Granger因果关系时空图推理的群体行为分析

谢 昭[1),2),3)] 李 骏[3)] 吴克伟[1),2),3)] 焦 畅[3)]

[1)](大数据知识工程教育部重点实验室(合肥工业大学) 合肥 230601)

[2)](工业安全与应急技术安徽省重点实验室(合肥工业大学) 合肥 230601)

[3)](合肥工业大学计算机与信息学院 合肥 230601)

摘 要 因果关系普遍存在于群体交互行为中,体现出主动体行为对被动体行为的有向影响.因果关系检测的难点在于交互双方的行为具有复杂的时间动态性.现有方法使用循环神经网络,来描述交互关系的时间变化特性,并使用时间注意力机制,来描述时间依赖关系.上述方法忽视了对多人依赖关系的分析,难以区分交互双方中的主动行为者和被动行为者.本文设计了一种基于Granger因果关系的时空图推理模型,来学习交互双方的主动和被动关系.为了实现Granger因果关系检测,该模型对单个个体时序特征进行自回归建模,来描述行为对个体自己的依赖.该模型对两个个体时序特征进行相关回归建模,来描述行为对两个个体的依赖.该模型通过比较自回归误差和相关回归误差,当自回归误差明显大于相关回归误差,则说明相关个体改变了另一方个体的行为特征,从而检测出相关个体为主动个体,另一方为被动个体.相关回归模型考虑了多种时间延迟量的两个个体的时序特征序列,用于学习两个个体之间行为的时间延迟量.该时间延迟量用于将主动个体时间特征与被动个体时间特征进行对齐.时间对齐后的主动个体特征提供了被动个体的时间和空间上下文特征,并与被动个体特征进行通道级的融合.为了充分描述个体之间的外观模式,位置约束,因果关系的交互关系,该模型构建多尺度外观的因果图,并使用图推理学习融合上下文的个体特征和群体特征.本文对Granger因果关系检测进行消融分析,并说明时间延迟量,交互融合通道比例,多尺度图推理,能够有效改善个体特征、群体特征的描述能力.本文方法在Volleyball和Collective Activity数据集上优于现有群体行为识别方法.本文的可视化结果说明Granger因果关系可以捕获群体中关键的交互关系.

关键词 群体行为识别;Granger因果关系;时间延迟依赖;时空上下文;图卷积推理

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2023.00856

Spatial-temporal Graph Inference with Granger Causality Relation for Group Activity Analysis

XIE Zhao[1),2),3)] LI Jun[3)] WU Ke-Wei[1),2),3)] JIAO Chang[3)]

[1)](*Key Laboratory of Knowledge Engineering with Big Data*, *Hefei University of Technology*, *Ministry of Education*, *Hefei* 230601)

[2)](*Anhui Province Key Laboratory of Industry Safety and Emergency Technology* (*Hefei University of Technology*), *Hefei* 230601)

[3)](*School of Computer Science and Information Engineering*, *Hefei University of Technology*, *Hefei* 230601)

Abstract Causality reflects the directional effect from the active actor to the passive actor and commonly exists in group interactions. The difficulty in causality detection lies in the complex temporal dynamics of sequential features of the interacting actors. Existing methods use recurrent neural networks to describe the temporal dynamics of the interaction relations. Some methods use

收稿日期:2022-03-29;在线发布日期:2022-11-29.本课题得到安徽省重点研究与开发计划(202004d07020004)、安徽省自然科学基金项目(2108085MF203)、中央高校基本科研业务费专项资金资助(PA2021GDSK0072,JZ2021HGQA0219)资助.**谢 昭**,博士,副研究员,主要研究方向为计算机视觉,图像分析与理解,Email:xiezhao@hfut.edu.cn.**李 骏**,硕士研究生,主要研究方向为计算机视觉,图像分析与理解.**吴克伟**(通信作者),博士,副研究员,主要研究方向为计算机视觉,图像分析与理解,Email:wu_kewei1984@163.com.**焦 畅**,硕士研究生,主要研究方向为计算机视觉,图像分析与理解.

temporal attention mechanisms to describe temporal dependencies. They neglect to analyze the dependency between two actors, and are hard to distinguish the active actor and passive actor in the interaction. In this work, we design a Granger causality-based spatiotemporal graph model to learn the active-passive relations between interacting actors. To detect the Granger causality, the model designs an autoregression function for single individual temporal sequential features to describe the dependence of action on the individual itself. The model designs a correlative regression function for two individual temporal sequential features to describe the dependence of action on two individuals. The model detects the correlative individual as an active individual and the other as a passive individual by comparing the autoregressive error with the correlative regression error, when the autoregressive error is significantly larger than the correlative regression error, which indicates that the correlative individuals change the action of the other individual. The correlative regression function considers two individual temporal sequential features with multiple time delays, which can be used to learn the amount of time delay for actions between two individuals. This time delay amount is used to align the active individual time features with the passive individual time features. The temporally aligned active individual features provide the temporal and spatial contextual features of the passive individual and are fused with the passive individual features at the channel-wise level. The model constructs causal graphs of multi-scale spatiotemporal features to fully describe the interaction between appearance patterns, location constraints, and Granger causality among individuals. The multi-scale causal graph embeds the contextual features into the individual features and group features with graph inference. Experiments compare with state-of-the-art methods on Volleyball and Collective Activity datasets. (1) The spatial relation pooling model, such as Hierarchical Deep Temporal Model (HDTM). (2) The spatial relation graph models include Social Scene Understanding model (SSU), Convolutional Relational Machine (CRM), Hierarchical Relational Machine model (HRN), Actor Relation Graph model (ARG), Graph Attention Interaction Model (GAIM), Actor-Transformer (AT), Position Distribution and Appearance Relation model (PDAR), Multi-level Interaction Relation model (MLIR). (3) The spatial-temporal relation model includes Confidence-Energy Recurrent Network (CERN), spatial-temporal attentive graph network (stagNet), Progressive Relation Learning model (PRL), Graph LSTM-in-LSTM model (GLIL), Visual Context model (VC), GroupFormer (GF), Partial context embedding (PCE), Coherence Constrained Graph LSTM (CCGLSTM). Our Granger causality-based relation detector can describe the relations between potential active actors and passive actors. The channel-wise temporal causality graph inference module can enhance the feature of the passive actor by fusing a temporal delayed feature of the active actor. The graph model use Granger causality relation can describe effective interaction between actors and provide contextual features for group activity recognition.

Keywords group activity recognition; Granger causality relation; temporal delay dependency; spatial-temporal context; graph convolutional inference

1 引 言

群体行为识别通过对多人场所的视频进行分析，并识别其突发性群体行为，有利于维护公共场所安全，避免人员和财产损失，被广泛应用于视频监控、视频摘要、视频检索等领域[1-2]. 与个体行为识别不同，群体行为识别的关键在于描述群体行为的交互关系，该交互关系提供丰富的行为的上下文特征，具有重要的研究价值.然而，在真实视频中，群体中

的多人关系是无标记的，具有多变的时空特性，大量复杂的时空要素干扰了多人关系的估计.因此，群体行为识别仍然是对时空数据敏感的病态问题.

群体行为识别的关键是如何实现群体交互关系的建模.现有模型关注于使用图卷积(GCN：Graphical Convolutional Network)模型分析多人交互特征.模型使用的特征包括外观特征、位置特征、运动特征.外观特征可以描述人体姿态信息，用于分析多人的交互关系[3-5].位置信息可以度量人体之间的距离.由于不同行为发生的人体距离是不同的，距离约束可以限制交互关系类型[6-7].外观和位置特征可以学习群体结构的静态空间特性，但是，难以描述群体结构的动态时间特性.对时序外观特征使用循环神经网络建模，能够捕获一定的人体运动信息，可以区分复杂运动的交互关系[8-10].现有方法认为个体之间的交互关系是对称的，对交互双方的约束是相同的，不能有效描述交互行为双方的主动和被动关系，因此，仍然无法有效解释主动行为体到被动行为体的有向因果关系.

图1给出了群体行为中的多人交互关系，多人的交互满足特定的行为因果关系，例如，扣球(因)和拦网人员(果)的交互，一传手(因)和二传手(果)的交互.从图1中可以看出群体中的多人关系具有以下三个特点.

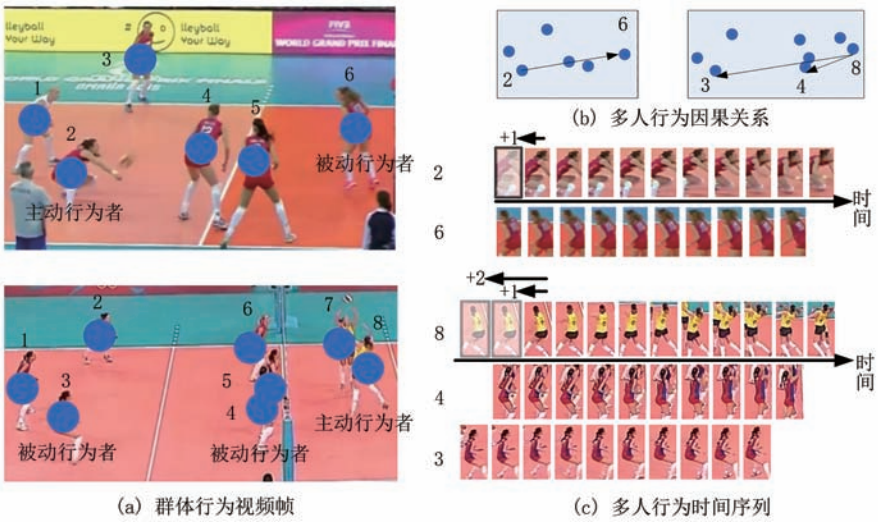


图1 群体行为分析中的多人因果关系

(1)多人行为的因果关系是有向的，交互双方被区分为主动和被动人员.主动行为者(图1(a)中上方2号传球手，下方8号扣球手)能够自主地决定自己的行为，而被动行为者(图1(a)中上方6号二传人员，下方4号拦网和3号垫球人员)会关注主动行为者并改变自己的行为.这种主动被动关系可以被描述为多人行为的因果依赖，如图1(b)，其中主动行为者的原因是自己，而被动行为者的原因包括自己和环境中的人员.通过对图1(a)的观察，我们发现因果双方的外观特征是复杂的，使用单个视频帧的外观特征，难以区分多个视频帧的动态行为因果关系.本文认为这种动态行为因果是群体行为中的关键交互，需要分析多个视频帧的行为特征(图1(c)).

(2)多人行为的因果关系时间不同步.由于被动行为者感受到交互行为的快慢不同.主动者在发出行为后，需要经过时间延迟，被动者的行为才会受到改变.交互双方具有不同的空间距离，进一步加剧了时间延迟效应.图1(c)中，下方垫球人员的反应速度，比拦网人员的反应速度要迟，这是因为球的运动是先遇到拦网人员，后遇到垫球人员，提前执行动作可能会造成失误.如果不考虑时间延迟效应，难以实现交互双方行为的时间匹配.使用未匹配的双方特征来估计边关系，会降低边关系的强度，从而降低了对关键交互行为的响应.

(3)多人行为的因果关系具有时空约束.不同交互行为具有不同的空间距离.例如，图1(a)中一传和二传人员的距离，扣球和拦网人员的距离，扣球和垫球人员的距离是不同的.距离约束可以筛选出不同类型的交互行为.通过有限的距离筛选，还可以减少环境中无关行为的干扰.使用距离约束来挑选出合理的交互行为双方，可以发现群体中的关键交互关系.因此，需要根据上述群体因果关系的三个特性，来重新设计具有因果关系的图卷积模型，才能捕获合理的群体交互关系.

为了解决上述问题，本文提出一种基于Granger因果时空图推理的群体行为识别方法.Granger因果分析利用序列信号自回归关系，学习历史特征对当前特征的影响.不同于因果有向图模型[11]，本文旨在分析不同个体形成的时间特征序列表达之间的影响，描述行为的因果关系，是对Granger自回归模型的扩展.利用因果关系可以发现主动行为者和被动行为者，并联合主动行为者和被动行为者的特征可以更好地描述被动行为者的特征，从而降低被动行为者行为识别的不确定性.在获得更可靠的被动行为者特征的情况下，综合分析多人交互关系，从而得到可靠的群体行为类别.

本文方法将群体交互关系建模拆分为三个子任务.首先，为了实现多人因果关系检测，本文利用Granger因果条件分析多人行为的相关性[12-13].具体来说，Granger使用过去时刻的状态来回归当前时刻的状态，如果回归误差小，则说明过去状态是当前状态的Granger原因.Granger因果条件可以拓展用于分析多人之间的关系，即如果环境中多人的过去时刻状态来回归个体的当前状态，其回归误差如果小于自回归误差，则说明环境是个体状态的Granger原因.因此，本文在多人时序特征基础上，构建了Granger因果检测模块，该模块包括时序特征的自回归模型，环境时序特征的相关回归模型和Granger原因测度估计.根据Granger检测结果构建Granger因果关系图.

其次，为了解决多人行为的时间延迟问题，本文设计了时间延迟图来记录群体交互关系的时间延迟关系.为了自适应地估计时间延迟量，将Granger因果检测的固定时间窗，拓展为多个时间平移的时间窗，并使用最小相关回归误差，来确定时间延迟图中交互关系的延迟量.为了实现因果双方的时间同步，本文使用时间延迟量平移时序特征，以获得时间同步后的时序特征.为了同时考虑当前特征和同步特征的时间上下文，采用通道级方式将Granger因果双方特征进行融合，以同时描述特征的时间延迟关系和因果关系.

最后，为了解决多人行为的多时空交互约束，本文设计多尺度空间条件来筛选交互行为双方个体，通过融合多尺度空间关系和因果关系，来实现因果关系的多时空约束.不同于现有的图卷积模型，本文的多时空因果图卷积中的节点特征和边特征都使用了因果关系.具体来说，节点特征是时间同步后的特征，边特征是Granger因果条件.因此，本文的多时空因果图卷积能够更好地描述交互行为，产生合理的群体行为交互关系.

本文的主要贡献如下：(1)本文针对多人之间具有的复杂时空关系，提出一种基于Granger因果时空图推理的群体行为识别.该模型使用多人的Granger因果检测，同时解决了时空图推理中的三个问题：多人行为的因果有向依赖检测，时间不同步检测；时空因果图融合.(2)针对单个视频帧的外观特征无法描述时序动态特征之间的因果有向依赖，通过分析个人时序行为受到空间上下文时序行为的主动被动关系，本文设计了一种时序的Granger因果关系模块.该模块对行为特征进行时序回归建模，估计自回归误差和空间上下文的相关回归误差.为了分析两者之间的差异程度，该模块设计了基于F统计量的Granger因果依赖测度.若两者差异性较大，说明空间上下文明显改变行为特征，则行为具有较强的空间上下文约束.(3)针对多人行为空间交互推理，无法描述行为之间的时间依赖不同步问题，本文设计了一种通道级时间因果图推理模块.该模块利用Granger因果检测，来生成具有时间依赖的空间关系图，并采用通道平移策略，实现时间依赖图的多人特征融合，以生成融合时间上下文的个人行为特征.(4)针对群体的时间因果图，无法描述多人之间的距离约束的问题，本文设计了一种多尺度距离约束的因果关系图.该图模型使用距离和外观来约束多人之间的交互行为，通过不同的图卷积参数，来学习多种线索约束下的多人关系，并实现基于多种线索图卷积推理的个人行为和群体行为表达.

2　现　状

本文将群体行为识别任务拆分为行为特征提取，群体交互关系建模，因果交互关系分析三个子任务.行为特征提取提供个体行为表达，交互关系实现对个体之间关系的描述，因果交互关系进一步分析个体之间交互的主动行为者和被动行为者的关系.

2.1　行为特征提取

群体行为识别需要学习有判决能力的视频特征.行为识别可以将视频帧作为一个整体进行特征学习，以解释场景中外观相关的空间模式，或场景中运动相关的时间模式.早期空间模式使用手工视觉特征算子[14].手工特征不利于表达复杂空间模式，卷积神经网络借助深度学习的反馈传播机制可以学习出行为相关的空间模式.双流网络学习关键帧的RGB图

像特征和光流图像的特征，以识别不同的行为[15].时间分段网络(TSN：Temporal Segment Network)采用多个时间帧的双流特征并对多个时间帧的特征进行加权池化来识别行为[16].然而，TSN的时间加权对于视频帧采用相同的权重.如果需要对视频帧设置不同的时间权重，可以将2D卷积神经网络扩展为3D卷积神经网络.膨胀三维卷积网络(I3D：Inflated 3D ConvNet)将2D CNN参数膨胀后作为3D卷积神经网络的初始化参数[17].Non-local网络估计空间位置权重，可以提供更好的个体特征用于行为识别[18].在I3D特征基础上，学习时间注意力有利于找出视频中的关键帧，学习空间注意力利于发现图像中的关键内容[19].

3D卷积中时间维度的权重仍然是线性，它难以描述序列特征中的复杂时间变化.为了实现更复杂的时间依赖分析，在深度学习框架中，通常采用时间卷积来分析[20]，还可以通过膨胀的时间卷积算子来分析跨时间的依赖[21].时间卷积被用于解决帧播放顺序的问题[22].循环Transformer融合时间位置的编码信息来描述时间依赖[23].视频级行为的时间依赖，包括发生时刻相同的多标签行为依赖，也包括发生时刻不同的行为关系[24].

由于循环神经网络关注于视频帧的状态变化，因此，其提取的特征能更好地反映特征的时间变化.长短期记忆网络(LSTM：Long Short-Term Memory)具有长短时特征变化的学习能力，是循环神经网络中的典型方法.在群体行为识别中，LSTM可以对个体和群体建立独立的模型，来分别描述个体和群体的特征时间变化.个体和群体之间可以建立层次的LSTM网络(HDTM：Hierarchical Deep Temporal Model)[25].社会场景理解网络(SSU：Social Scene Understanding)[26]将两个时间帧特征的位置差值和特征差异作为匹配关系加入循环神经网络，以学习出更可靠的个体和群体的时间模式.

2.2 群体交互关系建模

群体行为识别的细节在于群体中个体之间的交互关系，因此，其群体中的交互关系分析是群体行为识别的关键问题.群体交互关系学习需要依赖于空间关系图模型，图模型中节点表示个体特征，节点之间的边描述个体之间的交互关系.层次关系网络(HRN：Hierarchical Relational Network)使用固定的群体结构来学习交互关系强度[5].角色关系图(ARG：Actor Relation Graph)同时考虑外观特征和位置特征来学习交互关系[27].卷积关系机(CRM：Convolutional Relational Machine)使用多阶段深度特征的误差来优化群体交互关系[28].

图注意力交互模型(GAIM：Graph Attention Interaction Model)将群体节点加入图模型，并利用自注意力同时学习个体之间和个体与群体之间的关系[3].多人特征融合中，可以通过运动特征计算个体注意力并进行个体选择[29].角色变换模型(AT：Actor-Transformer)使用自注意力对空间位置编码特征进行相关性估计，并使用额外的前向多层网络来优化空间位置关系[6].群体模型(GF：GroupFormer)[30]对个体特征进行时空划分，并设计交叉时空匹配的Transformer来学习个体特征的交互关系.多层交互关系(MLIR：Multi-Level Interaction Relation)，通过关键节点选择可以生成不同的群体结构，将不同群体结构视为池化和反池化关系，可以学习出多层次的交互关系[4].研究个体注意力和个体-环境注意力的估计方法，可以实现个体特征的分层注意力池化[1].位置与外观关系网络(PDAR：Position Distribution and Appearance Relation)使用边界框的位置特征可以估计出位置相关性，并构建位置图卷积网络用于群体行为识别[7].

群体交互关系学习依赖于个体特征，上述方法的个体特征使用卷积神经网络学习个体特征.为了进一步描述个体的时间模式，循环神经网络被用于进一步优化卷积神经网络特征，从而获得具有时空特性的个体特征用于描述交互关系.置信度能量循环网络(CERN：Confidence-Energy Recurrent Network)在LSTM的时间动态个体特征基础上构建图模型，并引入了能量层来估计图结构的稳定程度，用于优化图模型的交互关系[31].行为顺序关系学习网络(PRL：Progressive Relation Learning)将关系估计视为Markov Process过程，并采用Action-Reward的强化学习机制来渐进学习交互关系，并将Action操作设计为每个节点的gating操作[9].图LSTM嵌套LSTM网络(GLIL：Graph LSTM in LSTM)[10]在个体之间的关系估计时，研究基于外观和位置的边关系权重，在群体和个体的关系估计时，研究基于群体特征和个体特征的节点权重.时空注意力图网络(stagNet：spatio-temporal attention graph network)在视觉特征以外，对个体行为的语义标记进行编码来描述节点.该网络为节点特征构建RNN来学习其动态模式，并分别学习特征的时间相关性和空间相关性，构建时空注意力图模型[8].视觉上下文(VC：Visual Context)[2]在推理中研究个体

和群体特征来学习个体注意力，研究个体之间关系的估计，同时使用位置编码的点对关系进行图关系约束.部分上下文编码方法（PCE：Partial context embedding）[32]利用Transformer分析视频帧中多个标记(token)之间的关系.一致性图LSTM（CCGLSTM：Coherence Constrained Graph LSTM）认为交互需要时空上下文特征和场景图像全局特征，使用图LSTM来学习时空交互关系，并添加场景图像的全局特征进行群体行为识别[33].

2.3　因果交互关系分析

群体行为的因果交互关系是指个体之间的依赖.不同个体之间的行为依赖，可以描述个体与个体之间的主动与被动关系，例如，物体受到外力而发生的行为[34].主动行为者的控制效果可以描述为与或图的形式[35]，也可以使用贝叶斯进行归纳学习[36].因果推理要区分出有向关系，这不同于无向的统计相关性[11].本文专注于群体中两个个体的主动和被动关系，即研究两个时间序列之间因果关系.Granger理论将时间序列进行自回归建模和相关回归建模，来分析时间序列回归误差[12-13].如果添加环境中的个体，回归误差变小，则说明环境中的个体引起了主动的影响.在脑电EEG时间序列数据中，Granger理论可以通过逆向建模来分析有向的因果关系[37]，通过延迟变量来分析延迟的因果关系[13].Granger因果的强度可以通过统计测试的方式来确定[38].在目标识别任务中，因果关系分析可以用于特征选择和度量学习[39].上述Granger因果关系主要用于脑电数据分析，并没有用于行为特征的分析.为了有效描述不同个体运动之间复杂的有向因果关系，本文将因果关系描述为不同的时间延迟情况，并设计具有时间延迟依赖的Granger模型，以提取群体中的Granger因果关系图.本文还将进一步将Granger因果关系图，拓展为时空因果关系图，并设计时空因果关系的特征融合和特征推理方式，以便最终实现群体行为识别任务.

3　Granger因果关系时空图模型

本文将群体行为特征提取拆分为时间因果检测、时间同步因果融合、多尺度时空因果图推理三个子任务.图2给出了本文的基于Granger因果时空图推理的模型框架.第一阶段，关注于两个个体之间的因果关系提取.本文研究行为特征的回归约束，设计了一种时间延迟的序列回归模块，实现Granger因果关系检测；第二阶段，需要解决两个个体的特征同步问题，以避免不同步特征的时空推理中引入的误差.本文使用检测的因果关系构建因果图模型，该图模型包含有向关系和延迟量，并进一步使用延迟量进行时间同步的交互特征融合.我们的时间同步融合同时考虑了交互个体之间的通道比例，来分析因果双方在因果推理的重要性；第三阶段，将时间因果图拓展到多尺度时空因果图，并学习具有时空特征的图特征.时空因果图，能够额外使用距离来筛选出不同类型的交互关系，以有针对性地学习图推理后的群体行为特征.

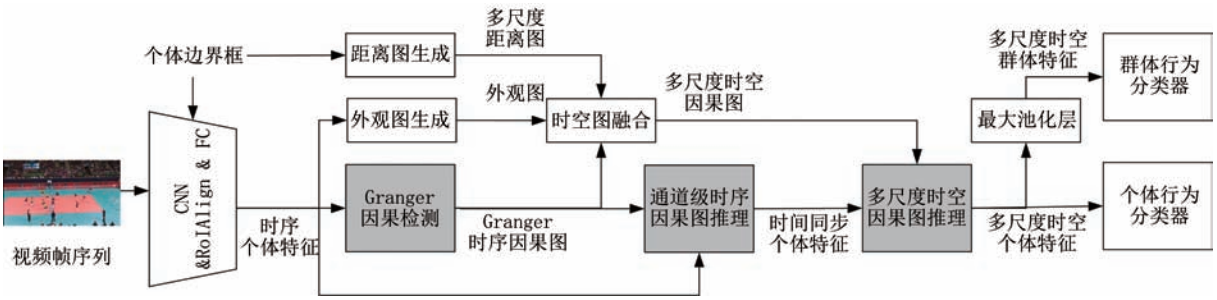


图2　基于Granger因果关系的时空图推理模型框架

3.1　Granger因果检测

Granger因果条件用于判断两个个体行为之间是否存在影响[12-13].如果个体j影响个体i的行为，那么称个体j是个体i的Granger原因，个体i是个体j的Granger结果.由于个体行为具有多变的局部空间交互关系，需要提取局部细节的行为模式，例如，腿部的跑步，跳起动作，躯干的弯腰，手部的击打、抬举动作等.因此，本文利用局部的多通道特征描述不同的局部状态，充分考虑时间序列回归模型在多通道的一致性约束，建立Granger因果检测模型.该模型包括两个子模型，来分别分析个体行为被自己历史状态的约束，和个体行为被环境状态的约束.

为学习个体对自己的约束，本文建立一种时间序列的自回归模型，即通过自身历史时刻的状态预

测出行为最后的状态.为学习环境对个体的约束,本文建立一种时间序列的相关回归模型,在自身特征基础上,引入环境中相关的个体特征,来回归预测出自身的最后状态.如果个体不被环境影响,个体会根据自己的目的,来设计自己的行为;此时,自回归模型可以准确地回归自己的行为,相关回归模型的预测结果不会产生明显的改进.如果个体被环境中相关个体影响,相关个体改变了被动个体的行为;此时,自回归模型的预测误差较大,相关回归模型的预测结果误差较小.

为了实现验证上述的Granger因果关系[12-13],我们进一步引进两个假设条件.原假设:个体j和个体i在行为上是无关的,即个体j不是个体i的Granger原因.对立假设:个体j和个体i在行为上是有关的,即,个体j是个体i的Granger原因.我们还设计了Granger因果关系的假设验证条件用于检测群体中多个体之间的因果关系.

给定视频帧序列和其中个体的标记框,我们使用2D CNN来学习视频帧的特征,并使用ROI Align(Regine of Interesting Align)方法提取个体标记框中的特征.在个体特征提取过程中,由于网络会对图像的分辨率规范化,标记框会无法对齐到缩放后的图像网格上.此时,ROI Align方法使用插值处理,从未对齐的缩放图像网格特征上,估计出缩放后的个体标记框网格的特征,随后使用sum pooling对网格特征进行聚合,获得个体特征向量.我们在ROI Align基础上,额外使用FC层进行特征变换,进一步增强该特征对行为的描述能力.

此时,我们获得视频帧序列的多通道个体特征集合$X=\{x_{i,t,c}\}$,其中i是个体编号,t是时间编号,c是通道编号,该特征集合的每个通道是一个时空立方体.我们认为每个通道所描述的行为模式是相对独立的,并建立时间回归模型来分析每个通道中个体之间的因果关系.如图3所示,本文的Granger因果检测从时空立方体中抽取两个个体特征,并构建自回归模型和相关回归模型.图3上方是单个个体的Granger自回归预测模型$\phi_a(\cdot)$,即利用个体自身历史特征预测个体当前的特征:

$$\begin{aligned} x_{i,k,c} &= \phi_a([x_{i,k-m:k-1,c}],\theta^a_{m,i}) \\ &= \sum_{q=1}^{m} w^a_{m,i,k-q} x_{i,k-q,c} + b^a_{m,i} \end{aligned} \tag{1}$$

其中i是个体编号,k是当前时刻,m是时间窗口宽度,r是时间窗的下标,下标对应的范围是$q=[1:m]$,历史特征的时间窗口是$[k-m:k-1]$.$\theta^a_{m,i}$是自回归模型的参数,具体包括$\theta^a_{m,i}=\{w^a_{m,i,k-q}, b^a_{m,i}\}$分别是自回归模型中的一次项参数和常数项参数.

由于单个视频行为的目标唯一,为了保证多通道的一致性约束,不同通道间应该具有相同的参数,因此,在训练该自回归模型时,其训练集包含所有通道的不同时刻k下的序列样本.自回归采用误差平方和作为损失函数,并采用最小二乘法进行参数学习:

$$\hat{\theta}^a_{m,i}=\arg\min_{\theta^a_{m,i}} \sum_{k,c} \left\| x_{i,k,c} - \phi_a([x_{i,k-m:k-1,c}],\theta^a_{m,i}) \right\|_2^2 \tag{2}$$

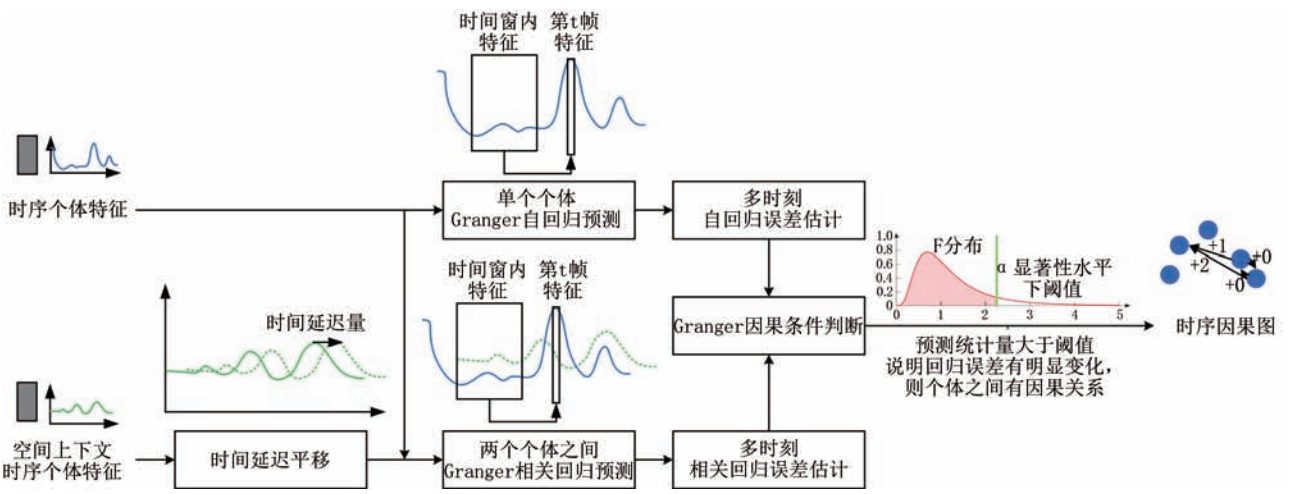


图3 Granger因果检测中的自回归模型和相关回归模型

图3下方是两个个体之间的Granger相关回归预测模型$\phi_r(\cdot)$,即同时考虑当前个体和环境中相关个体的历史特征来预测.相关回归模型考虑时间延迟量来描述个体之间有向影响的延迟现象.

$$x_{i,k,c}=\phi_r([x_{i,k-m:k-1,c},x_{j,k-delay-m:k-delay-1,c}],\theta^{r,j\to i}_{delay,m,j\to i})$$
$$=\sum_{q=1}^{m}w^{r,j\to i}_{delay,m,i,k-q}x_{i,k-q,c}+\sum_{q=1}^{m}w^{r,j\to i}_{delay,m,j,k-delay-q}x_{j,k-delay-q,c}+b^{r,j\to i}_{m,delay} \tag{3}$$

其中i是当前个体编号，j是环境中相关个体编号，k是当前时刻，m是时间窗口宽度，q是时间窗内的帧下标，被动行为者的时间窗口是$[k-m:k-1]$，$delay$是主动行为者的时间延迟量，主动行为者的时间窗口是$[k-delay-m:k-delay-1]$.$\theta^{r,j\to i}_{delay,m}$是相关回归模型的参数，其中$j\to i$表示，分析个体$j$是否是个体$i$的有向Granger原因.该参数包括$\theta^{r,j\to i}_{delay,m}=\{w^{r,j\to i}_{delay,m,i,k-q},w^{r,j\to i}_{delay,m,j,k-delay-q},b^{r,j\to i}_{delay,m}\}$分别是模型中的当前个体的一次项参数、环境中相关个体的一次项参数和常数项参数.

在训练该模型时，为了保证多通道的一致性约束，其训练集包含所有通道的不同时刻k下的时间序列样本.模型采用误差平方和作为损失函数，并采用最小二乘法进行参数学习：

$$\hat{\theta}^{r,j\to i}_{delay,m}=\arg\min_{\theta^{r,j\to i}_{delay,m}}\sum_{k,c}\left\| x_{i,k,c}-\phi_r([x_{i,k-m:k-1,c},x_{j,k-delay-m:k-delay-1,c}],\theta^{r,j\to i}_{delay,m})\right\|_2^2 \tag{4}$$

在利用多通道多时刻数据训练后的自回归模型和相关回归模型基础上，本文进一步分析误差的概率分布，并设计了基于Granger的F分布统计量假设检验模型[10-11].由于服从Gaussian分布的随机变量的平方和分布服从χ^2分布，那么函数形式是平方和的预测误差也服从χ^2分布.Granger的原因条件的原假设是分析两个误差是否具有明显差异，即分析两个样本集合的方差是否相同，此时，采用F分布进行假设验证.个体j是否是个体i的Granger原因的F统计量为：

$$\begin{cases} f^{r,j\to i}_{delay,m}=\dfrac{(ssr^{a}_{m,i}-ssr^{r,j\to i}_{delay,m})/m}{ssr^{r,j\to i}_{m}/(n^{r}_{m}-\upsilon^{r}_{m})} \\ ssr^{a}_{m,i}=\sum_{k,c}\left\| x_{i,k,c}-\phi_a([x_{i,k-m:k-1,c}],\hat{\theta}^{a}_{m,i})\right\|_2^2 \\ ssr^{r,j\to i}_{delay,m}=\sum_{k,c}\left\| x_{i,k,c}-\phi_r([x_{i,k-m:k-1,c},x_{j,k-delay-m:k-delay-1,c}],\hat{\theta}^{r,j\to i}_{delay,m})\right\|_2^2 \end{cases} \tag{5}$$

其中，$ssr^{a}_{m,i}$是时间窗宽度m情况下，个体i在自回归模型的预测误差的平方和(SSR：Sum of Square Residual)，$ssr^{r,j\to i}_{delay,m}$是时间窗宽度m和延迟量$delay$情况下，相关回归模型的预测误差的平方和.该误差求和过程，在处理视频分割中存在两种不同因果关系(先被动后主动)时，会产生大小接近的自回归误差和相关回归误差，从而降低F统计量.为了避免视频分割中存在两种不同的因果关系，在实际处理过程中，应该选择较小的视频分割段，并选择关键帧作为该分割段的时间中心点.

分母是自回归模型的部分，n^{r}_{m}是相关回归模型的样本数量，根据自回归模型的预测误差公式，可以看出样本数量为$n^{r}_{m}=k\cdot c.\upsilon^{r}_{m}$是相关回归模型的自由度，即模型的参数个数.由于自相关模型共包括m个一次项参数和1个常数项参数，则自回归模型的自由度为$\upsilon^{r}_{m}=2m+1$.分子是相关回归模型部分，对于$ssr^{a}_{m,i}-ssr^{r,j\to i}_{delay,m}$来说，也服从$\chi^2$分布，其自由度为相关回归模型和自回归模型的差值，其差值正好是时间延迟量m.此时，我们获得个体j影响个体i情况下的Granger原因F分布统计量，该统计量服从自由度为$(m,k\cdot c-2m-1)$的F分布$f^{r,j\to i}_{delay,m}\sim F(m,k\cdot c-2m-1)$.为了分析不同显著性强度下的多人之间的Granger原因，我们设计Granger原因的概率值：

$$p^{r,j\to i}_{delay,m}=\int_0^{f^{r,j\to i}_{delay,m}}\psi_{F(m,k\cdot c-2m-1)}(z)dz \tag{6}$$

其中，$\psi_{F(m,k\cdot c-2m-1)}(z)$是F分布的概率密度函数，$z$是概率密度函数的随机变量.如果两个模型的预测误差差值越大，则Granger原因F分布统计量越大，则Granger原因的概率值越大，则接受原假设，拒绝对立假设，此时个体j是个体i的Granger原因.

本文提取的Granger原因概率值具有以下3个特点.(1)个体之间的Granger原因是有向的，可以用于描述扣球和拦网行为之间的关系.扣球是主动行为者，拦网是被动行为者，拦网行为需要根据扣球的方式进行调整，包括跳起的位置、高度、姿势等.同时，扣球个体受到拦网个体的影响不是很明显.(2)个体之间的Granger原因是有延迟的，在扣球行为中，延迟量和球的运动路径长度和运行时间有关，具体来说拦网个体需要选择合适的时间起跳，才能和排球同时到达预定的位置，成功实现拦网.本文的Granger原因强度是随着时间延迟量变化的，因此可以根据最大强度来确定个体之间最好的时间延迟关系.(3)由于F分布统计量和显著性水平的关系是非线性的，不利于直接观察群体中交互行为的关系.本文将F分布统计量转化为Granger原因的概率值，更有利于直观描述个体之间的因果强度，以便分析群体中典型交互行为具有何种强度的因果关系.

3.2　通道级时间因果图推理

本文设计的Granger原因强度是随着时间延迟量变化的.时间延迟量过大或过小的情况下，Granger

原因强度会减弱，这也反映了个体特征之间的匹配程度会减弱.因此，本文利用Granger因果图中的时间延迟量，来进行时间同步处理，以减少由于时间不同步引起的特征融合误差.

图4中给出了时序因果图，图中的节点是个体，图中的边是Granger原因关系，边上的箭头从主动行为者指向被动行为者，边上同时带有Granger原因关系强度最大时的时间延迟量.+1表示时间延迟量为1.给定视频序列中多个延迟量的Granger原因概率矩阵，根据矩阵中每对元素，来描述个体之间的关系.为了分析运动的不同时间延迟关系，我们使用延迟量0来分析时间同步的行为关系，使用延迟量1，2，3来描述不同长度的延迟关系，并根据Granger原因的最大概率值确定最佳时间延迟量，有

$$[delay^{*}_{j\to i}, p^{r,j\to i}_{delay^{*}_{j\to i},m}] = \arg\max_{delay} p^{r,j\to i}_{delay,m} \quad (7)$$

其中，$p^{r,j\to i}_{delay^{*}_{j\to i},m}$表示在最佳时间延迟量下$delay^{*}_{j\to i}$的Granger原因强度，此时获得延迟矩阵$D^{*}=\{delay^{*}_{j\to i}\}$.通过设置因果关系阈值，进一步筛选出关键的因果关系：

$$e^{Granger}_{j\to i}=\begin{cases}1 & p^{r,j\to i}_{delay^{*}_{j\to i},m}>\tau\\ 0 & otherwise\end{cases} \quad (8)$$

其中，τ是因果关系阈值，从而获得时序因果图$E^{Granger}=\{e^{Granger}_{j\to i}\}$.在因果图中，本文通过考虑多种延迟量的情况，来分析群体中可能存在的复杂延迟关系.

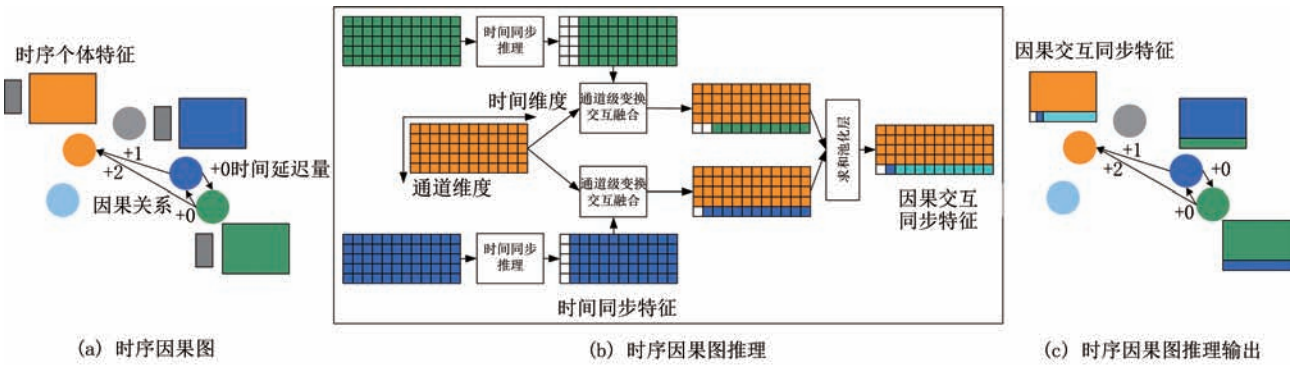


图4 时间延迟特征与通道级时间因果图推理

本文设计因果图推理的目的在于获得时间同步特征，以便更准确地反映特征的时间上下文关系.首先，选择出有因果关系的节点对，并将主动行为者的特征进行时间平移，

$$x^{shift,j\to i}_{i,k,c}=shift(x_{j,k,c}, delay^{*}_{j\to i}) = \begin{cases}x_{j,1,c} & k-delay^{*}_{j\to i}<1\\ x_{j,k-delay^{*}_{j\to i},c} & otherwise\end{cases} \quad (9)$$

其中，根据时间延迟量$delay^{*}_{j\to i}$对主动行为者j的特征进行平移，获得被动行为者i的$x^{shift,j\to i}_{i,k,c}$因果关系上下文特征，k为当前时刻.其中主动行为者j先发生行为，被动行为者i后发生行为，因此，主动行为者的特征需要延迟平移.通过对主动行为者的特征序列，按照时间序列的下标k，向后平移$delay^{*}_{j\to i}$个时间单位.由于平移后初始时刻会出现空缺，对于这个空缺的部分，仍然使用第1帧的特征填充.

除了时间上下文以外，我们在同步过程中还需要考虑两个个体之间融合方式.串联方式可以兼顾两个个体的所有特征，然而，所有通道中存在一些相关性不高的特征，降低了计算效率.因此，本文引入通道降维操作，将因果图推理设计为一种通道级时间同步融合方式：

$$x^{con,j\to i}_{i,k}=concat_{k}(w^{d}_{i}x_{i,k}, w^{d}_{j\to i}x^{shift,j\to i}_{i,k}) \quad (10)$$

其中$x^{con,j\to i}_{i,k}$是时间同步融合特征，该特征既考虑了个体i的特征，也考虑了个体j对个体i的融合，使用$concat_{k}(\cdot)$在维度k上进行串联操作融合两个个体特征.为了分析主动行为在和被动行为者在交互过程中的重要性，设置通道比例因子，对融合后的特征通道进行约束.w^{d}_{i}是个体i的通道降维参数，个体i特征降维后的通道数量是$C-C/d$，其中C是特征的总通道数量，d是通道比例因子.$w^{d}_{j\to i}$是个体j的通道降维参数，个体j特征降维后的通道数量是C/d.提高通道比例因子，可以进一步削弱主动行为者的影响，当比例因子提高到无穷大，则不考虑环境中的个体影响，模型退化为无因果关系的图模型.

本文的时间同步融合，可以处理具有多个Granger原因的个体.首先找出个体i的所有Granger原因的个体，其次根据Granger延迟量计算每个Granger因果对之间的时间同步特征，最后采用平均操作处理

多个时间同步特征，有

$$x_{i,k,c}^{syn}=\frac{1}{\sum_j e_{j\to i}^{Granger}}\sum_j x_{i,k,c}^{con,j\to i} \tag{11}$$

其中，$e_{j\to i}^{Granger}$是Granger因果关系.由于对个体i的所有Granger原因，在Granger因果关系中都被设置为1，因此，分母部分对其求和，可以得到影响个体i的Granger原因数量.

3.3　距离和外观约束的因果图模型

上述的时间因果图根据因果强度来进行推理，但是，只考虑因果关系无法细分出交互行为的类型.外观和距离作为细节属性，可以进一步区分交互行为类别.外观可以指出球员是否为同一阵营，外观可以描述姿态，发现姿态之间的相关性.距离可以描述两个个体之间的远近.例如，拦网需要在前排尽可能地靠近扣球，避免扣球方向的变化；而扣球人员需要根据对方站位确定落点，扣球和对方后排的垫球人员的交互关系就比较远.因此，图5在因果图基础上，进一步融合外观图和距离图，以便将上述信息融合在交互特征中.

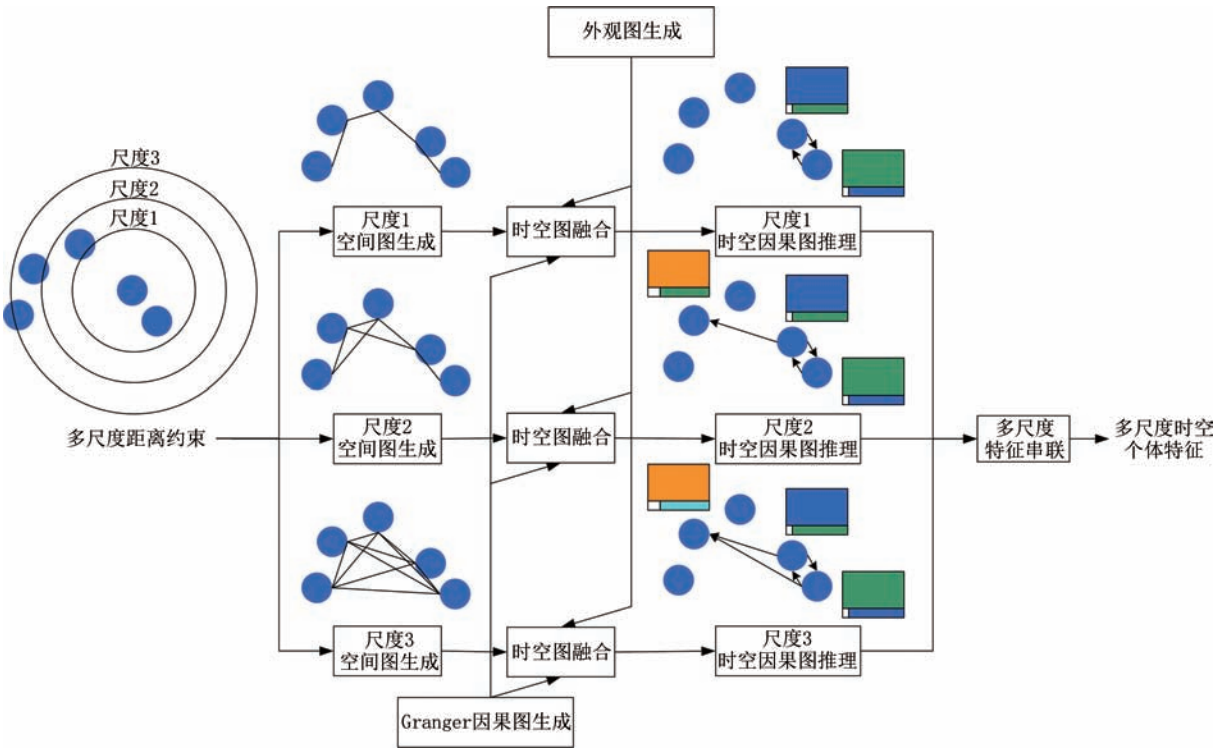


图5　多尺度距离、外观、时间因果关系的图推理

给定个体的特征，使用点积相似度，来描述外观图中两个个体特征之间的相似性.

$$e_{i,j,k}^{app}=\frac{1}{C}\left(w_1^{app}\cdot x_{i,k,1:C}\right)^{Trans}\cdot\left(w_2^{app}\cdot x_{j,k,1:C}\right) \tag{12}$$

其中，C是特征的通道数量，分母用于对相似度进行归一化，$x_{i,k,1:C}$表示个体i时刻k的所有通道的特征向量，Trans表示矩阵的转置操作.w_1^{app}，w_2^{app}是外观相似度模块的参数，用于学习比原始特征更好的嵌入特征空间.采用点积操作，具有容易计算的梯度形式，有利于参数的训练.

给定个体的边界框位置，考虑多个距离超参数，来生成多尺度的距离图：

$$e_{s,i,j,k}^{dist}=\begin{cases}1 & dist_{i,j}\leqslant\lambda_s\cdot width\\ 0 & otherwise\end{cases} \tag{13}$$

其中，s是距离尺度编号，$dist_{i,j}$是个体i和个体j边界框中心点之间的欧式距离，λ_s是尺度s的距离超参数，$width$是图像分辨率的宽度.

多时空约束图模型，综合考虑外观图、距离图、因果图.具体来说，外观图提供的边的权重，而距离图、因果图是两个指示函数，用于修剪群体图交互关系，以有针对性的分析典型的交互行为.多时空约束图的邻接关系为：

$$e_{s,j\to i,k}^{fuse}=\frac{e_{j\to i}^{Granger}\cdot e_{s,i,j,k}^{dist}\cdot e_{i,j,k}^{app}}{\sum_j e_{j\to i}^{Granger}\cdot e_{s,i,j,k}^{dist}\cdot e_{i,j,k}^{app}} \tag{14}$$

其中，分母是归一化项，即考虑在距离尺度s的距离约束下，影响个体i的个体j的数量.

本文的多时空约束图，是对每个时刻分别处理

的，允许每个时刻发生变化.本文进一步使用学习的多时空约束图，进行图卷积操作：

$$X_{s,k}^{graph}=E_{s,k}^{fuse}X_k^{syn}W_k^{graph} \tag{15}$$

其中$X_{s,k}^{graph}=\{x_{s,i,k,c}^{graph}\}$是图卷积之后的特征，$E_{s,k}^{fuse}=\{e_{s,j\to i,k}^{fuse}\}$是多时空约束的邻接关系图，$X_k^{syn}=\{x_{i,k,c}^{syn}\}$是时间同步后的特征.$W_k^{graph}$是图卷积的参数，该参数允许学习群体中通道特征之间的关系.

对于个体行为识别，首先将多尺度的个体特征在维度s上进行串联$concat_s(\cdot)$，使用两层全连接层作为分类器$FCs(\cdot,w^{ind})$，实现对每帧个体特征的分类，其中w^{ind}是个体行为分类器的参数.最后，将视频帧序列的平均预测结果作为个体行为识别的结果.

$$y_i^{ind}=\frac{1}{T}\sum\nolimits_k FCs(concat_s(x_{s,i,k,1:C}^{graph}),w^{ind}) \tag{16}$$

对于群体行为识别，由于多时空约束图已经融合了个体之间的图推理过程，能够有效推荐出关键的个体来描述群体场景中的主要行为.因此，本文先将个体行为按照个体编号进行max pooling找出最大响应的特征作为代表特征，随后使用串联方式来同时考虑多个尺度，使用两层全连接层完成群体行为识别任务$FCs(\cdot,w^{group})$，其中w^{group}是群体行为分类器的参数.

$$\begin{cases}y^{group}=\dfrac{1}{T}\sum\nolimits_k FCs(concat_s(x_{s,k,1:C}^{group}),w^{group})\\ x_{s,k,1:C}^{group}=\max pooling_i(x_{s,i,k,1:C}^{graph})\end{cases} \tag{17}$$

3.4 损失函数

本文提取的多时空约束图特征可以同时用于个体特征表达和群体特征表达.我们使用个体行为标签来鼓励找出被动行为者相关的主动行为者，从而本文的图推理特征能够区分易混淆的被动个体行为.同理，群体行为标签能够鼓励图推理特征区分易混淆的群体行为.本文使用的损失函数同时考虑了个体行为识别损失和群体行为识别损失：

$$L=\frac{1}{N^{train}}\sum\nolimits_u\sum\nolimits_i L_E(y_{i,u}^{ind},y_{i,u,gt}^{ind})+\frac{1}{N^{train}}\sum\nolimits_u L_E(y_u^{group},y_{u,gt}^{group}) \tag{18}$$

其中，u是训练集中视频的编号，N^{train}是训练集视频的数量，i是个体的编号，$L_E(.)$是交叉熵形式的损失函数.第一项是个体行为识别损失，$y_{i,u,gt}^{P}$是个体行为的人工标记，第二项是群体行为识别损失，$y_{u,gt}^{G}$是群体行为人工标记.交叉熵函数形式为：

$$L_E(y_{i,u}^{ind},y_{i,u,gt}^{ind})=y_{i,u,gt}^{ind}\cdot\log(y_{i,u}^{ind})+(1-y_{i,u,gt}^{ind})\cdot\log(1-y_{i,u}^{ind}) \tag{19}$$

4 实 验

4.1 数据集

我们在两个公开的群组行为识别数据集(Volleyball数据集和Collective Activity数据集)上进行了实验，Volleyball数据集由55场排球比赛中收集的4830个视频片段组成，其中有3493个训练片段，1377个测试片段.在每个视频片段中，视频的中间帧标注了个体的边界框，个体行为标签和群体行为标签.其中个体行为标签有9种，分别是：Blocking、Digging、Falling、Jumping、Moving、Setting、Spiking、Standing、Waiting.群体行为标签有8种，分别是：Right set、Right spike、Right pass、Right winpoint、Left set、Left spike、Left pass、Left winpoint.在我们的实验中，我们使用一个长度为T=10的时间窗口，对应于标注帧的前5帧和后4帧.未被标注的个体边界框数据从该数据集提供的轨迹信息数据获取.

Collective Activity数据集由44个视频组成，总共2511个视频片段，实验使用1673个训练片段和838个测试片段.每个视频片段每10帧有一个标注，标注包含个体的边界框、个体行为和群组行为标签.6种个体行为标签，分别为：NA、Crossing、Waiting、Queueing、Walking、Talking.共5个群组活动标签，分别为：Crossing、Waiting、Queueing、Walking、Talking.

4.2 实验细节

本文实验使用ImageNet数据集上预训练的Inception-v3网络提取图像特征，特征维数是1024，即特征通道数是1024.对于输入的个体边界框，以个体中心点为中心，裁剪分辨率为5×5的图像块，选取图像块的2×2的采样点来提取不同局部的特征.由于网络输入需要缩放图像，个体采样点坐标会产生浮点数，而不是整数.ROI Align方法首先根据采样点坐标确定其临近的四个的图像网格坐标，使用双线性插值，分别计算各个采样点坐标的特征；最后将2×2采样点的特征串联成一维向量，获得个体特征.

本文的训练过程包括三个阶段.第一阶段不考虑GCN模块，直接将个体特征进行一层FC降维到256个特征通道，随后进行个体行为识别和群体行为识别.第二阶段，将场景的个体依次配对并构建训练集，使用最小二乘法训练Granger因果检测中的

回归模型参数.在第一阶段和第二阶段预训练参数的基础上，第三阶段进一步添加通道级时序因果图推理，多尺度时空因果图推理，并使用多尺度时空融合后的特征，进行个体行为识别和群体行为识别.第一阶段和第三阶段使用相同的损失函数，第三阶段网络将第一阶段网络的参数固定.

在Volleyball数据集上，batch size为8，分类器层的dropout参数为0.3.实验采用Adam优化器，学习率初始设置为1e-4，网络训练180个周期，每30个周期学习率将为之前的0.5倍，学习率在四次衰减后停止衰减.在距离约束的因果图模型多尺度的距离超参数分别是0.1，0.2，0.3，0.4.

在Collective Activity数据集上，batch size为16，分类器层的dropout参数为0.5.实验采用Adam优化器，初始学习率为1e-3，网络训练80个周期，每10个周期学习率将为之前的0.1倍，学习率在四次衰减后停止衰减.在距离约束的因果图模型多尺度的距离超参数分别是0.1，0.2，0.3，0.4.

实验在64位Ubuntu 16.04上进行，编程环境选择Python 3.8，实验采用Pytorch 1.8深度学习平台，配置Intel Core i9-10900X@3.7 GHz处理器和64 GB内存，配有1块GeForce RTX 3090显卡.

4.3　评价方法

对个体行为和群体行为识别，采用两种评价方法：(1)每类平均正确率(MPCA：Mean Per-Class Accuracy)，先计算每个类别的正确率=该类别正确检测样本数/该类别样本总数，然后求其平均值；(2)多类正确率(MCA：Multi-Class Accuracy)，先求出所有类别的正确样本数，并除以所有类别的样本总数来获得多类正确率.

4.4　消融实验

实验在Volleyball数据集开展消融实验，来分析各个模块参数对因果关系提取，对因果图推理的影响.

4.4.1　Granger因果检测模型对因果关系检测的影响

Granger因果检测模型的3个主要参数是时间窗口尺寸、时间延迟量、因果关系阈值.为了分析视频中多人因果关系对参数的依赖，我们在Right spiking视频中，分析和spiking相关的三种主要交互关系的因果关系检测结果.表1、表2、表3给出spiking-blocking，spiking-moving，spiking-digging的多人关系的统计矩阵，即分析训练视频中发现的相应交互关系的数量.

表1　Right spiking中spiking-blocking的因果关系检测结果

因果关系阈值	0.90	0.90	0.95	0.95	0.98	0.98
时间窗尺寸	3	4	3	4	3	4
时间延迟量						
0	70	127	56	103	42	83
1	40	79	29	58	19	40
2	22	42	12	24	5	12

表2　Right spiking中spiking-moving的因果关系检测结果

因果关系阈值	0.90	0.90	0.95	0.95	0.98	0.98
时间窗尺寸	3	4	3	4	3	4
时间延迟量						
0	63	105	47	93	35	73
1	34	58	25	43	18	32
2	18	37	12	24	8	13

表3　Right spiking中spiking-digging的因果关系检测结果

因果关系阈值	0.90	0.90	0.95	0.95	0.98	0.98
时间窗尺寸	3	4	3	4	3	4
时间延迟量						
0	29	46	27	37	21	26
1	14	33	14	19	12	13
2	7	16	6	14	2	4

我们发现：

(1)时间窗尺寸为4比时间窗尺寸为3发现更多的因果关系，除了直接的因果关系外，也可能发现一些间接的因果关系.例如，两个人都在摆臂，是因为同时被第三个人的行为影响，但是，他们双方的直接原因不是对方.具体来说，随着时间窗口尺寸m值变大，由于F值计算过程中，相关回归的样本是子回归样本的两倍，因此，F值也随之变大.当F值转化为概率p时，需要观察F分布的统计量$F(m, k\cdot c-2m-1)$，其F分布的均值为$E(F)=(k\cdot c-2m+1)/(k\cdot c-2m-1)$，均值依赖于其第二个自由度.该均值随着时间窗$m$值变大而减小，会造成从0到固定点的$p$值变大.最终在F值变大和F分布均值变小的共同作用下，p值进一步变大.当m值较大，所有p值达到接近于1的饱和情况.当m值较小，也会造成p值变小.我们也发现时间窗m值增大，其计算量更大，不利于模型实现.为了平衡因果关系数量和计算量，我们最终选择时间窗尺寸为4.

(2)随着延迟量的增加，因果关系数减少.延迟量为0的时候发现最多的因果关系，说明行为的反应是及时的，运动员需要根据扣球人的行为积极调整位置.延迟量为0的因果关系，发现的spiking和blocking之间的关系是有限的，主要是因为视频帧

中，前半部分扣球人在起跳，blocking人员还在观望并没有明显的行为变化，所以造成F值较小，没有发现他们之间的因果关系.当延迟量增加的时候，能发现上述引起的延迟的弯曲膝盖、起跳行为，主要的延迟量在1和2之间.

(3)因果关系阈值用于描述交互双方行为的回归关系，当阈值下降会发现更多的因果关系.虽然都是数量增加，但是因果关系阈值和时间窗口大小变化，发现的因果关系有本质的不同.因果关系强度下降发现相关性弱的交互，时间窗增加发现较长时间内的因果相关性交互.因此，因果关系阈值取值为0.95，来严格判断因果关系，避免引入噪声关系.

4.4.2 Granger因果检测模型对群体行为识别的影响

在通道级时间同步特征融合中，我们分析三种不同的时间处理方式的影响.(1)特征不根据延迟量平移(无平移)，(2)特征采用固定的延迟量平移(平移1帧)，延迟量固定为1，(3)特征采用自适应延迟量平移(自适应平移)，通过比较多个延迟量的因果关系，选择因果关系最强的延迟量进行平移.由于自适应平移需要考虑多个延迟量组合下的因果关系矩阵，为了避免因果关系矩阵造成的影响，我们分析相同的因果关系矩阵情况下，不同的时间处理方式.

表4给出了上述时间处理方式，在多个延迟量组合情况下的群体行为识别正确率和个体行为识别正确率.其中Granger因果检测模型使用4.4.1节中确定的最佳参数.融合时，通道比例因子为2，使用单尺度图，距离超参数为0.4.

表4 不同时间处理方式在多延迟量组合情况下群体行为识别和个体行为识别结果

多延迟量组合	群体行为MCA			个体行为MCA		
	无平移	平移1帧	自适应平移	无平移	平移1帧	自适应平移
delay=[0]	91.0	89.7	—	81.5	81.2	—
delay=[0,1]	91.8	91.2	92.1	82.1	81.7	82.3
delay=[0,1,2]	91.9	91.4	92.4	82.2	81.8	82.6
delay=[0,1,2,3]	91.9	91.4	92.4	82.2	81.8	82.6

我们发现：(1)群体识别中，在延迟量*delay*=0时，不平移特征比平移特征的结果好，这是因为*delay*=0检测出的关系就是不平移的，此时，特征平移会产生错误的时间对齐.(2)在*delay*=[0,1]时，自适应平移是最好，因为可以自动根据最佳的延迟时间进行对齐.不平移特征比固定平移特征好，这是因为检测出的*delay*=0的因果关系数量，多于*delay*=1的因果关系数量.(3)在*delay*=[0,1,2]时，其自适应特征的正确率趋于饱和，这是因为*delay*=2发现的因果关系数量有限.*delay*=[0,1,2,3]时，其自适应特征不再增加，这是因为*delay*=3发现的因果关系没有明显增加.(4)个体行为低于群体行为识别，是因为个体之间具有遮挡情况，遮挡个体的标签容易预测错误.使用多个个体特征的池化融合，能够避免少数个体特征中存在干扰.

4.4.3 通道比例对群体行为识别的影响

由于被动行为者的行为受到主动行为者影响，我们将因果关系的两个个体的特征进行通道融合，来描述被动行为者的行为.我们设置通道比例，来限制主动行为者特征的参与量.比例因子为d，则主动行为者的特征参与比例为$1/d$，当d越大，则主动行为者参与量越小.当趋向于无穷大时，被动者行为特征为0，此时仍然使用不进行因果融合的特征，在表5中描述为不进行多人融合.

表5 不同通道比例下群体行为识别和个体行为识别结果

通道比例	群体行为MCA	个体行为MCA
d=2	92.4	82.6
d=3	92.6	82.8
d=4	92.9	83.0
d=5	93.1	83.2
d=6	93.3	83.3
d=7	92.0	82.2
d=8	91.8	82.0
不进行多人融合	91.6	81.8

表5给出了不同通道比例因子下的群体行为识别和个体行为识别结果，其中Granger因果检测模型使用4.4.1节中确定的最佳参数，延迟量为[0,1,2]，自适应时间平移，使用单尺度图，距离超参数为0.4.

我们发现：(1)不进行多人融合，此时个体行为只考虑自己的特征，此时个体行为识别和群体行为识别准确率较低.(2)d=2时，主动行为者特征和被动行为者特征的通道数相等.在交互过程中主动行为者的特征不发生因果融合，被动行为者的特征根据因果融合，同时考虑主动和被动个体的特征.因此，d=2时，个体行为识别、群体行为识别准确率比不进行因果融合的模型更高.(3)最好的比例出现在d=6，此时环境中的主动行为者特征占1/6，被动行为者特征占5/6.说明环境中的原因需要参与交互双

方的行为，但是，仍然是因被动者自己的行为为主，被动者在受到外界环境干扰下，仍然有能力决定自己的行为.(4)随着d的增加，主动行为者特征减少，行为被动者特征增加.行为被动者特征在只描述结果的情况下，正确率略有下降.

4.4.4　多尺度距离约束对群体行为识别的影响

上述消融实验，使用的图卷积条件都是单尺度图距离超参数为0.4.表6分析距离超参数的目的是用距离来选择出不同的Granger因果关系，以便使用不同的图模型来学习其中的特征.我们固定最佳的Granger因果检测超参数、通道融合超参数，表6中的最后一行使用4个尺度交互关系并联的图卷积网络.

表6　不同尺度和距离情况下的群体行为识别和个体行为识别结果

距离超参数	群体行为MCA	个体行为MCA
0.1	92.1	82.4
0.2	92.7	82.9
0.3	93.1	83.2
0.4	93.3	83.3
[0.1,0.2,0.3,0.4]	94.3	84.2

实验表明：(1)不同距离可以筛选出不同数量的因果关系.单尺度情况下，0.4尺度空间比0.1尺度空间包含更多的因果关系，因此，0.4尺度空间中个体行为和群体行为识别准确率都比0.1尺度空间的模型要高.(2)在群体行为中，交互行为多数发生在较近的社交距离内，因此，尺度为0.1能发现较多交互关系.随着尺度增加，新发现的交互关系数量下降.0.4尺度比0.3尺度情况下，新增的因果关系有限，因此，正确率提高也有限.(3)多尺度模型中包含了所有尺度的因果关系，而且使用不同图卷积参数，来学习不同距离的交互行为，因此，多尺度模型比0.4尺度模型有明显的提高.

4.4.5　因果关系对群体行为识别的影响

在现有的图卷积群体行为识别中，图中的关系使用外观和距离线索，而本文模型在上述线索之外，使用个体的多帧运动特征来提取因果关系，用于对图模型的修正.表7进一步讨论因果关系对图模型的影响.其中因果关系参数为时间窗为4，延迟量为[0,1,2]，因果关系阈值为0.95，多尺度的距离超参数为[0.1,0.2,0.3,0.4].

表7　因果关系和无因果关系下的群体行为识别和个体行为识别

Method	♯Param(M)	FLOPs(G)	Time/ms	群体行为MCA	个体行为MCA
Backbone-Inception	0.261	0.634	59.2	89.8	80.9
+CR	1.311	0.886	64.0	91.2	81.6
+MSGCN [27]	25.191	6.042	86.8	92.1	82.2
+CR+MSGCN	26.241	6.294	95.9	94.3	84.2

(1)Backbone-Inception模型，包括FC进行特征变换，个体行为分类器和群体行为分类器.(2)Backbone+CR (Causality Relation)，提取了额外的时间延迟量和因果关系，进行了时间平移多人特征融合，更好地描述了被动个体的时间特征，此时的个体行为识别结果优于Backbone-Inception的结果.(3)Backbone+MSGCN (Multiple Scale Graph Convolutional Network)在Backbone-Inception模型基础上，额外添加外观和距离线索构建多尺度图模型.根据ARG模型的设置，采用16个并行的图模型学习复杂的个体之间的关系.时间因果关系对外观和距离不敏感，在发现有效关系时，也可能引入噪声关系.因此，该模型使用外观和距离，对时间因果图进行噪声关系删除，从而其结果优于时间因果图.(4)Backbone+CR+MSGCN在因果图基础上，考虑时间平移多人特征融合，提取个体的时空上下文特征，同时构建同时满足外观、距离、因果的个体关系图进行图推理，进一步学习个体的空间上下文特征，从而，Backbone+CR+MSGCN的识别准确率最高.相对于Backbone+MSGCN模型、Backbone+CR+MSGCN的计算量增加4%(0.252G/6.042G)，每个视频的预测时间为95.9ms，仍然具有较好的推理速度.

4.5　现状对比实验

表8提供了Volleyball数据集的群体行为识别和个体行为识别结果.对比实验中考虑了不同的主干网络(Backbone)提供的行为特征，考虑了是否(Y/N)使用光流特征.

(1)空间关系池化模型认为交互双方是同等重要的关系，不使用参数来学习交互关系.层次深度时间模型(HDTM)[25]对个体的动态特征利用池化进行交互关系建模.社会场景理解网络(SSU)[26]对个体的动态匹配特征，利用池化进行交互关系建模.本文方法通过时空特征的回归建模，学习个体之间特

表8 本文方法和现有方法在Volleyball数据集上的对比结果

	Backbone	光流特征	FLOPs(G)	群体行为 MCA	群体行为 MPCA	个体行为 MCA
空间关系池化模型						
Hierarchical Deep Temporal Model (HDTM) 2016[25]	AlexNet	N	8.6	81.9	—	—
Social Scene Understanding (SSU) 2017[26]	Inception-v3	N	—	90.6	—	81.8
空间关系图模型						
Hierarchical Relational Network (HRN)2018[5]	VGG19-code	N	0.9	89.5	—	—
Graph Attention Interaction Model (GAIM) 2020[3]	Inception-v3	N	—	91.9	—	—
Position Distribution and Appearance Relation (PDAR) 2021[7]	Inception-v3	N	—	92.2	—	—
Actor Relation Graph (ARG) 2019[27]	Inception-v3	N	6.0	92.5	—	82.8
Multi-Level Interaction Relation (MLIR) 2021[4]	Inception-v3	N	—	92.6	92.8	82.8
Actor-Transformers (AT) 2020[6]	I3D	Y	—	93.0	—	83.7
Convolutional Relational Machine (CRM) 2019[28]	I3D	Y	—	93.4	—	—
时空关系图模型						
Confidence-Energy Recurrent Network (CERN) 2017[31]	VGG16	N	—	83.3	83.6	—
Spatial-Temporal Attentive Graph Network (stagNet) 2020[8]	VGG16	N	—	89.3	—	82.3
Coherence Constrained Graph LSTM (CCGLSTM) 2022[33]	AlexNet	N	—	89.3	—	—
Partial context embedding (PCE) 2022[32]	Transformer	N	—	90.5	—	—
Progressive Relation Learning (PRL)2020[9]	VGG16	N	—	91.4	91.8	—
Graph LSTM-in-LSTM (GLIL) 2021[10]	Inception-v3	N	—	—	93.0	—
Visual Context (VC) 2021[2]	Inception-v3	N	—	93.3	93.4	
GroupFormer (GF) 2021[30]	Inception-v3	N	408.5	94.1	—	83.7
本文方法	Inception-v3	N	6.3	94.3	94.4	84.2

征的时间依赖关系，用于描述个体的主动/被动交互关系.因此，本文模型的群体行为识别的多类正确率(MCA)(94.3%)，显著超过HDTM[25](81.9%)和SSU[26](90.6%)方法.

(2)在空间关系图模型中，位置与外观关系网络(PDAR 2021)[7]认为交互关系依赖于外观线索和位置线索，并构建基于双线索的双分支图模型来学习个体交互关系.本文方法使用个体的时空特征来学习Granger因果关系，并将因果关系、外观关系、位置关系融合在因果图模型中，从多种线索来描述个体之间存在的交互关系，从而实现群体中交互关系的有效提取.多层交互关系(MLIR 2021)[4]认为交互关系存在于主要的个体之间，通过设计节点池化方法来找出图结构中的关键节点，形成多个层次的关系图，用多层次的图串联来学习不同层次的交互关系.本文方法使用因果关系来限制个体之间交互，因果关系可以找出主动行为者和被动行为者，从而在群体多人关系中发现主要的交互关系.在Volleyball数据集上，本文方法的群体行为识别的多类正确率(MCA)(94.1%)胜出位置与外观关系网络(PDAR 2021)[7](92.2%)和多层交互关系(MLIR 2021)[4](92.6%).

(3)在时空关系图模型中，图LSTM嵌套LSTM网络(GLIL 2021)[10]设计个体LSTM来学习个体之间的交互关系.本文方法通过对时空特征进行时间回归建模，认为当两个个体之间(个体i和个体j)的相关回归模型误差小于(个体i)自回归模型误差时，两个个体之间存在明显的因果关系，从而学习个体的主动(个体j)和被动关系(个体i).主动/被动关系的学习，更容易解释群体中的交互关系，产生合理的上下文特征.Visual Context(VC)[2]认为交互关系依赖于场景图像的全局特征，将场景图像特征嵌入到图模型来学习交互关系.一致性图LSTM(CCGLSTM)[33]认为交互需要时空上下文特征和场景图像全局特征，使用图LSTM来学习时空交互关系，并添加场景图像的全局特征进行群体行为识别.本文方法关注于场景中关键个体之间的主动/被动关系.实验说明主动/被动关系比场景全局特征，能够更有效描述场景中的关键交互.GroupFormer(GF)[30]认为交互关系中的时间特征和空间特征存在交叉查询过程，设计了时空Transformer模块来考虑个体特征的交叉查询过程.本文方法使用时间回归模型来模型个体特征之间的因果关系，实验说明时间回归因果相关性，比查询匹配的相关性，更有效地描述了个体之间的交互关系，从而在Volleyball数据集

上，本文方法胜出GroupFormer(GF)[30](94.3% vs 94.1%).表8给出了现有开源方法HDTM、HRN、ARG、GF方法的FLOPs.本文方法能够兼顾时间复杂度和模型准确性.

表9提供了Collective Activity数据集的群体行为识别和个体行为识别结果.本文方法使用时间回归模型来估计两个个体时空特征的依赖关系，明显胜出无参数的交互关系估计的空间关系池化模型HDTM[25].本文方法提取的时间回归个体依赖关系，描述了个体的主动和被动关系.该交互关系考虑多个时刻，并考虑多种时间窗口，多种时间延迟量，来处理群体中可能存在复杂时间变化下的交互关系.因此，本文方法使用复杂时间交互关系，从而胜出现有的空间关系图模型(GAIM[3]、PDAR[7]、ARG[27]、AT[6]、CRM[28]).本文方法的复杂时间交互关系，也胜出了现有的基于LSTM的时间交互关系(CERN[31]、GLIL[10])，以及基于Transformer的时间交互关系(VC[2]、GF[30]).

表9　本文方法和现有方法在Collective Activity数据集上的对比结果

	Backbone	光流特征	群体行为MCA	群体行为MPCA
空间关系池化模型				
Hierarchical Deep Temporal Model (HDTM) 2016[25]	AlexNet	N	81.5	—
空间关系图模型				
Convolutional Relational Machine (CRM) 2019[28]	I3D	Y	85.8	—
Multi-Level Interaction Relation (MLIR) 2021[4]	Inception-v3	N	90.2	90.0
Position Distribution and Appearance Relation (PDAR) 2021[7]	Inception-v3	N	90.3	—
Graph Attention Interaction Model (GAIM) 2020[3]	Inception-v3	N	90.6	—
Actor Relation Graph (ARG) 2019[27]	Inception-v3	N	91.0	—
Actor-Transformers (AT) 2020[6]	I3D	Y	92.8	—
时空关系图模型				
Confidence-Energy Recurrent Network (CERN) 2017[31]	VGG16	N	87.2	88.3
Spatial-Temporal Attentive Graph Network (stagNet) 2020[8]	VGG16	N	89.1	—
Progressive Relation Learning (PRL) 2020[9]	VGG16	N	—	93.8
Graph LSTM-in-LSTM (GLIL) 2021[10]	Inception-v3	N	—	94.9
Coherence Constrained Graph LSTM (CCGLSTM) 2022[33]	AlexNet	N	93.0	
GroupFormer (GF) 2021[30]	Inception-v3	N	93.6	—
Visual Context (VC) 2021[2]	Inception-v3	N	95.1	—
本文方法	Inception-v3	N	95.5	96.3

4.6　可视化

4.6.1　因果边关系可视化

图6中给出了Volleyball数据集中Right spike、Left pass群体行为类别的个体关系图，包括时间因果关系图、位置+外观约束的关系图、位置+外观+时间因果约束下的关系图.所有关系图使用单尺度，其距离超参数为0.4，因果关系中时间窗口为4，延迟量为[0,1,2]，因果阈值为0.95.位置关系只有0和1两种取值，多线索融合后的关系在0到1区间内，我们可视化阈值为0.2.

(1)在Right spike中，时间因果图使用时序特征发现右侧扣球人员(8号)和左侧前排拦网(4号、5号)、左侧后排垫球人员(3号)的时间因果关系，说明找到了和扣球相关的人员.但是，在外观关系图中，这些个体的交互关系不明显，最后没有保留在位置+外观+时间因果图中.外观和时间因果共同认为有关系的交互行为是右侧中排的人员(9号、10号).他们根据对方拦网方式，前向迈步找落球点的位置，其迈步行为具有同步性.时间因果图可以检测出时间不同步的因果关系.右侧的扣球人员(8号)和二传手(7号)处于起跳状态，其行为特征和外观特征都具有相关性在.二传手和扣球人员之间，二传手在传球后身体下落，随后扣球人员起跳，具有时间延迟关系，是单向的因果关系.时间因果图可以检测出两个被动个体之间的配合因果关系，例如，左侧前排拦网人员(4号、5号)受到扣球人员行为影响，是被动个体，两者之间具有相关的跳起行为，相互配合实现拦网.图6(a)中，时间因果关系图也发现了8号和3号人员的关系.此时，3号人员的接球行为目的，仍然是一个较大的群体行为识别干扰.当使用外观-时间因果关系图时，外

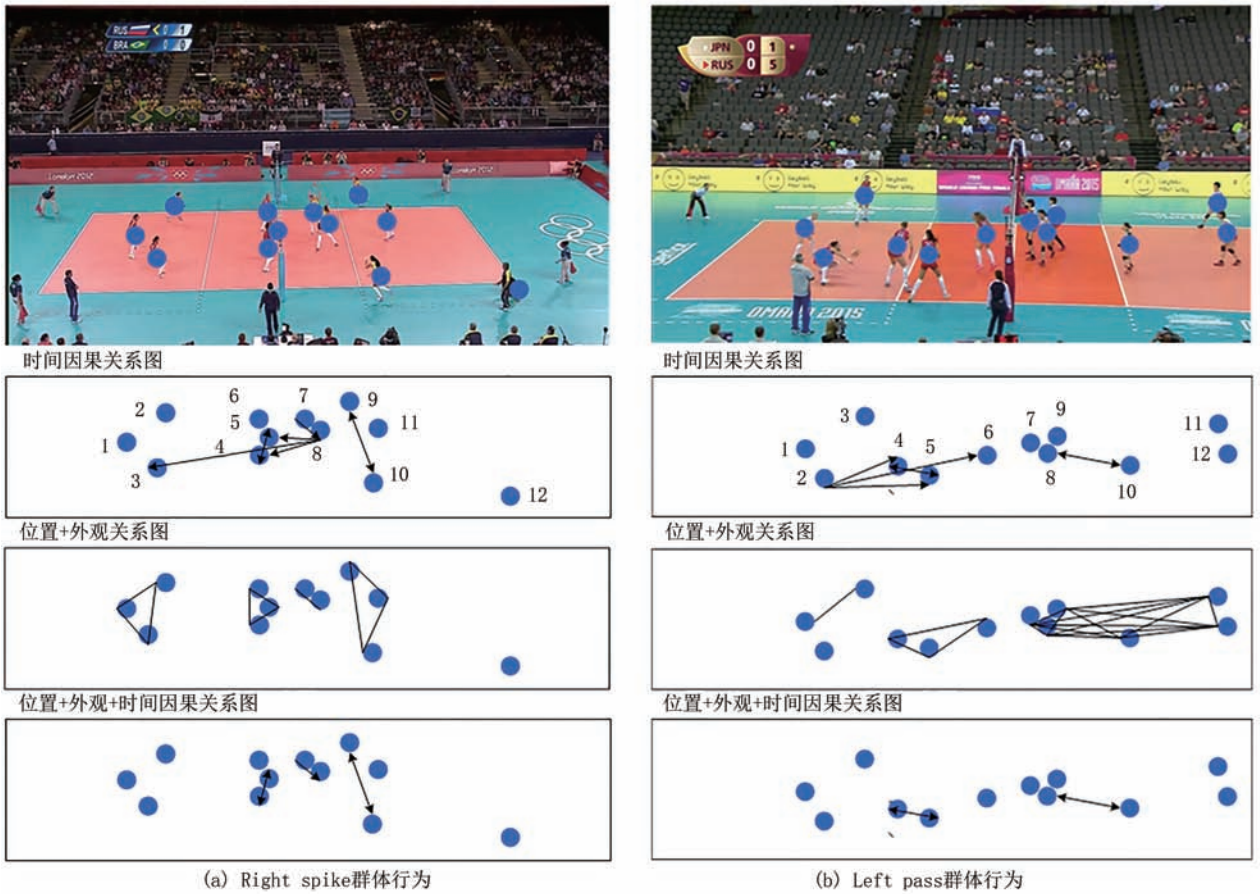


图6 外观关系图和外观—时间因果关系图

观发现主要的活动在右侧,增强了右侧群体行为的描述.同时,外观关系描述了左方后排人员之间的典型关系,另外,两个人员还没有开始准备接球行为.因此,弱化了3号人员对群体行为的干扰.因此,外观-时间因果关系图有效实现了图6(a)的群体行为识别.

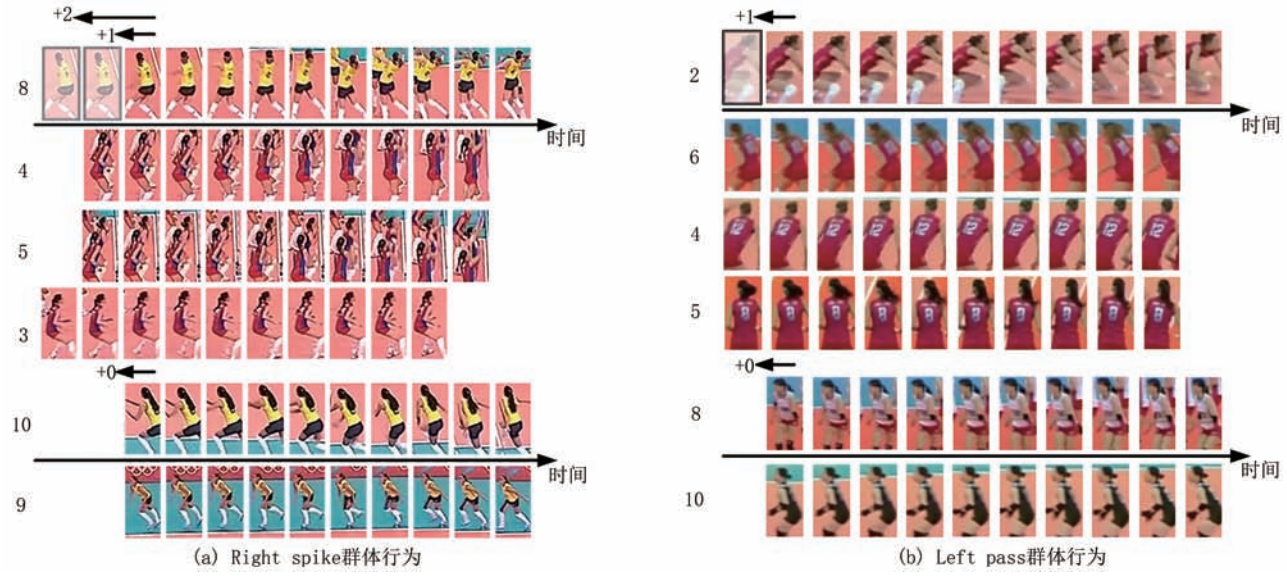
(2)在Left pass中,外观-时间因果关系图认可的关系是左侧两个扣球人员(4号、5号).他们都在同步向下移动,给二传留出位置.此外,右侧防守方发现两个人员(8号、10号)在同步向下走动,他们在判断扣球的路线,以便进行拦网和垫球.在图6(b)的Left pass群体行为中,外观图模型有效描述了主要活动者是左侧2号人员的垫球状态.仅使用时间因果关系时,2号和6号之间有较强的因果关系,6号作为被动个体被2号特征的关系增强,此时,6号产生的群体预测 Left set 被一定程度的增强.此时,如何结合外观图和时间因果图来共同描述2号人员和6号人员,兼顾描述6号个体行为,和2个个体行为的主要作用,从而实现了可靠的图6(b)群体行为识别.

4.6.2 因果延迟关系可视化

图7可视化个体之间的时间因果关系和时间延迟量.时间因果关系能发现行为依赖的多人关系,尤其是具有时间延迟量的多人行为关系.

(1)在Right spike中,8号的扣球行为为球场上的主要行为,被8号影响的人员包括,左侧前排拦网人员4号、5号.根据8号在第3帧起跳,拦网人员4号和5号在第4帧都开始起跳行为,其时间延迟量为1.其中,4号行为特征明显,而5号身体由于部分遮挡,特征被一定程度干扰,会造成5号的时间因果关系,略低于4号的时间因果关系.左侧后排垫球人员3号,也被8号影响.垫球人员3号通过判断落点,在第5帧向下侧面移动,被检测出时间延迟量为2.在Right spike中,也发现了无时间延迟量的因果关系,即进攻方9号和10号.他们在判断对方拦网后的落点,都在向网前迈步.

(2)在Left pass中,左侧垫球人2号是球场上的主要行为,与2号具有明显时间因果关系的是二传手6号,其主要行为是转身来观察垫球方式,以判断排球的落点,当垫球人员在第6帧将球反弹后,二传手通过判断落球点,在第7帧准备开始接球,具有时间延迟量为1.同时,2号也影响了主攻人员4号和副



(a) Right spike群体行为　　(b) Left pass群体行为

图7　时间因果关系和时间延迟量

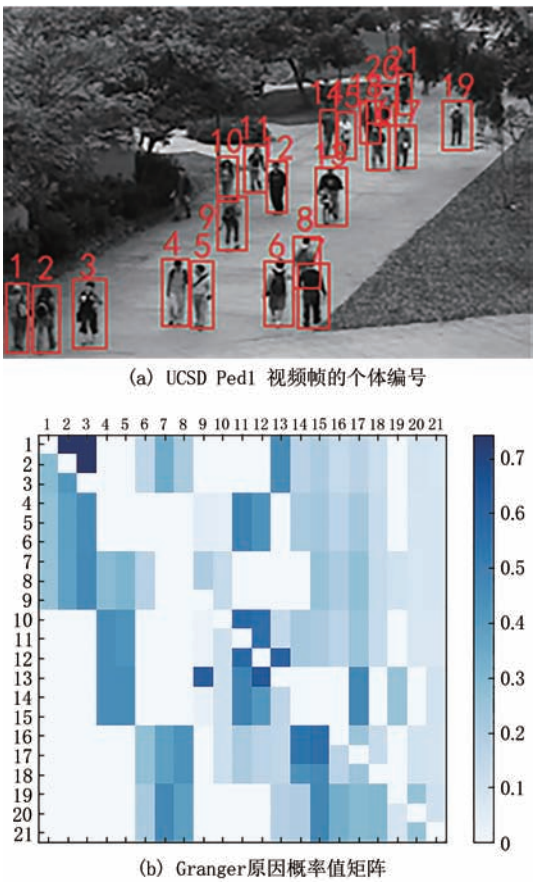
攻人员5号的行为，两人转身观察并向下迈步以给二传手接球的空间位置.在Left pass中，也发现了8号和10号的无时间延迟量的因果关系，他们在判断落点后，向下移动达到预期位置.

(3)在不同的Volleyball群体行为中，人员的关系随着行为的执行而发生变化.例如，扣球早期关注于扣球人员和二传手之间的关系，随后，转换为扣球人员和拦网人员，扣球人员和垫球人员的关系.在群体行为识别中，因果关系检测仍然是群体行为识别的主要挑战.

4.6.3　多人场景因果关系可视化

为了进一步验证多人场景中的因果关系检测，图8给出了UCSD Ped1的本文Granger模型的个体关系检测结果.我们额外标记个体边界框，用于个体之间关系的分析.由于分辨率较低，造成了较低的Granger原因概率值.因此，我们选取阈值0.5来分析个体的因果关系.

图8中，Granger发现了由于避免冲撞而产生的因果关系.(1)本文方法检测出个体3向下行走，需要关注于向上行走的个体1和个体2.同时，由于个体1的步幅比个体2明显，个体1是个体2的行为原因.(2)个体13是向下的骑车人员，会影响到接近的个体9和个体12的行为.个体13在骑行过程中，避免冲撞个体12，也需要考虑个体12的行为.(3)个体10、11、12都是向下行走，其中个体10在前方而且运动幅度较明显.本文方法检测出个体10是个体11和个体12的原因，也检测出个体11和个体12之间也相互避免冲撞.(4)个体14、15朝向个体16、17的方向行走.本文方法可以检测出个体16、17是个体14的原因，也是个体15的原因.



(a) UCSD Ped1 视频帧的个体编号

(b) Granger原因概率值矩阵

图8　UCSD Ped1视频帧的Granger因果关系检测

5　总　结

本文针对群体行为中难以利用时序特征检测多

人因果关系的问题，提出一种基于Granger因果时空图推理方法实现群体行为识别.本文提出了三个模块，来依次解决多人行为的因果关系检测，因果特征融合，时空因果图推理.

为了实现因果关系检测，本文设计了一种时序数据分析的Granger因果关系模型.该模型在时间序列回归模型基础上，分析个体行为是否被环境所影响.具体包括两个子模型，自回归模型用于分析行为被自己历史状态的约束，相关回归模型用于分析被环境状态的约束.通过对两个个体的多帧时序数据进行回归分析，来判断相关回归误差是否明显小于自回归误差.本文的Granger因果关系模型利用F分布的统计量来计算原因的概率值.本文的Granger因果关系可以有效解释被动行为者的运动受到主动行为者的影响.

为了解决多人行为不同步的问题，我们将Granger因果关系推广到带时间延迟量的时间因果关系.我们设计了通道级的时间因果特征融合，来实现主动行为者和被动行为者的特征融合.该模块将主动行为者的运动特征转化为被动行为者的上下文特征，能够更全面了描述被动行为者的运动.

我们设计多尺度时空因果图推理模块，使用因果关系描述主动和被动关系，使用外观关系描述群体中常见的个体的交互外观模式，使用位置关系描述个体之间的距离约束.该模块构建多尺度外观的因果图，能够提取个体之间同时满足外观模式、位置约束、因果关系的交互关系，并使用图推理提取融合上下文的个体特征和群体特征.本文方法在Volleyball和Collective Activity数据集上优于现有群体行为识别方法.

本文方法在Volleyball数据集上能够处理12人的群体行为识别，在Collective Activity数据集能够处理2-14人的群体行为识别.其中，2人可以形成小规模群体活动，6人可以形成中等规模的群体活动.在人员较多的场景中，个体外观分辨率、遮挡、运动不确定性等因素，会造成个体外观特征较大程度地退化.针对外观特征退化的问题，可以进一步考虑运动特征，轨迹特征，来构建因果关系模型.

参 考 文 献

[1] Li Ding, Zhang Wensheng. Attentive pooling for group activity recognition. SCIENTIA SINICA Informationis, 2021, 51(03): 399-412 (in Chinese)
(李定，张文生. 面向群体行为识别的注意力池化机制. 中国科学：信息科学，2021，51(03)：399-412)

[2] Yuan Hangjie, Ni Dong. Learning visual context for group activity recognition // Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event, 2021: 3261-3269

[3] Lu Lihua, Lu Yao, Yu Ruizhe, Di Huijun, Zhang Lin, Wang Shunzhou. GAIM: graph attention interaction model for collective activity recognition. IEEE Transactions on Multimedia, 2020, 22(2): 524-539

[4] Lu Lihua, Lu Yao, Wang Shunzhou. Learning multi-level interaction relations and feature representations for group activity recognition // Proceedings of the 27th Conference on Multimedia Modeling. Prague, Czech Republic, 2021: 617-628

[5] Ibrahim Mostafa S., Mori Greg. Hierarchical relational networks for group activity recognition and retrieval // Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany, 2018: 742-758

[6] Gavrilyuk Kirill, Sanford Ryan, Javan Mehrsan, Snoek Cees G. M.. Actor-transformers for group activity recognition // Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 836-845

[7] Pei Duoxuan, Li Annan, Wang Yunhong. Group activity recognition by exploiting position distribution and appearance relation // Proceedings of the 27th Conference on Multimedia Modeling. Prague, Czech Republic, 2021: 123-135

[8] Qi Mengshi, Wang Yunhong, Qin Jie, Li Annan, Luo Jiebo, Gool Luc Van. StagNet: an attentive semantic RNN for group activity and individual action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(2): 549-565

[9] Hu Guyue, Cui Bo, He Yuan, Yu Shan. Progressive relation learning for group activity recognition // Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 977-986

[10] Shu Xiangbo, Zhang Liyan, Sun Yunlian, Tang Jinhui. Host-parasite: graph LSTM-in-LSTM for group activity recognition. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(2): 663-674

[11] Zhang Chi, Jia Baoxiong, MarkEdmonds, Zhu Song-Chun, Zhu Yixin. ACRE: abstract causal reasoning beyond covariation // Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 10643-10653

[12] Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 1969, 37(3): 424-438

[13] Siggiridou Elsa, Kugiumtzis Dimitris. Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model. IEEE Transactions Signal Process, 2016, 64(7): 1759-1773

[14] Lan Tian, Wang Yang, Yang Weilong, Robinovitch Stephen N., GregMori. Discriminative latent models for recognizing contextual group activities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(8): 1549-1562

[15] Simonyan Karen, Zisserman Andrew. Two-stream convolutional

networks for action recognition in videos // Proceedings of the Conference on Neural Information Processing Systems 2014. Montreal, Quebec, Canada, 2014: 568-576

[16] Wang Limin, Xiong Yuanjun, Wang Zhe, Qiao Yu, Lin Dahua, Tang Xiaoou, Gool Luc Van. Temporal segment networks for action recognition in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2740-2755

[17] Carreira João, Zisserman Andrew. Quo vadis, action recognition? A new model and the kinetics dataset // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4724-4733

[18] Li Ding, Ma Jing, Yang Menglin, Zhang Wensheng. Nonlocal based deep model for group activity recognition. Journal of Image and Graphics, 2019, 24(10): 1728-1737. (in Chinese)
(李定, 马静, 杨萌林, 张文生. 面向群体行为识别的非局部网络模型. 中国图象图形学报, 2019, 24(10): 1728-1737)

[19] Xie Zhao, Zhou Yi, Wu Ke-Wei, Zhang Shun-Ran. Activity recognition based on spatial-temporal attention LSTM. Chinese Journal of Computers, 2021, 44(02): 261-274 (in Chinese)
(谢昭, 周义, 吴克伟, 张顺然. 基于时空关注度LSTM的行为识别. 计算机学报, 2021, 44(02): 261-274)

[20] Carreira João, Patraucean Viorica, Mazaré Laurent, Zisserman Andrew, Osindero Simon. Massively parallel video networks // Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany, 2018: 680-697

[21] Chang Shuo-Yiin, Li Bo, Simko Gabor, Sainath Tara N., AnshumanTripathi, Oord Aäron van den, OriolVinyals. Temporal modeling using dilated convolution and gating for voice-activity-detection // Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2018. Calgary, Canada, 2018: 5549-5553

[22] Cheng Changmao, Zhang Chi, Wei Yichen, Jiang Yu-Gang. Sparse temporal causal convolution for efficient action modeling // Proceedings of the 27th ACM Multimedia 2019. Nice, France, 2019: 592-600

[23] Dai Zihang, Yang Zhilin, Yang Yiming, Carbonell Jaime G., Le Quoc V., RuslanSalakhutdinov. Transformer-XL: attentive language models beyond a fixed-length context // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 2978-2988

[24] Tirupattur Praveen, Duarte Kevin, Rawat Yogesh Singh, MubarakShah. Modeling multi-label action dependencies for temporal action localization // Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 1460-1470

[25] Ibrahim Mostafa S., Muralidharan Srikanth, Deng Zhiwei, ArashVahdat, GregMori. A hierarchical deep temporal model for group activity recognition // Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1971-1980

[26] Bagautdinov Timur M., Alahi Alexandre, Fleuret François, Fua Pascal, Savarese Silvio. Social scene understanding: end-to-end multi-person action localization and collective activity recognition // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3425-3434

[27] Wu Jianchao, Wang Limin, Wang Li, Guo Jie, Wu Gangshan. Learning actor relation graphs for group activity recognition // Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9964-9974

[28] Azar Sina Mokhtarzadeh, Atigh Mina Ghadimi, Nickabadi Ahmad, Alahi Alexandre. Convolutional relational machine for group activity recognition // Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7892-7901

[29] Wang Chuan-xu, Xue Hao. Group activity recognition based on GFU and hierarchical LSTM. ACTA ELECTRONICA SINICA, 2020, 48(08): 1465-1471 (in Chinese)
(王传旭, 薛豪. 基于GFU和分层LSTM的组群行为识别研究方法. 电子学报, 2020, 48(08): 1465-1471)

[30] Li Shuaicheng, Cao Qianggang, Liu Lingbo, Yang Kunlin, Liu Shinan, Hou Jun, Yi Shuai. GroupFormer: group activity recognition with clustered spatial-temporal transformer // Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 13648-13657

[31] Shu Tianmin, Todorovic Sinisa, Zhu Song-Chun. CERN: confidence-energy recurrent network for group activity recognition // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4255-4263

[32] Kim Dongkeun, Lee Jinsung, Cho Minsu, Kwak Suha. Detector-free weakly supervised group activity recognition. arXiv preprint, 2022, arXiv: 2204.02139

[33] Tang Jinhui, Shu Xiangbo, Yan Rui, Zhang Liyan. Coherence constrained graph LSTM for group activity recognition. IEEE Trans on Pattern Analysis and Machine Intelligence. 2022, 44(2): 636-647

[34] Mottaghi Roozbeh, Rastegari Mohammad, Gupta Abhinav, Farhadi Ali. "What happens if..." learning to predict the effect of forces in images // Proceedings of the 2016 European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 269-285

[35] Fire Amy Sue, Zhu Song-Chun. Learning perceptual causality from video. ACM Transactions on Intelligent Systems and Technology, 2016, 7(2): 23:1-23:22

[36] Edmonds Mark, Ma Xiaojian, Qi Siyuan, Zhu Yixin, Lu Hongjing, Zhu Song-Chun. Theory-based causal transfer: integrating instance-level induction and abstractlevel structure learning // Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York City, USA, 2020: 1283-1291

[37] Winkler Irene, Panknin Danny, Bartz Daniel, Müller Klaus-Robert, Haufe Stefan. Validity of time reversal for testing granger causality. IEEE Transactions Signal Process, 2016, 64(11): 2746-2760

[38] Chopra Ribhu, Murthy Chandra Ramabhadra, Rangarajan Govindan. Statistical tests for detecting granger causality. IEEE Transactions Signal Process, 2018, 66(22): 5803-5816

[39] Bhattacharya Gautam, Ghosh Koushik, Chowdhury Ananda S.. Granger causality driven AHP for feature weighted kNN. Pattern Recognition, 2017, 66: 425-436

XIE Zhao, Ph. D., associated researcher. His research interests include computer vision, and image analysis and understanding.

LI Jun, M. S. candidate, His research interests include computer vision, and image analysis and understanding.

WU Ke-Wei, Ph. D., associated researcher. His research interests include computer vision, and image analysis and understanding.

JIAO Chang, M. S. candidate, His research interests include computer vision, and image analysis and understanding.

Background

Group activity recognition aims at detecting the activity of multiple people in a crowded scene, which is conducive to maintaining the safety of public places. Group activity recognition has many potential applications in Video surveillance, video summarization, video retrieval. Unlike action recognition, group activity recognition needs to extract the interaction relation between actors. The interaction relation provides context information for activity. However, the interaction relation is hidden in the activity and its spatial-temporal feature is complex. Therefore, spatial-temporal feature extraction in a group activity is still an ill-posed problem.

Many methods use a graph model to estimate the interaction relation in a group. They analyze the relation with appearance feature, position feature, and motion feature. They estimate symmetrical interaction relations between two actors and consider the interaction is the same as each other. They cannot describe the positive actor and the passive actor in the interaction and cannot generate a reasonable interaction relation.

In this work, we design a Granger causality spatial-temporal graph model to learn the complex interaction relation. First, we design a Granger causality-base relation detector, which has two regression sub-models. The self-regression sub-model measures the temporal dependency with the actor itself temporal features, and the related-regression sub-model considers the actor as a passive one and measures the temporal dependency of a directed relation with the temporal features of both itself passive actor and its potential active actor. When the regression error of the related-regression sub-model is significantly smaller than the regression error of the self-regression sub-model, this suggests the potential active actor is true and the relation is Granger causality relation. Second, we design a channel-wise temporal causality graph inference with the temporally delayed features. We extend the related-regression sub-model in the relation detector to a temporal delayed related-regression sub-model, which can provide the temporal delay information and generate a temporal delayed causality graph. We use the temporal delay information to obtain a temporal delayed feature of the active actor and fuse it into the feature of the passive actor with channel-wise concatenate operation. To learn a spatial-temporal fusion feature for the passive actor, we embed both the feature of the passive actor and the delayed feature of the active actor with a channel-wise dimension transformation and further discuss the channel ratio between the passive actor and the active actor. Third, we embed our Granger causality relation with multiple distance scale and appearance relation. The enhanced graph divides the relation into multiple distances and uses multiple graph convolutional parameters to analyze these relations separately.

This research was supported by the Anhui Province Key research and development plan (202004d07020004), Anhui Natural Science Foundation (2108085MF203), the Fundamental Research Funds for the Central Universities of China (PA2021GDSK0072,JZ2021HGQA0219).