群上下文特征(Disentangled Graph Collaborative Filtering, DGCF)模型来识别群组行为,首先依据视频序列(包含 bounding box)得到每帧内每个成员的位置信息,并计算成员速度和方向获得个人和子组信息,从而得到每一帧的信息;然后为了提取整个视频帧的特征,引入了 GRU(Gate Recurrent Unit)网络进行逐层群组行为特征提取,从而实现群组行为的分类.

除了利用 LSTM/GRU 网络,GAN(Generative Adversarial Networks)网络也被用于层次结构的群组行为识别. Gammulle 等[9]提出了一种基于 LSTM 结构的多级顺序生成对抗性网络,该算法首先利用 LSTM 获得"成员级"和"场景级"的时序特征,经过门控融合单元将上述特征进行聚合,并将其作为 GAN 的生成器输入,由生成器预测当前时序过程的单人和群组行为属性;另外,生成器的预测结果与"场景级"的时序特征经过另一路门控融合单元聚合,由鉴别器对生成器的群组行为预判结果的真伪进行甄别和反馈,最终达到平衡时,得到当前的群组行为判断结果.

另外,由于语义信息对群组行为识别更具有指导性的作用,Li 等[10]提出了一种基于语义的两层结构的群组行为识别模型:第一层为标题生成层,即利用 CNN 分别提取光流和 RGB 特征,并借助 LSTM 对其序列和光流、RGB 信息生成语意标题;第二层为行为预测层,该层也是利用 CNN 和 LSTM 对标题进行推理从而生成群组行为识别标签.

上述多层递次的模型不断迭代应用,不仅可以逐层提纯和融合每层的特征,还能够进一步剖析群组行为内部潜在的高层次语义关系,达到提升群组行为识别精度的效果. 但其相对于轻量级数据集(集体行为数据集[11]和排球数据集[6])由于样本数据量不丰富,而无法保证训练过程中的鲁棒性;再则,这种层次结构越高意味着网络深度越大,对设备的要求也就越高,因此,能够设计一种高效轻量级的组群时序特征提纯模型对群组行为的识别会更有意义,这样就催生了注意力机制下的组群时空特征描述算法.

2.3　基于注意力机制提纯群组行为时空特征的行为识别

群组行为分为两种,如图5所示,一种是大多数人做的相同的动作(图5(a)),另一种是多人协同完成一种行为(图5(b)). 针对后者,群组行为往往并不是由场景中的所有人都参与完成的,而是仅仅由少数的关键人参与并决定的,而那些与群组行为无关的人及动作

则会干扰对群组行为的推断，即在图5(*b*)中“扣球”行为中，并非双方全体球员都参与了这次行为，相反，仅仅由“二传手”和“扣球手”两人来完成的. 因此，剔除团队中无关的人物，确定团队中关键人物成了群组行为识别的重要方法，即群组行为中的关键人物注意力机制.

(*a*) 大多数人做相同动作“Walking”实现的群组行为 (*b*) 由少数人参与并决定的群组行为

图5 两种群组行为类型比较

随着深度学习的发展，注意力机制已在图像识别、字幕识别[12]、机器翻译[13]、人体行为识别[14,15]等领域取得了良好的效果，也为群组行为识别提供了新的思路. Ramanathan等[16]引入注意力机制，通过BiLSTM计算每个个体的行为对群组行为的影响和重要程度，从而区分不同的个体行为在群组行为中发挥的作用. Karpathy等[17]等通过RNN(Recurrent Neural Network)网络实现对视频序列中人物的跟踪及对时变权重的自动学习，提取每个BiLSTM和轨迹信息，通过注意力获取关键人的信息. Lu等[18]提出了一种基于时空注意力机制的GRU模型，通过基于姿势的注意力机制捕捉到每个成员重要的关节点，并通过第一阶段的GRU网络实现对个人动作的识别，然后借助群组级的池化策略找到空间中重要成员并提取时间序列信息，使用时间注意力机制找到关键帧，从而得到最终的群组行为类别. Tang等[19]通过CCG-LSTM模型捕捉与群组行为相关人的运动，并通过注意力机制量化个体行为对群组行为的贡献，通过聚合LSTM聚合个人运动状态，从而实现对群组行为类别的判断. 王传旭等[20]将注意力机制、CNN网络和LSTM网络结合从而提取群组中关键成员的时空信息.

基于注意力机制的群组行为识别方法，不仅能够考虑到所有人的特征，同时还可以依据每个成员在不同时间点上对群组行为的贡献程度，进行空间上和时间上的特征优化，剔除了与群组行为无关的人和帧，有效提纯了组群信息，提高了识别精度.

上述3种架构下的无交互关系建模群组行为识别算法主要是对场景中组群的整体特征进行多线索/多层级的提取和融合，旨在获得组群全面的、显著的行为描述，实现较好的识别效果. 但该类方法所提取的信息仍然局限于组群宏观的整体底层特征描述及其融合，缺乏对群组内部成员之间协同并存、彼此依存关系这一核心信息的挖掘，即缺少成员之间交互关系的建模，最终限制了其识别精度的提升.

3 基于交互关系建模的群组行为识别

与单人行为识别方法不同，群组行为是由多人共同参与完成的，因此，群组行为识别不仅要考虑个体行为、空间位置等信息，还要重点考虑群体中人与人的交互信息. 本文定义群组的交互关系是指群体成员之间互动关系总和，它表现为某一行为过程中的成员间彼此影响、相互制约，并通过该彼此关联信息把整个组群交融成一个整体. 上述无交互关系建模的群组行为识别算法只是从整体上对群组时空特征进行描述，忽略了运动过程中人与人之间的互动关系. 因此，随着群组行为识别研究的深入，建立并推理群组中的交互关系成了群组行为建模的核心任务. 本节依据交互关系建模方法的不同，将其归纳为“基于群组成员交互关系全局化建模的行为识别”“基于群组分组交互关系建模的行为识别”“基于群组关键成员间交互关系建模的行为识别”3种类别分别概述.

3.1 基于群组成员交互关系全局化建模的行为识别

构建群组行为交互关系的过程是具有挑战性的，不仅要考虑到个体自身的信息，还要考虑如何量化人与人之间的关系及关系变化. 典型代表是一些学者提出的全连接图形化交互关系图模型，以此刻画场景中成员整体的交互关系，实现群组行为的“细腻化”描述，如图6所示，每个“蓝色圆点”代表“一个成员”，彼此“连线”代表“交互关系”，该“连线的粗细”表示交互关系的

强弱,它是随时间不断更新的.

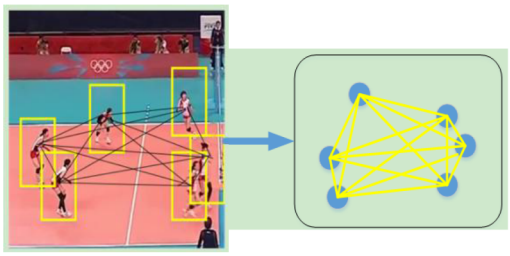


图6　组群整体交互关系图形化描述

典型地,Liu等[21]提出基于全连接的条件随机场模型(Full Connected Conditioned Random Field,FC-CRF)捕捉并推理群组成员间的交互关系,如图7所示.首先,输入的视频图像经过基于卷积神经网络和长短时记忆网络的时序模型,得到群组行为中每个人(图中用i表示)的观测信息x_i,及每个人行为类别y_i的初步预测;然后,基于得到的单人行为信息,使用全连接条件随机场分析人与人之间丰富的交互关系,对每个人的行为类别y_i和群组行为的场景类别进行重新判定.

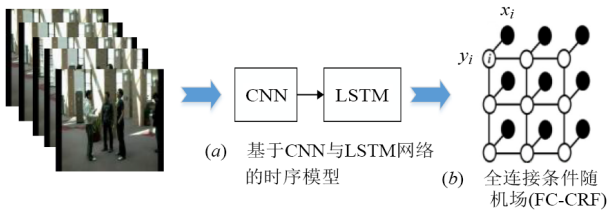


图7　全连接图形化交互关系建模框架图[21]

文献[21]用条件随机场中的二元势函数$\psi_p(y_i,y_j)$描述了人与人之间的交互关系,即

$$\psi_p(y_i,y_j)=\mu(y_i,y_j)k(\boldsymbol{f}_i,\boldsymbol{f}_j) \tag{1}$$

其中,$\mu(y_i,y_j)$是标签兼容函数(label compatibility function),由Potts模型$\mu(y_i,y_j)=[y_i\neq y_j]$给出[22],对于相似度高但是分配了不同标签的人引入这种惩罚机制;而向量$\boldsymbol{f}_i$和$\boldsymbol{f}_j$分别表示第i和第j个人的特征向量,它们实际上是上一阶段基于CNN与LSTM网络的时空特征输出;$k(\boldsymbol{f}_i,\boldsymbol{f}_j)$代表高斯核函数,是根据位置信息向量$\boldsymbol{p}_i,\boldsymbol{p}_j$和特征向量$\boldsymbol{f}_i,\boldsymbol{f}_j$来定义的,即

$$\begin{aligned}k(\boldsymbol{f}_i,\boldsymbol{f}_j)=&w_1\exp\left(-\frac{|\boldsymbol{p}_i-\boldsymbol{p}_j|^2}{2\theta_1^2}-\frac{|\boldsymbol{f}_i-\boldsymbol{f}_j|^2}{2\theta_2^2}\right)\\&+w_2\exp\left(-\frac{|\boldsymbol{p}_i-\boldsymbol{p}_j|^2}{2\theta_3^2}\right)\end{aligned} \tag{2}$$

可以看出,核函数被观测信息所影响,即当同一个群组中具有相近位置和相似特征信息的两个人,他们拥有较强的势函数值,表示此时两者之间交互关系比较强.最后群组行为的识别是通过由该二元势函数参与计算的吉布斯能量概率值实现判别的.

此外,Cheng等[23]通过高斯过程来描述个体运动轨迹,并通过设计的个体行为模式、二元行为模式和分组行为模式3种描述符来捕捉群体行为中人与人潜在的关系.Zhang等[24]通过对群组构造加权关系图,并通过该加权图捕捉每个人的运动和上下文信息,最后通过支持向量机对群组事件进行分类.Lan等[25]提出了一种基于上下文的判别模型,在结构、功能和混合模型3种不同的方法来模拟整个群组中人与人之间的交互关系.Qi等[26]通过节点RNN和边RNN构建个体间交互的语义关系图,从而推理得到每个子组行为和整个群组行为标签.

上述方法虽然能够构建交互关系,但提取的交互关系依然是浅层的、单层次的,这导致其关系表示不够紧凑和深入.因此,为了获取紧凑细致的交互关系表征,Ibrahim等[27]通过关系层来细化关系图,并且关系层中的每对单独的交互特征都映射成一个共享的新特征,并借助去噪自动编码器变体,推断上下文交互信息实现对群组行为的识别.为了加强交互关系描述时的多信息集成,Xu等[28]提出了一种时空注意力机制的多模态交互关系表示模型,首先,引入关系对象模型实现对几何关系和运动特征的建模,再通过关系GRU和Opt-GRU分别对个体间的关系和运动进行编码,从而实现对群组整体交互关系特征的补充.Shu等[29]提出了一种宿-寄结构的基于图LSTM-in-LSTM的网络,首先通过残差LSTM提取每个人的CNN特征,并作为Person-LSTM的输入,从而提取人与人之间的交互关系,然后利用组级记忆单元提取每帧的全局交互关系信息,最后实现群组行为识别.丰艳等[30]提出一种基于伪3D残差网络(Pseudo 3D CNN Network)的群组行为识别模型,一支路通过P3D网络与图卷积网络提取群组中的交互关系特征,另一支路则通过P3D网络本身提取全局时空特征,分别对两支路信息进行识别,最后通过决策融合得到对群组行为的识别.

总而言之,上述通过对整体成员之间的交互关系进行提取和推理,提供了群组成员之间全面的交互关系,可为群组行为识别提供重要的线索.但是,如果场景中参与成员的数量过多,那么在建立群组关系时,其参数量是巨大的,尤其是时空全连接的网络架构会导致网络负荷过大,影响群组行为识别算法的训练,进而影响识别精度.因此,构建高效轻量级交互关系模型成为后续的研究重点.

3.2　基于组群分组交互关系建模的行为识别

成员的数量可能会随着数据集的不同产生差异,从而对群组参与者之间的交互关系图的构建、整体关系特征提取和推理造成影响,尤其是当复杂组群成员众多时的全局交互关系建模,常常会导致设计的网络

参数巨大. 为了降低交互关系建模时的参数量,也为了更好地构建群组交互关系,研究者通常会对成员进行分组交互关系建模,而后再进行组间交互关系融合,从而达到“分而治之”的精准建模效果,这类算法的原理示意如图8所示. 场景中的成员可以按照诸如运动方向、行为属性以及空间距离等特征,再借助聚类算法实现小组群的划分,如图8所示,该场景中的6个人聚合为红、黄、蓝3个小组;然后,对每个小组分别进行交互关系建模;最后,再实现组间交互关系的高层次融合,达到对整个组群特征的多维度多层次描述,进而实现其行为属性判断.

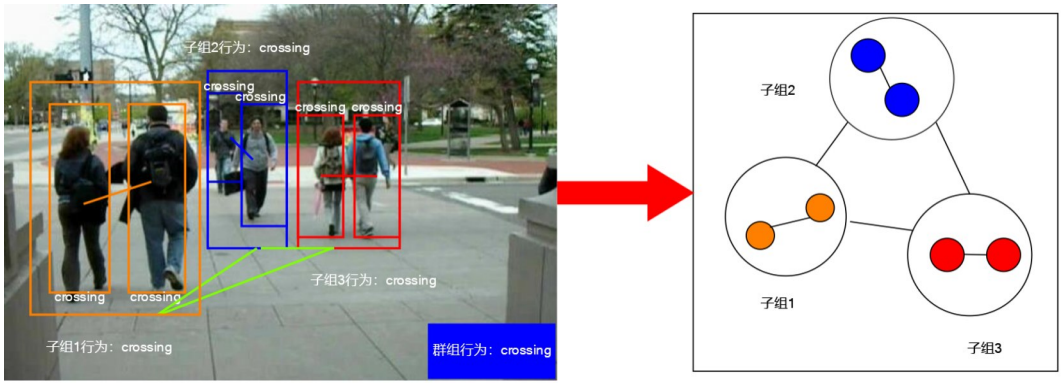


图8 分组交互关系图

典型地,Ehsanpour等[31]认为通常情况下,社会组群需要被分成若干子群体,每个子群体可能从事不同的社会活动. 该算法的子组分割以及单人/群组行为识别原理如图9所示,主要包含以下4个步骤:第一步通过I3D网络提取场景时空特征,并借助ROI Align等模块获得单人特征;第二步从初始化的成员全链接关系图经过图注意力模块迭代,可以获得交互关系强弱不同的交互关系图;第三步则是利用光谱图聚类算法将成员全链接关系图分割聚类为多个子图,这些子图内部成员交互关系相对密切,可以看作“自成一体”;最后则是根据单人特征预测出单人行为,由第一步中的场景特征和第三步中的组群特征合并构成的整体场景特征,进而分类得到组群的行为属性.

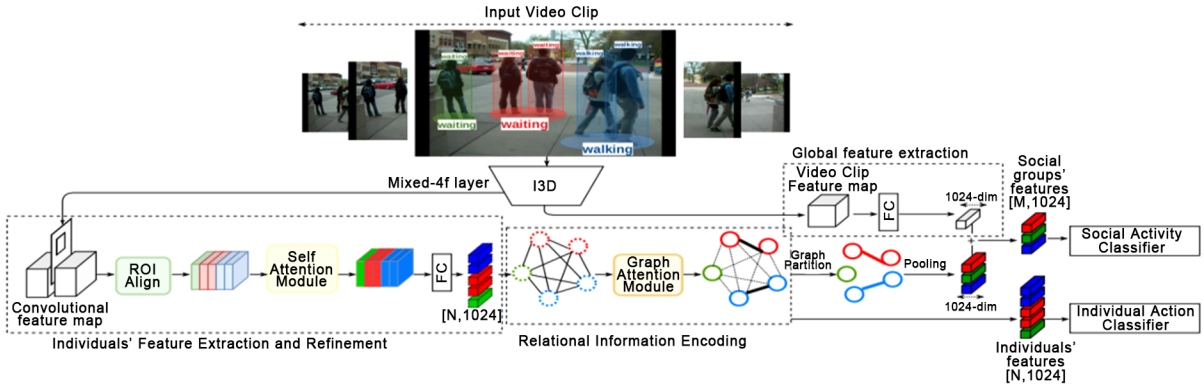


图9 组群分组以及子组交互关系建模下的群组行为识别算法框图[31]

此外,Sun等[32]通过构建潜在的图模型来同时解决多目标跟踪、子群体检测和行为识别问题. 该方法把多目标跟踪和群聚类相结合,即依据成员运动轨迹间的相似度进行聚类实现群组成员的分组,再以星状结构刻画整体组群的分组状态;然后再分别编码组内成员的交互关系和组间的交互关系特征,从而实现对群组行为的识别. Yin等[33]首先通过最小生成树算法将个体聚类成几个子组,然后借助社交网络分析的特征描述提取全局和局部特征,最后采用高斯过程动力学模型来分别建模不同子组的行为. 类似地,Azorin-Lopez等[34]提出了一种群体行为描述向量(Group Activity Descriptor Vector,GADV)表示方法来分析和识别群体活动,该GADV包含3部分:小组行为描述符向量ADV、小组内成员关系描述符IntraGD、组间关系描述符InterGD. 其中的小组行为描述符向量ADV的建立过程如下:先将场景空间等分为若干小单元,计算每个单元内成员轨迹并将它们作为“小组群元”;再通过聚类算法把这些“小组群元”聚合为若干小组群,实现整个场景的组群分组;最后把该小组视为一个实体,计算其中心点的运动轨迹,并进而构建其时空描述符. 而小组内成

员关系描述符 IntraGD 主要是依据其内部成员的运动轨迹计算彼此之间的 4 种交互信息:关联位移(Coherent Displacement)、非关联上移(Incoherent Up)、非关联下移(Incoherent down)、反向程度(Opposite). 组间关系描述符 InterGD 包括如下 4 部分:组间相干性(Coherence of the group)、组间无关性(Incoherence of the group)、组间吸引度 A(Attraction)、组间排斥度(Repulsion). 最后该研究选择了 4 个自组织网络分类器实现单人和组群的分类,这 4 个分类器为自组织网络(Self-Organizing Map)、Neural GAS 网络、监督自组织网络(Supervised Self Organizizing Map)、自组织行为描述符网络(Self Organizing Activity Description Map).

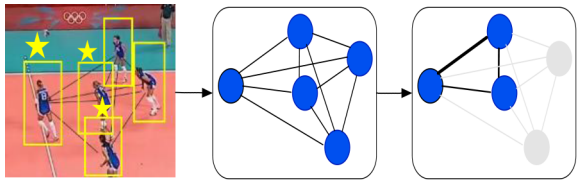
除了上述利用运动轨迹特征实现组群结构分析外,Tran 等[35]通过社会信号线索来测量个体之间的交互程度,并利用图聚类算法来发现场景中具有强相互作用的子群,并丢弃的弱交互作用的子群,从而提取不同子组间的交互关系,进而实现群组行为识别. 还有,Zhang 等[36]提出了一种结构可变的金字塔层级模型来稀疏地表示组群结构. 他们把组群结构的建立看成一个 NP-hard 优化问题,并通过二步迭代算法实现组群成员的结构化分组;而对于成员之间的交互关系建模,提出了 6 种类型的势函数,即成员-场景势函数、成员-成员势函数、群体-成员势函数、群体-群体势函数、行为-群体势函数和群体-场景势函数;最后根据 SVM 分类实现群组行为属性识别.

上述方法能够实现组群自动分组,实现组群的结构化,进而提取组内、组间的交互关系,起到化整为零、降低模型复杂度、更好地应对场景中复杂群组行为的分析的作用. 但这类方法在提取的过程中需要的计算量大,其准确度也有待提升,并且存在交互关系的信息冗余,构建的交互关系网络也会不够简约,最终影响识别的精度. 因此,如何进一步精简场景交互关系建模仍是一个有待深入探讨的问题.

3.3　基于群组关键成员间交互关系建模的行为识别

上述方法主要是借助对群组成员间交互关系的描述达到群组行为识别的目的,但在群组行为识别过程中,并非所有成员对群组的行为识别都是有用的,而通常仅仅是由某些少数成员的行为来决定,这些成员即被称为"关键人物". 为了能抑制无关人员信息从而构建更简约的组群交互关系,研究者们提出了一系列以关键人物为核心的交互关系建模的群组行为识别方法,其思想可以概括为如图 10 所示的原理图.

图 10(a)为排球数据集中防守方场景图,其中带星的为重要的群组成员;图 10(b)为构建的初始全局交互关系图,节点为各个成员,边则为各个成员之间的交互关系;图 10(c)为通过对原始关系图的推理得到关键人物及其交互关系,使得重要的节点和边被加强,不重要的节点和边则被淡化删除,从而得到核心成员的交互关系图.



(a) 防守方场景图　(b) 初始交互关系图　(c) 仅有核心成员构成的交互关系图

图 10　以关键人物为主的核心成员之间的交互关系图

典型地,Yan 等[37]认为群组行为虽然是多人协同参与的复杂行为,但实际上仅仅由核心成员起主导作用,其他成员的作用可以忽略不计. 该算法根据成员的运动状况将"核心成员"定义为如下两种人:其一是在整个行为实施过程中保持稳定且长时间的运动的人;其二是在某个重要时刻有剧烈的运动产生的人. 为此,其提出了一个基于参与贡献度的群体行为时序动态模型(Participation-Contributed Temporal Dynamic Model, PCTDM),如图 11 所示,其算法包含如下几个步骤:首先,在上层的"one" Network 层的 LSTM 模块负责每个成员的时序运动特征提取,并统计每个人的稳定长时间运动程度和突发剧烈运动程度,确定成员的重要性顺序;其次,按照成员重要性次序,利用 Bi-LSTM 模块为成员之间的交互关系建模;然后,利用聚合 LSTM 模块,逐步聚合具有可训练注意力权重的 Bi-LSTM 潜在输出状态,即如果某个成员的行为与群体行为更为关联一致,那么他相应的习得注意系数就会更大,反之亦然;最后,针对图中 Volleyball 数据集场景,聚合 LSTM 模块将场景中的关键人物聚合为左右半场的多个子组,并将它们的 Bi-LSTM 潜在输出状态级联作为分类器输入,实现群组行为识别.

此外,Deng 等[38]通过多层感知机实现个体间的互动及其关系的建模,但这样无法构建其上下文交互关系. 为此,他们进一步提出一种结构推理机(Structure Inference Machine, SIM)[39],利用循环神经网络构建同一场景中个体之间的交互关系,并通过可训练的门控功能来抑制无关人员的影响,突显其他重要成员的贡献. Wu 等[40]提出了一种基于学习的有向角色关系图(Actor Relation Graph, ARG)对整体成员的交互关系进行推理,通过图卷积网络实现对角色关系图中节点间信息的融合,得到信息量较多的节点,即关键成员,通过关键成员节点信息得到群组行为的类别. 为了能够更高效地提取初始特征和推理交互关系,Kuang 等[41]将骨干网络改进为轻量级的 MobileNet 网络,从而更高效

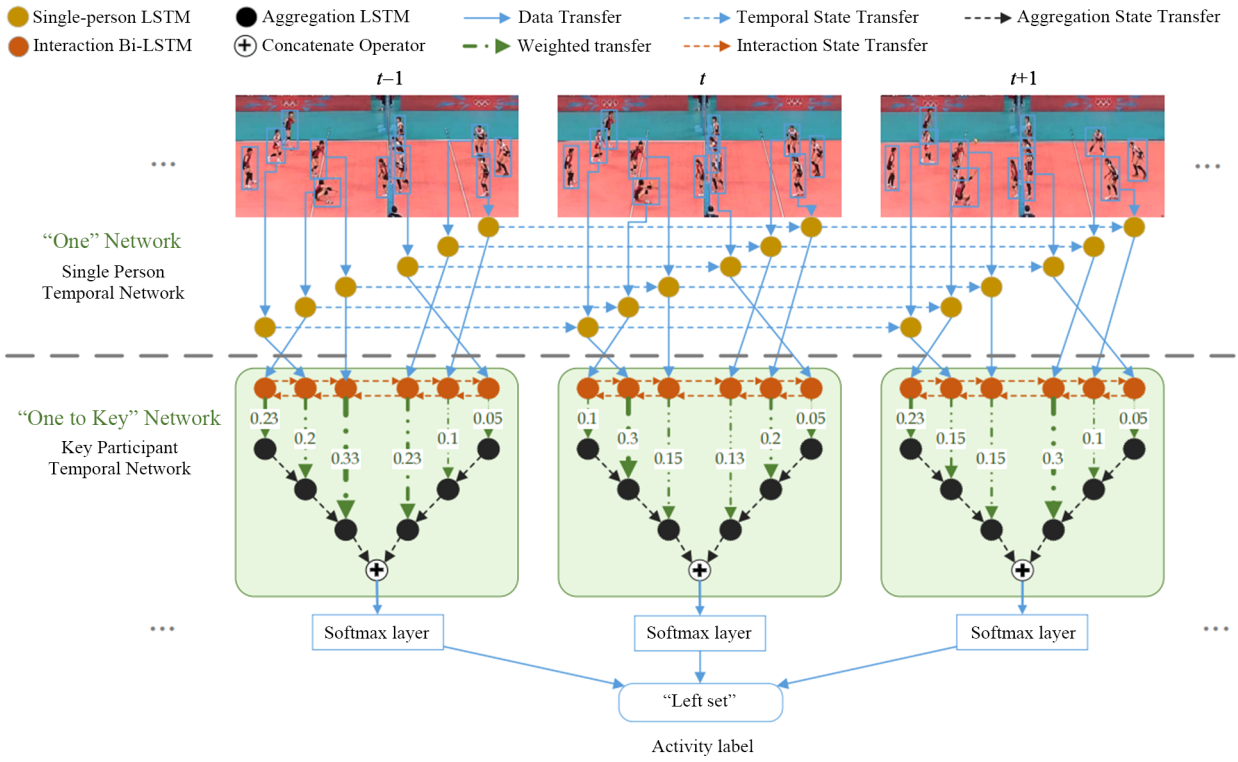


图 11 基于参与贡献度的群体行为时序动态模型[37]

地提取初始特征,并使用归一化互相关和绝对差异之和计算成对的外观相似性来构建角色关系,通过 GCN 网络对关系图进行推理,得到关键人的信息,最后借助关键人的信息得到群组行为标签. Hu 等[42]提出了一种渐近交互关系模型,首先利用帧蒸馏代理网络提取具有信息量的帧,然后通过关系代理网络提取关键人之间的交互关系,实现了对群组行为的分类.

由于受到注意力机制的启发,不少研究人员开始通过注意力机制抑制与群组活动无关的人员和交互关系的干扰,突显群组中重要的成员并进行核心成员关系建模. Zhang 等[43]在图卷积中引入交互注意力机制构成图注意力网络,进而提取群组中重要人物的交互信息. 由于语义特征具有增强网络表达和指导的作用,同时受到 Zhang 等[44]的启发,Tang 等[45]提出了一种基于语义保留的注意力机制模型,该模型包含教师网络和学生网络,使用带有注意力机制的 GCN 分别对动作标签和 RGB 信息进行关系推理,找出重要的标签和个体,并通过教师网络对学生网络进行纠正,提高了识别的准确率. 另外,Yang 等[46]通过 Agnet(Approach Group Net)和 AGTransformer(Approach Group Transformer)提取关键人和关键帧的信息,以实现基于注意力的群组行为识别;同时借助 MST-GCN(Multi-Spatial-Temporal Graph Convolutional Networks)提取每个成员和新来成员的关键关节点实现对人体动作的识别,并利用 G-GCN(Group GCN)提取人与人之间的交互关系从而实现群组行为识别.

总而言之,组群核心成员的交互关系不仅包含关键成员的个人信息,还包括关键人物之间的互动关系,这样不仅可以抑制无关人员对群组活动的影响,还可以提升组群特征描述的精准性,进而提高了群组行为识别精度.

4 群组行为数据集及不同算法实现性能的分析比较

随着对群组行为识别技术不断深入地研究,群组行为数据集也相继推出. 目前,用于群组行为识别的经典数据集如表 1 所示.

表 1 列出了群组行为识别的相关数据集. 不难发现,随着群组行为数据集的规模不断扩大,数据集的种类也不断更新,为群组行为识别未来的发展提供了可靠的数据支撑,下面将重点介绍其中几个典型的数据集.

4.1 集体行为数据集及扩展数据集

CAD(Collective Activity Dataset)数据集包含由低分辨率手持相机收集的 44 个视频剪辑,共有 2 500 个片段,如图 12 所示. 其包含 6 类个人动作标签,即 NA, Crossing, Queuing, Walking, Talking, Waiting,同样包含 5 类群组行为标签,即 Crossing, Queuing, Walking, Talk-

表1　群组行为识别数据集

数据集名称	视频数量	片段数量	个人标签种类	群组标签种类	时间	视频来源	视频类型
NUS-HGA[47]	—	476	—	6	2009年	Youtube	监控数据集
BEHAVE[48]	—	174	—	10	2009年	Youtube	
CAD	25	2 500	6	5	2009年	Youtube	
CAED	30	3 300	8	6	2011年	Youtube	
nCAD	32	2 000	3	6	2012年	Youtube	
Volleyball	55	4 830	9	8	2016年	Youtube	运动数据集
NCAA Basketball	257	6 553	—	11	2016年	Youtube	
C-sports	257	2 187	5	11	2020年	Youtube	
NBA dataset	181	9 172	—	9	2020年	Youtube	

ing,Waiting. 由于相机在采集数据集时角度是固定的,背景是静态的,动作变化也是缓慢的,数据集相对较小,通常会使用早期的深度学习网络来评估. 在实验过程中,一般将70%作为训练集,其余作为验证集和测试集.

鉴于CAD数据集规模较少,因此,提出CAED(Collective Activity Extended Dataset)数据集对其进行了拓展. 该数据集将Walking动作从CAD中移除,并补充了两个新的动作类型,分别是Dancing和Jogging,因此,CAED数据集共有6种行为标签,分别是Crossing,Queuing,Dancing,Talking,Waiting,Jogging. 每个人都分配有一个行为标签,每一帧图像也包含一个群组行为标签.

同样,nCAD(new Collective Activity Dataset)数据集依然是CAD数据集的扩展,包含了6个集体行为类别(Crossing,Queuing,Dancing,Walking,Waiting,Jogging),8种姿势标签(right,right-front,…,right-back). 除了上述标签外,增加了所有序列中的动作标签、交互标签、以及每个人体目标与这两者标签的对应关系标注. 8种交互标签为Approaching(AP),Leaving(LV),Passing-by(PB),Facing-each-other(FE),Walking-side-by-side(WS),Standing-in-a-row(SR),Standing-side-by-side(SS),No-interaction(NA).

(*a*) "Waiting"群组行为　(*b*) "Moving"群组行为

图12　CAD数据集中的2个群组行为类别举例

在集体行为数据集中,集体行为的属性主要是依据大多数人的行为来进行判断的,即大多数人的行为标签即为群组行为标签.

4.2　排球数据集VD(Volleyball Dataset)

群组行为的定义并非仅仅是对大多数人做相同行为的描述,而更多的是对组群成员协同完成复杂行为的刻画. 为此,为了评估深度学习模型的泛化性,许多学者对运动数据集进行了提升,最常用的运动数据集为排球数据集(VD)[49~51]. 该数据集是基于公开的Youtube排球比赛视频收集而成的,如图13所示,共有4 830帧,55段视频. 对于每一帧,每个人都被赋予一个动作类型(Waiting,Setting,Digging,Failing,Spiking,Blocking,Jumping,Standing,Moving),同时包含每组的群组行为类型之一(right-pass,right-spike,right-set,right-winpoint,left-pass,left-spike,left-set,left-winpoint). 由于相机采集数据集时为可调的,故视频中参与者的运动变化相对适中. 通常该数据集的72%用于训练,28%用于验证和测试.

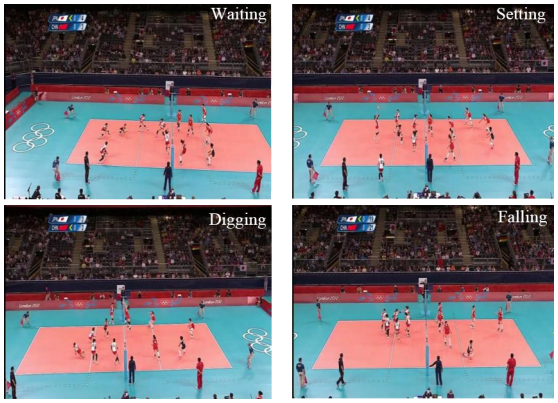


图13　排球数据集中的4个群组行为类别举例

图13为排球数据集,主要展示了数据集中的4种群组行为类型:"左扣杀""右赢球""右扣杀"和"右发球". 而在排球数据集中,排球运动为所有人共同完成或关键人的动作决定,因此,其群组行为的标签为关键人的行为标签.

4.3　NBA数据集(NBA Dataset)

大多数用于群组行为识别的数据集对个人和群组行为都进行了标注,但NBA数据集则仅有视频级标注,并没有单人级的标注信息,其更适合于弱监督下的群组行为识别[52]. 该数据集包含了9 172个视频剪辑,共

包含了9种群组行为:2p-succ,2p-fail-off,2p-fail-def,2p-layup-succ, 2p-layup-fail-off, 2p-layup-fail-def, 3p-succ, 3p-fail-off,3p-fail-def. 在实验过程中,通常将该数据集的83%作为训练集,17%则作为测试集.

4.4 C-Sports数据集(Collective Sports Dataset)

现有体育运动数据集大多数是只针对一种运动进行分类,其种类有限,缺乏多样性,无法支持复杂和有代表性的模型的训练,为此,Zalluhoglu等[53]提出了一种新的群组行为数据集——Collective-Sports数据集(简称"C-Sports"),有效解决了现有数据集存在的局限性问题,该数据集中包含11个团体体育运动标签(A. Football, Basketball, Dodgeball, Football, Handball, Hurling, IceHockey, Lacrosse, Rugby, Volleyball, Waterpolo)和5种群组行为标签(Gather, Dismissal, Pass, Attack, Wander),其中数据集的80%作为训练集,20%则作为测试集,如图14所示.

图14中,从左到右、从上到下,运动类别分别为美式足球、篮球、躲避球、足球、手球、投掷、冰球、长曲棍球、橄榄球、排球、水球,其群组行为类型为"gather""pass""wander""dismissal""wander""dismissal""attack""wander""gather""gather""wander".

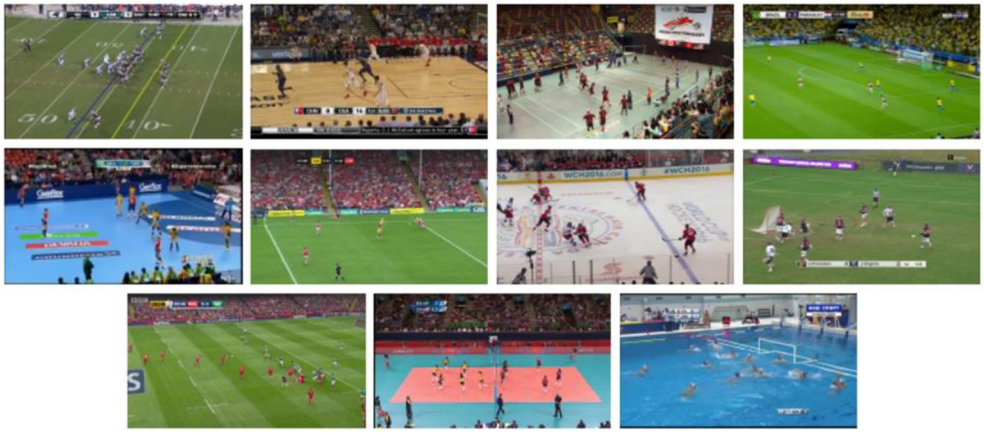


图14 C-sports数据集中的不同群组行为类别举例

4.5 典型群组行为识别方法性能的比较和分析

本文分别从无交互关系建模的群组行为识别方法和基于交互关系建模的群组行为识别算法对群组行为进行了阐述,并对相关数据集进行了介绍. 为了验证各种算法在不同数据集上的性能,表2和表3对其进行了展示. 其中,OF表示光流图像,代表相邻时间图像之间的瞬时速度;Pose代表姿态信息;"—"表示为没有进行实验;其百分比表示每类算法在该数据集上的平均精确度MPCA.

表2所示的为基于无交互关系建模的方法,通过对场景信息、运动信息等组群整体信息的提取,实现群组行为的识别. 不难发现,大部分输入仍旧为RGB图像,取得了一定的效果,但算法MCN[9]除了RGB信息外,还引入了光流、姿态两路视觉信息,这3路特征信息最后经过池化融合后作为场景组群的时空特征,在CAD数据集上获得了95.26%的高平均识别精度,同时在Volleyball数据集也获得了90.42%的较高平均识别精度.

另外,表2中MLS-GAN[9]提出的基于LSTM结构的多级顺序生成对抗性网络,利用两个层次的"成员级"和"场景级"时序特征,经过门控融合单元聚合后,再利用GAN生成器对该融合特征进行单人和群组行为属性预判;另外鉴别器对生成器的群组行为预判结果的真伪进行甄别和反馈,最终达到平衡时,得到当前的群组行为判断结果,由于该算法经过GAN网络的多次矫正优化,在Volleyball数据集获得了92.40%的较高平均识别精度.

相比之下,表3中的算法通过对群组交互关系进行了提取和分析,细化了群组行为特征,在相同的数据集上平均识别精度均有明显提高. 其中,算法XU[28]利用两个层级模块重点挖掘和充分利用了成员交互关系,具体地,将成员外观特征和位置信息输入到关系模块(Relational model)获得初始交互关系表达,然后,该特征与光流特征分别输入到Relation-GRU和Opt-GRU模块,两者融合得到帧级交互关系描述,最后利用注意力机制进行时序特征聚集,将不同权重的帧级特征整合视频级表示,作为分类器输入实现群组行为的识别. 可见正是由于该算法对交互关系进行多层次提取融合,才得到了组群时空特征的精准表示,在Volleyball实际上取得了93.49%的高平均识别精度. 另外,GLIL[29]在CAD和Volleyball数据集都取得了94.40%和93.04%较高平均识别精度,主要是得益于提出的GLIL(Graphical LSTM-In-LSTM)网络架构,它被形象地比喻为"宿主-寄生"体系结构,"寄生"模块是负责建立每个成员之间的

表2　无交互关系建模的群组行为识别方法在不同数据集下的性能比较

Method	Date	Input	CAD	CADE	UCL Courtyard	Volleyball	NBA dataset	Multi-camera Futsal Game dataset
Choi[54]	2012	RGB	80.2%	83.0%	—	—	—	—
DLM[55]	2012	RGB	78.4%	—	—	—	—	—
SIM[56]	2015	RGB	83.40%	90.23%	—	—	—	—
Zappardino[49]	2021	RGB+OF	—	—	—	91.00%	—	—
XU[28]	2020	RGB+OF	91.2%	—	—	93.49%	—	—
GLIL[29]	2020	RGB	94.40%	—	—	93.04%	—	—
LARG[40]	2019	RGB	92.60%	—	—	91.00%	—	—
DRGCN[57]	2020	RGB	89.60%	—	—	92.20%	—	—
STPS[45]	2018	RGB	95.70%	—	—	90.00%	—	—
GAIM[58]	2020	RGB	90.60%	91.20%	—	91.90%	—	—
CRM[59]	2019	RGB	94.20%	—	—	93.04%	—	—
SAM[52]	2020	RGB	—	—	—	—	47.50%	—

表3　基于交互关系建模的方法在不同数据集下的性能比较

Method	Date	Input	CAD	CADE	UCL Courtyard	Volleyball	NBA dataset	Multi-camera futsal game dataset
Canon[7]	2017	RGB	—	—	—	—	—	63.40%
Gavrilyuk[60]	2020	RGB	80.60%	—	—	—	—	—
Region based multi-CNN[3]	2019	RGB,OF	88.9%	—	—	72.40%	—	—
SRNN[61]	2018	RGB	—	—	—	89.90%	—	—
MCN[62]	2018	RGB,OF,Pose	95.26%	—	—	90.42%	—	—
Ibrahim[6]	2016	RGB	81.50%	—	—	51.50%	—	—
Wang[63]	2017	RGB	89.40%	—	—	—	—	—
PCTDM[37]	2018	RGB	92.20%	—	—	88.10%	—	—
StagNet[26]	2018	RGB	—	90.20%	86.90%	89.90%	—	—
Lu[18]	2019	RGB	—	—	—	91.90%	—	—
Tang[19]	2019	RGB	93.00%	—	—	90.70%	—	—
MLS-GAN[9]	2018	RGB	91.20%	—	—	92.40%	—	—
Lu[64]	2021	RGB	91.31%	—	—	92.35%	—	—

交互关系建模,而“宿主”模块负责群体级行为建模,即将多个成员运动信息根据其对群体行为的贡献,选择性地整合并存储到“宿主”中,实现对全局交互关系的关键时空特征的选择和提纯,保障了较高的识别精度.

总的说来,通过对群组交互关系的提取和分析,可以达到细化群组行为特征的效果,使得在相同的数据集上相比较粗狂的无交互关系建模的方法,平均识别精确度都会有提高,因此,基于交互关系的群组行为识别的方法从整体上优于无交互关系的群组行为识别.

除此之外,从两个表格中发现,大多数算法是基于CAD和Volleyball数据集进行研究的,其识别效果大多数在80%以上.但也能发现,每种数据集仅能表示某一类的群组行为,缺乏多样性.因此,研究者们不断引入NBA,BFH[65]和C-Sports等数据集以便应用其他场景中.然而这些新数据集的引入并没有达到经典数据集的识别效果,在群组行为识别的效率和识别精度都有待提高.

5　总结与展望

5.1　总结

本文首先对群组行为识别的研究背景和研究意义进行了阐述,然后依据群组行为识别方法中是否包含“成员交互关系建模”,将其分为无交互关系建模的群组行为识别和基于交互关系建模的群组行为识别两大类;最后,介绍了相关的数据集以及两类群组行为识别方法在不同数据集下的性能比较.下面进一步对这两类算法的各自优势进行总结.

(1)无交互关系建模的群组行为识别方法可以从

视频序列提取场景特征并进行识别. 其中,基于多流网络的群组行为识别,能够通过不同信息的互相补充,从而丰富群组特征;基于层次结构的群组行为识别,能够通过逐层聚合获得群组特征;而基于注意力机制的方法,能够抑制场景中的冗余信息,从而提取群组中重要的时空特征. 这3类算法的先进性总的说来是逐步提升的.

(2)交互关系为群组行为的关键信息,因此,通过捕获群组行为过程中的交互关系,能进一步细化群组特征. 其中,基于整体交互关系建模的群组行为识别能够提取并推理成员整体的交互关系,从而为群组行为识别提供全面的关系特征;基于分组交互关系建模的群组行为识别通过对群组成员进行分组关系建模并融合,能够化整为零,从而达到"分而治之"的效果;基于以关键人物为核心的交互关系建模的群组行为识别方法,能够捕获群组中关键成员以及与其密切相关的其他成员的特征、以及他们的交互关系,抑制与群组行为无关成员的信息,从而降低了群组行为识别过程中的噪声干扰,提高了群组行为识别效率. 上述这3类算法的先进性总体上也是逐次进步的.

综上,无交互关系建模方法只是对场景整体信息进行笼统地提取而实现群组行为识别,其缺陷就是忽视了群组成员间的交互关系,使得其组群时空特征更多地只关注了底层特征,缺少高层交互以及语义特征的刻画;而基于交互关系建模的群组行为识别则更加细化了成员之间的互动以及语义表达,因此,它优于无交互关系建模的群组行为识别方法.

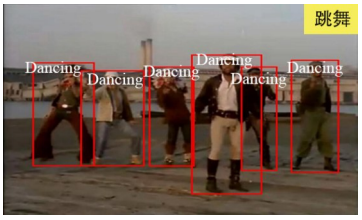
另外,基于交互关系建模实现群组行为识别的方法也有其不足,可以归纳为两点. 其一是需要较多底层特征的支持,因为交互关系建模主要是两两成员间(pair-wise)的交互关系描述,除了基本的CNN/LSTM时空特征外,还需要成员的位置信息、运动轨迹、邻域上下文信息等信息,以便构建成员彼此之间关系[21],但这些信息需要多目标跟踪算法作为底层特征提取的保障,但是这些底层算法的精度却是有限的,因此,导致成员间交互关系的精度不高;其二表现为多层级交互关系的冗余,具体地,交互关系除了上面的两两之间交互关系外,往往还需要构建不同子组群之间的交互关系,以及最后融合为整个组群的交互关系特征,上述多层级上的交互关系是有交集的、非正交的,而最后融合得到的不同特征间集合也难以保证彼此的独立性,故这类交互关系信息不是最简洁的. 上述这两个缺点会一起制约交互关系组群特征的区别性和显著性,进而影响群组行为识别精度的提升.

5.2 存在的问题与展望

虽然群组行为识别取得了显著的效果,但仍然存在不少问题,现总结如下.

(1)不同场景下群组行为类别定义与判别方式的差异性

现有的群组行为识别数据集大致概括为两大类. 其一为场景中的大部分人做相似的行为,如图15所示. 在图15(a)中近镜头处6个女士在"Dancing",则此场景的群组行为属性即定义为"Dancing";类似地,在CAD数据集场景中,图15(b)中近镜头处几个人,除了有两位在"Standing"外,其他成员在"Walking",故该场景群组行为即定义为"Walking". 其二,群组行为的定义取决于场景中的"标志性行为",而忽略其他"大众性平淡无奇的行为",如图16所示. 在图16(a)场景中标志性行为是"两个人在打架",而周围有较多"站立围观者".



(a) 群组行为类型"dancing"

(b) 群组行为类型"walking"

图15 群组行为属性取决于场景中大部分人的相同行为的类别

从信息量的角度定义场景群组行为也应该为"打架",而非"站立",因为"打架"行为是标志性的,是高信息量的;类似地,在Volleyball数据集场景中,如图16(b)所示,左边球员"扣球"行为是Volleyball场景中的"标志性行为",而其他球员大都在"Waiting"和"Standing",同样地,该高信息量的"Spiking"也应该定义为此时的群组行为类别.

总而言之,目前群组行为类别根据不同场景可以分为如上两类,在进行算法验证时也是按照该标准进行群组行为属性的识别. 需要注意的是,如果把诸如"CAD"和"Volleyball"两种群组行为定义完全不同的数据集,同时用来测试某个算法性能时,群组行为的判别方式也应该区别对待. 另外,针对第一类(图15)的群组行为,由于组群的构成有一定随机性,组群成员之间基本不发生交互关系,因此,基于"无交互关系建模的群

组行为识别方法”更适合对其识别,并且,还减少了对“交互关系建模”的计算负荷,提升了识别速度. 而针对第二类(图16)的群组行为识别,其显著特征是组群的构成不具有“随机性的”,而是有“组织性的”,成员之间彼此有分工与合作,因此,“基于交互关系建模的群组行为识别方法”更适合该类情况下的群组行为识别任务.

(*a*)　Fighting

(*b*)　Spiking

图16　群组行为的属性取决于场景中“标志性行为”的类别

(2)成员之间交互关系强弱度量的不统一性以及交互关系属性的多样性

组群中成员交互关系建模包含两层含义即属性和强弱. 目前的算法主要是针对交互关系强弱的定量分析较多,如文献[59]认为同一个群组中具有相近位置和相似特征信息的两个人,拥有较强的交互关系;Ehsanpour等[31]则是通过图注意力模型衡量成员间交互关系的强弱. 这些类似的算法各有不同的交互关系强弱衡量准则,而度量方法差别也很大.

相比交互关系强弱的度量,对交互关系属性的甄别更为重要. 如林晓萌等[66]将组群成员的交互关系属性分为“合作”与“竞争”两种类别,并借鉴情感识别模型Bert网络,利用其能够识别人脸表情类别中的“Positive”与“Negative”特性,用来判别成员的交互关系属性是“合作”还是“竞争”,并同时度量其强度. Azorin-Lopez等[34]依据子组内部成员的运动轨迹计算彼此之间的4种交互信息即关联位移(Coherent Displacement)、非关联上移(Incoherent Up)、非关联下移(Incoherent down)、反向程度(Opposite),并将组间交互关系分为4种属性即组间相干性(Coherence of the group)、组间无关性(Incoherence of the group)、组间吸引度(Attraction)、组间排斥度(Repulsion). 可见交互关系建模是一个“私人定制”过程,可以有不同的交互关系属性定义,也有仅仅对交互关系强弱的不同度量准则. 总的说来,交互关系描述应该先定义其属性,再度量其相应大小,这样定性/定量同时描述出来的组群交互关系才是完备的.

(3)组群结构的时变性

多个人体目标或许本身就是一个整体;或许只有其中的若干成员产生交互关系构成场景中的一个组群,而其他在场人体目标仅仅是无关的过客. 另外,组群成员的交互关系也具有一定的随机性,会随着时间的推移发生改变,导致组群结构也随之变化. 这些问题可以归纳为组群结构化分组、组群结构的动态化维护.

目前,组群结构化分组的方法大都是聚类算法,其依据的信息主要是组群中个体的运动属性、彼此空间距离进行聚类,这些算法的分组精度相对较差. 后期的分组方法多是依据图模型,根据成员交互关系的强弱和交互关系属性(如合作/竞争关系属性、关联/非关联、反向程度等)进行分组,这类分组算法更符合场景中的实际情况. 但这需要对组群交互关系实时性描述,进而根据交互关系的密切程度增减子组群内成员的数量.

(4)全监督/弱监督学习在群组行为识别应用上的不平衡

虽然基于全监督的群组行为识别已经取得了显著的效果,但全监督算法最大的问题是依赖数据集繁琐的人工标注. 而群组行为数据集在采集和制作时,其标注代价相比较单人行为数据集要高出许多,主要是因为群组行为数据集标签的种类和数量都是繁多的,尤其是群组行为中由于参与的成员较多,并且所有成员均需要标注,更甚者是同一成员在不同帧中需要进行反复标注,从而大大增加了工作量,严重阻碍了群组行为监督学习算法的开发. 为了解决上述问题,许多学者转向弱监督算法进行研究,并为群组行为识别方法提供了一种新的思路.

弱监督方法能利用简单易用的视频级标注替代复杂多样的全信息标注,迂回实现群组行为的识别. Zhang等[43]提出了一种快速弱监督深度学习算法用于组群活动识别,为了实现快速推理,其将成员目标检测和弱监督群组行为推理通过共享卷积层的方式得以同步实现,即通过损失函数联合学习这两个任务,从而更有效地过滤掉无关的成员干扰;对于弱监督学习的实现,该算法提出了一种能直接挖掘成员与组群之间交互关系的潜在嵌入式方案,避免了繁琐的需要成员行为标签信息才可建立的成员之间交互关系建模这一环节,不仅实现了群组行为识别还提高了运行速度,其处理帧率为22.65 fps,在很大程度上使群组行为识别更接近实时应用.

另外,弱监督算法还能够利用部分已标注的数据实现对数据集的扩充.Gammulle 等[9]能够利用GAN网络中的生成器产生与原有的已标注的数据集相似的、无标注的噪声数据,并利用判别器判别数据集是否为真,达到对数据集扩充的目的,并实现了对大数据集的弱监督方式的群组行为识别.

虽然弱监督算法具有快速处理数据、节省人力资源等优点,但也产生了一定的问题.例如上述算法[9]的多层次序列GAN网络对群组的行为识别时,产生的噪声样本虽然扩大了数据集,但对于硬件设备性能的要求更高,算力成本更大;更甚者是新样本中的噪声容易干扰分类器训练,降低识别效果.

因此,如果仅用弱监督算法实现群组行为识别,往往导致精度不高.一种有效方法是将弱监督学习与聚类算法、半监督主动学习结合使用,以满足不同场景的需求.如Li[67]提出了一种无监督训练和稀疏监督指导相结合的行为识别方法,其包括两个主要组件.第一个通过编码-解码器RNN来学习获得未标记动作序列的潜在表示;第二个组件根据聚类和半监督分类,进而主动学习上一步中的未标记序列.合并这两步的习得数据完成模型训练,实现行为识别.

(5)视角变化以及场景因素对群组行为识别的影响

视角变化会导致人体姿态发生明显变化,必然会影响成员行为识别,进而影响群组行为判断.针对该问题的统一解决方法就是对多视角下同一行为的样本序列进行学习,获得所谓的"视角无关行描述符",进而达到对不同视角的包容和兼容.丰艳等[68]利用对视角变化不敏感的骨架信息作为输入,首先通过特定视角子网学习每个视角序列的判别性特征,同时利用空域注意力和时域注意力模块分别重点关注关键关节点和关键帧;然后特定视角子网的输出特征作为公共子网的输入,通过公共子网进一步学习角度无关性特征;最后输出行为分类结果.类似地,吴培良等[69]提出一种视角无关的时空关联深度视频行为识别方法,首先,运用深度卷积神经网络的全连接层将不同视角下的人体姿态映射到与视角无关的高维空间,以构建空间域下深度行为视频的人体姿态模型;其次,考虑视频序列帧之间的时空相关性,在每个神经元激活的时间序列中分段应用时间等级池化函数,实现对视频时间子序列的编码;然后,将傅里叶时间金字塔算法作用于每一个池化后的时间序列,并加以连接产生最终的角度无关性时空特征表示.

此外,场景其他信息如背景、光照变化、遮挡、相机运动等因素,在识别群组行为时也会有一定的影响.针对场景信息的应用,可以构建场景时空结构上下文描述符,进而实现成员与场景之间的关系推理.Deng等[70]将场景作为一个结点D,与成员结点A,B,C共同构建组群交互关系网络,如图17所示,其中成员A与场景的交互关系由彼此的交互信息模块AD和DA计算,同理,其他成员与场景的交互关系也可以类似得到.这样通过引入场景结点,达到扩展组群全局时空信息描述的维度,进而提升对场景信息的融合利用.

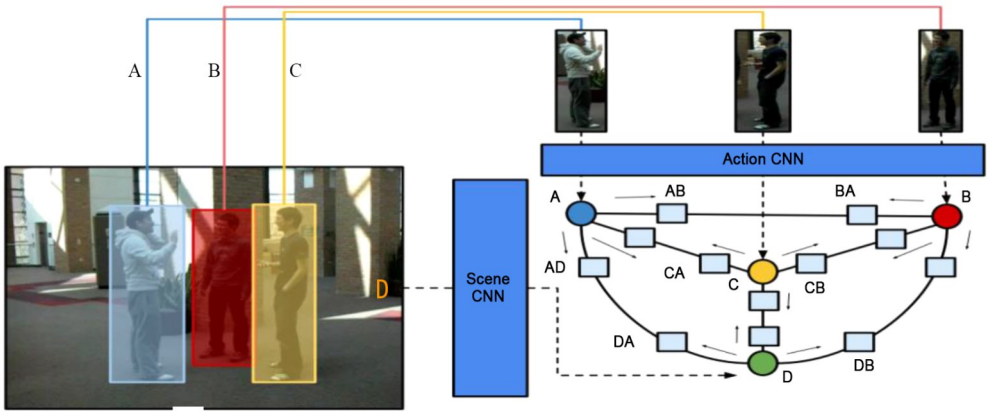


图17 构建场景-成员交互关系提升对场景信息的利用[70]

参考文献

[1] SIMONYAN K, ZISSERMAN A. Two-Stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 27: 568-576.

[2] BORJA-BORJA L F, AZORIN-LOPEZ J, SAVAL-CALVO M A, et al. Deep learning architecture for group activity recognition using description of local motions[C]//2020 International Joint Conference on Neural Networks. Glasgow: IEEE, 2020: 1-8.

[3] ZALLUHOGLU C, IKIZLER-CINBIS N. Region based

multi-stream convolutional neural networks for collective activity recognition[J]. Journal of Visual Communication and Image Representation, 2019, (60): 170-179.

[4] AZAR S M, ATIGH M G, NICKABADI A. A multi-stream convolutional neural network framework for group activity recognition[EB/OL]. (2018-12-26) [2021-10-09]. https://arxiv.org/abs/1812.10328.

[5] 王传旭, 胡小悦, 孟唯佳, 等. 基于多流架构与长短时记忆网络的组群行为识别方法研究[J]. 电子学报, 2020, 48(4): 178-185.
WANG C X, HU X Y, MENG W J, et al. Research on group behavior recognition method based on multi-stream architecture and long short-term memory network[J]. Acta Electronica Sinica, 2020, 48(4): 800-807.

[6] IBRAHIM M, MURALIDHARAN S, DENG Z, et al. A hierarchical deep temporal model for group activity recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 1971-1980.

[7] TAKAMASA T, YASUHIRO K, et al. Football action recognition using hierarchical LSTM[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017: 155-163.

[8] KIM P S, LEE D G, LEE S W. Discriminative context learning with gated recurrent unit for group activity recognition[J]. 2017 Pattern Recognition, 2018, 76: 149-161.

[9] GAMMULLE H, DENMAN S, SRIDHARAN S, et al. Multi-level sequence GAN for group activity recognition [C]//2018 ACCV Computer Vision. Perth: ACCV, 2018: 331-346.

[10] XIN L, CHUAH M C. SBGAR: Semantics based group activity recognition[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2895-2904.

[11] CHOI W, SHAHID K, SAVARESE S. What are they doing?: Collective activity classification using spatiotemporal relationship among people[C]//2012 IEEE International Conference on Computer Vision Workshops. Kyoto: IEEE, 2012: 1282-1289.

[12] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// 2015 International Conference on Machine Learning. Lille: ICML, 2015: 2048-2057.

[13] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]// 2016 International Conference on Learning Representations. San Diego: ICLR, 2015: 1713-1717.

[14] YAN S, SMITH J S, LU W, et al. CHAM: Action recognition using convolutional hierarchical attention model [C]//2017 IEEE International Conference on Image Processing. Beijing: ICIP, 2017: 3958-3962.

[15] WANG Y L, WANG S H, TANG J L, et al. Hierarchical attention network for action recognition in videos [EB/OL]. (2016-07-21)[2021-10-09].https://arxiv.org/abs/1607.06416.

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

[17] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 664-676.

[18] LU L, DI H, LU Y, et al. Spatio-temporal attention mechanisms based model for collective activity recognition[J]. Signal Processing Image Communication, 2019, 74: 162-174.

[19] TANG J, SHU X, YAN R, et al. Coherence constrained graph LSTM for group activity recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 44(2): 636-647.

[20] 王传旭, 龚玉婷. 基于注意力机制的群组行为识别方法[J]. 数据采集与处理, 2019, 34(3): 38-45.
WANG C X, GONG Y T. Group activity recognition method based on attention mechanism[J]. Journal of Data Acquisition and Processing, 2019, 34(3): 38-45.

[21] LIU J C, WANG C X, GONG Y T, et al. Deep fully connected model for collective activity recognition[J]. IEEE Access, 2019, 7: 104308-104314.

[22] BOYKOV Y Y, JOLLY M P. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images[C]//2001 Proceedings Eighth IEEE International Conference On Computer Vision. Columbia: ICCV, 2001: 1(105-112).

[23] CHENG Z, QIN L, HUANG Q, et al. Group activity recognition by Gaussian processes estimation[C]//2010 International Conference on Pattern Recognition. Istanbul: ICPR, 2010: 3228-3231.

[24] ZHANG Y, GE W, CHANG M C, et al. Group context learning for event recognition[C]//2012 Proceedings of the IEEE Workshop on the Applications of Computer Vision. Breckenridge: IEEE 2012: 249-255.

[25] LAN T. Discriminative latent models for recognizing contextual group activities[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(8): 1549-1562.

[26] QI M, JIE Q, LI A, et al. StagNet: An attentive semantic RNN for group activity recognition[J]. 2020 IEEE Transactions on Circuits and Systems for Video Technology. 2020, 30(2): 549-565.

[27] IBRAHIM M S, MORI G. Hierarchical relational networks for group activity recognition and retrieval[C]// 2018 European Conference on Computer Vision. Munich: ECCV, 2018: 742-758.

[28] XU D, FU H, WU L, et al. Group activity recognition by using effective multiple modality relation representation with temporal-spatial attention[J]. IEEE Access, 2020, (99): 1.

[29] SHU X, ZHANG L, SUN Y, et al. Host-Parasite: Graph LSTM-in-LSTM for group activity recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, (99): 1-12.

[30] 丰艳, 张甜甜, 王传旭. 基于伪3D残差网络与交互关系建模的群组行为识别方法[J]. 电子学报, 2020, 48(7): 1269-1275.
FENG Y, ZHANG T T, WANG C X. Group activity recognition method based on pseudo 3D residual network and interaction modeling[J]. Acta Electronica Sinica, 2020, 48(7): 1269-1275.

[31] EHSANPOUR M, ABEDIN A, SALEH J SHI F, et al. Joint learning of social groups, individuals action and subgroup activities in videos[C]//European Conference on Computer Vision. Virtual Conference. Glasgow: IEEE, 2020: 177-195.

[32] SUN L, AI H Z, et al. Localizing activity groups in videos [J]. Comput. 2016, 144: 144-154.

[33] YIN Y, YANG G, JIN X, et al. Small group human activity recognition[C]//2012 The 19th IEEE International Conference on Image Processing. Florida: ICIP, 2012: 2709-2712.

[34] AZORIN-LOPEZ J, SAVAL-CALVO M, FUSTER-GUILLO A, et al. Group activity description and recognition based on trajectory analysis and neural networks[C]// International Joint Conference on Neural Networks. Vancouver: IEEE 2016: 1585-1592.

[35] TRAN K N, GALA A, KAKADIARIS I A, et al. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling[J]. Pattern Recognition Letters, 2014, 44: 49-57.

[36] ZHANG C, YANG X K, ZHU J, et al. Parsing collective behaviors by hierarchical model with varying structure [C]//2012 The 20th ACM International Conference on Multimedia. Nara: ACM, 2012: 1085-1088.

[37] YAN R, TANG J, SHU X, et al. Participation-contributed temporal dynamic model for group activity recognition [C]//The 26th ACM International Conference on Multimedia. Seoul: ACM, 2018: 1292-1300.

[38] DENG Z, ZHAI M, CHEN L, et al. Deep structured models for group activity recognition[C]//2015 The British Machine Vision Conference. Swansea: BMVC, 2015: 1-12.

[39] DENG Z W, ARASH V, HU H X, et al. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 4772-4781.

[40] WU J C, WANG L M, et al. Learning actor relation graphs for group activity recognition[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition. California: CVPR, 2019: 9964-9974.

[41] KUANG Z J, TIE X R. Improved actor relation graph based group activity recognition[EB/OL]. (2020-12-29) [2021-10-09]. https://arxiv.org/abs/2010.12968v2.

[42] HU G, CUI B, HE Y, et al. Progressive relation learning for group activity recognition[C]//2020 Conference on Computer Vision and Pattern Recognition. Virtual Conference: CVPR, 2020: 980-989.

[43] ZHANG P Z, TANG Y Y, HU J F, et al. Fast collective activity recognition under weak supervision[J]. IEEE Transactions on Image Processing, 2020, 29(1): 29-43.

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

[45] TANG Y, WANG Z, LI P, et al. Mining semantics-preserving attention for group activity recognition[C]//2018 The 26th ACM international conference on Multimedia. Seoul: ACM, 2018: 1283-1291.

[46] YANG F K, YIN W J, et al. Group Behavior Recognition Using Attention-and Graph-Based Neural Networks[C]// the 24th European Conference on Artificial Intelligence. Santiago: IEEE, 2020: 1626-1633.

[47] NI B, YAN S, KASSIM A A. Recognizing human group

activities with localized causalities[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Florida: IEEE, 2009: 1470-1477.

[48] BLUNSDEN S J, FISHER R B. The BEHAVE video dataset: Ground truth video for multi-person[J]. Annals of the BMVA, 2010 (4): 1-11.

[49] FABIO Z, TIBERIO U, et al. Learning group activities from skeletons without individual action labels[C]//2021 International Conference on Pattern Recognition. Taichung: ICPR, 2021: 10412-10417.

[50] XU K, BA J, KIROS R, et al. Show, Attend and tell: Neural image caption generation with visual attention[J]. Computer Science, 2015 (37): 2048-2057.

[51] PEI D X, LI A, et al. Group activity recognition by exploiting position distribution and appearance relation[C]// 2021 International Conference on Multimedia Modelin. Manchester: ICMM, 2021: 123-135.

[52] YAN R, XIE L X, et al. Social adaptive module for weakly-supervised group activity recognition[C]//2020 European Conference on Computer Vision. Virtual Conference: ECCV, 2020: 208-224.

[53] ZALLUHOGLU C, IKIZLER-CINBIS N. Collective sports: A multi-task dataset for collective activity recognition[J]. Image and Vision Computing, 2020, (94): 103870.

[54] CHOI W, SAVARESE S. A unified framework for multi-target tracking and collective activity recognition[C]// 2012 European Conference on Computer Vision(ECCV). Florida: ECCV, 2012: 215-230.

[55] LAN T, WANG Y, et al. Discriminative latent models for recognizing contextual group activities[J]. 2012 IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(8), 1549-1562.

[56] DENG Z, VAHDAT A, HU H, et al. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 4772-4781.

[57] FENG Y Q, SHAN S M, et al. DRGCN: Deep relation gcn for group activity recognition[C]//2020 International Conference on Neural Information Processing. Transtations on Multimedin: ICONIP, 2020: 361-368.

[58] LU L H, LU Y, et al. GAIM: Graph attention interaction model for collective activity recognition[J]. IEEE Transtations on Multimedin, 2020, 22(2): 524-539.

[59] SINA M A, MINA G A, et al. Convolutional relational machine for group activity recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7892-7901.

[60] GAVRILYUK K, SANFORD R, JAVAN M, et al. Actor-transformers for group activity recognition[C]//2020 Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 839-848.

[61] SOVAN B, JUERGEN G. Structural recurrent neural network(SRNN) for group activity analysis[C]//Winter Conference on Applications of Computer Vision. Lake Tahoe: WCACV, 2018: 1625-1632.

[62] AZAR S M, ATIGH M G, NICKABADI A. A multi-stream convolutional neural network framework for group activity recognition[EB/OL]. (2018-12-26) [2021-10-09]. https://arxiv.org/abs/1812.10328.

[63] WANG M, NI B, YANG X. Recurrent modeling of interaction context for collective activity recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 7408-7416.

[64] LU L H, LU Y, et al. Learning multi-level interaction relations and feature representations for group activity recognition[C]//ACM International Conference on Multimedia. Chengdu: ACM, 2021: 617-628.

[65] LAN T, SIGAL L, MORI G. Social roles in hierarchical models for human activity recognition[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition [C]. Providence: IEEE, 2012: 1354−1361.

[66] 林晓萌 . 基于图模型和深度学习网络的群组行为识别算法研究[D]. 青岛: 青岛科技大学, 2021.
Lin X M. Group Activity Recognition Research Based on Graph Model and Deep Learning[D]. Qingdao: Qingdao University of Science and Technology, 2021.

[67] LI J, SHLIZERMAN E. Sparse semi-supervised action recognition with active learning[EB/OL]. (2020-12-03) [2021-10-09]. https://arxiv.org/abs/2012.01740.

[68] 丰艳, 李鸽, 原春锋, 等 . 基于时空注意力深度网络的视角无关性骨架行为识别[J]. 计算机辅助设计与图形学学报, 2018, 30(12): 2271-2277.
FENG Y, LI G, YUAN C F, et al. Spatio-temporal attention deep network for skeleton based view-invariant human action recognition[J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(12): 2271-2277.

[69] 吴培良, 杨霄, 毛秉毅, 孔令富, 侯增广 . 一种视角无关的时空关联深度视频行为识别方法[J]. 电子与信息学报, 2019, 41(4): 904-910.
WU P L, YANG X, KONG L F, et al. A perspective-independent method for behavior recognition in depth video

via temporal-spatial correlating[J]. Jorney of Electronic & Information Technology, 2019, 41(4): 904-910.

[70] DENG, Z W, ARASH V, HU H X, et al. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition[C]//2016 the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 4772-4781.

作者简介

邓海刚 男,1985年8月生,山东菏泽人.现为哈尔滨工业大学仪器科学与工程学院博士研究生.主要研究方向为仪器科学与技术.
E-mail: 307140082@qq.com

王传旭 男,1968年生,山东邹城人.现为青岛科技大学信息科学技术学院教授、硕士生导师.主要研究方向为计算机视觉与模式识别.
E-mail: wangchuanxu_qd@qust.edu.cn

李成伟 男,1963年生,黑龙江哈尔滨人.现为哈尔滨工业大学仪器科学与工程学院教授、博士生导师.主要研究方向为仪器科学与技术.
E-mail: chengweili@hit.edu.cn

林晓萌 女,1995年11月生,山东潍坊人.青岛科技大学信息科学技术学院硕士.研究方向为计算机视觉与模式识别.
E-mail: 1104612139@qq.com