软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2023,34(5):2350−2364 [doi: 10.13328/j.cnki.jos.006439]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 基于多模态多粒度图卷积网络的老年人日常行为识别*

丁 静, 舒祥波, 黄 捧, 姚亚洲, 宋 砚

(南京理工大学 计算机科学与工程学院, 江苏 南京 210094)

通信作者: 舒祥波, E-mail: shuxb@njust.edu.cn

**摘 要**: 随着人口老龄化问题日益严重, 人们对家庭环境中老年人的安全问题越来越重视. 目前, 国内外一些研究机构正在试图研究通过家用摄像头对老年人的日常行为进行智能化看护, 实现对一些危险行为的预警、报警与报备. 为了助推这些技术的产业化, 主要研究如何自动识别出老年人的日常行为, 如“喝水”“洗手”“读书”“看报”等. 通过对老年人的日常行为视频的调研发现, 老年人的日常行为语义具有非常明显的细粒度特性, 如“喝水”与“吃药”两种行为的语义高度相似, 且只有少量的关键帧能准确体现出其类别语义. 为了有效解决老年人行为识别问题, 提出一种新的多模态多粒度图卷积网络 (multimodal and multi-granularity graph convolutional networks, MM-GCN), 通过利用图卷积网络分别从人体骨骼点 (“点”) 和人体骨架 (“线”)、关键帧 (“面”) 和视频提名段 (“段”) 两种模态对老年人行为进行建模, 捕捉“点-线-面-段”这 4 种颗粒度对象下的语义信息. 最后, 在目前最大规模的老年人日常行为数据集 ETRI-Activity3D (11 万+视频段、50+行为类别) 上进行老年人行为识别性能评测, 相比于当前最好的方法, 提出的 MM-GCN 方法取得了最高的识别性能. 此外, 为了验证 MM-GCN 方法对常规人体行为识别任务的鲁棒性能, 在业界标准的 NTU RGB+D 数据集上进行实验, MM-GCN 方法也表现出了很不错的性能.

**关键词**: 老年人行为识别; 图卷积网络; 多模态; 多粒度

**中图法分类号**: TP183

中文引用格式: 丁静, 舒祥波, 黄捧, 姚亚洲, 宋砚. 基于多模态多粒度图卷积网络的老年人日常行为识别. 软件学报, 2023, 34(5): 2350–2364. http://www.jos.org.cn/1000-9825/6439.htm

英文引用格式: Ding J, Shu XB, Huang P, Yao YZ, Song Y. Multimodal and Multi-granularity Graph Convolutional Networks for Elderly Daily Activity Recognition. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2350–2364 (in Chinese). http://www.jos.org.cn/1000-9825/6439.htm

## Multimodal and Multi-granularity Graph Convolutional Networks for Elderly Daily Activity Recognition

DING Jing, SHU Xiang-Bo, HUANG Peng, YAO Ya-Zhou, SONG Yan

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

**Abstract**: With the problem of the aging population becomes serious, more attention is payed to the safety of the elderly when they are at home alone. In order to provide early warning, alarm, and report of some dangerous behaviors, several domestic and foreign research institutions are focusing on studying the intelligent monitoring of the daily activities of the elderly in robot-view. For promoting the industrialization of these technologies, this work mainly studies how to automatically recognize the daily activities of the elderly, such as “drinking water”, “washing hands”, “reading a book”, “reading a newspaper”. Through the investigation of the daily activity videos of the elderly, it is found that the semantics of the daily activities of the elderly are obviously fine-grained. For example, the semantics of “drinking water” and “taking medicine” are highly similar, and only a small number of video frames can accurately reflect their category semantics. To effectively address such problem of the elderly behavior recognition, this work proposes a new multimodal multi-granularity

* 基金项目: 科技创新 2030“新一代人工智能”重大项目课题 (2018AAA0102001); 国家自然科学基金 (62072245, 61932020, 62102182, 61976116)
收稿时间: 2021-04-02; 修改时间: 2021-06-06, 2021-08-08; 采用时间: 2021-08-29; jos 在线出版时间: 2022-09-30
CNKI 网络首发时间: 2022-11-15

© 中国科学院软件研究所 http://www.jos.org.cn

graph convolutional network (MM-GCN), by applying the graph convolution network on four modalities, i.e., the skeleton (“point”), bone (“line”), frame (“frame”), and proposal (“segment”), to model the activities of the elderly, and capture the semantics under the four granularities of “point-line-frame-proposal”. Finally, the experiments are conducted to validate the activity recognition performance of the proposed method on ETRI-Activity3D (110 000+ videos, 50+ classes), which is the largest daily activities dataset for the elderly. Compared with the state-of-the-art methods, the proposed MM-GCN achieves the highest recognition accuracy. In addition, in order to verify the robustness of MM-GCN for the normal human action recognition tasks, the experiment is also carried out on the benchmark NTU RGB+D, and the results show that MM-GCN is comparable to the SOTA methods.

**Key words**: elderly activity recognition; graph convolutional network (GCN); multimodal; multi-granularity

随着社会的飞速发展, 各个国家都出现了不同程度的人口老龄化问题. 人口老龄化是指一个国家或地区人口中 65 岁以上人口占比超过 7% 的一种社会现象[1]. 随着老龄化程度的加深, 空巢老年人的占比也在不断上升, 已经成为一个严重的社会问题. 由于老年人行动缓慢, 在发生危险时无法及时应变、无法及时向医护人员求救, 这可能会导致严重后果. 而随着空巢家庭数量的增加, 上述情况正在呈现逐年上升的趋势. 目前, 国内外一些研究机构正在试图研究对老年人日常行为进行智能化看护, 使得老年人在发生意外危险前进行预警或者在发生意外危险时发出求救信号.

当前, 受益于人工智能技术和深度学习理论的发展, 解决老年人日常安全看护问题的一个解决方案是利用基于深度学习的识别技术对摄像头下的老年人的日常行为进行监测与识别. 深度学习[2]的出现和发展极大地推动了近十年来机器学习各个领域的进步, 如自然语言处理[3]、计算机视觉[4]等. 而作为计算机视觉领域的研究热点, 基于深度的行为识别算法和各种任务的提出正在不断刷新和完善该领域的理论和技术体系[5]. 老年人日常行为识别是近年来一个新兴的行为识别任务, 通过理解和分析老年人日常行为, 能够为老年人安全看护系统提供关键支持.

由于老年人发生危险是突发情况, 因此需要对老年人的日常生活进行监控. 密切了解和监控老年人在日常生活中的实际行为对于老年人行为识别任务至关重要. 世界范围内, 已经有许多类似的研究工作. 如美国佐治亚理工大学开展了 Aware Home Research Initiative 项目[6], 旨在帮助人们通过摄像头看护老人的生活情况, 协助老年人完成日常活动, 确保老人独自在家中的安全. Intel 公司也开展了 Caregiver’s Assistant 项目的研究, 通过各种微型传感器获取老年人日常活动状态, 判断是否有进食、吃药等行为, 为老年人独自生活提供了巨大帮助. 此外, Jinhyeok 等人[7]还拜访了 50 名老年人的家, 仔细监测并记录了他们从早到晚的日常行为, 建立了 ETRI- Activity3D 数据集. 因此, 深入对老年人日常行为识别的研究, 不仅可以帮助解决独居老人日常看护问题, 还可以降低看护成本、提高生活质量, 具有重要的社会意义和研究价值.

通过对老年人的日常行为视频进行调研发现, 老年人行为的语义具有明显的细粒度性, 即多数老年人行为在大部分的视频时长内具有非常高的重合度, 真正区分类别的语义信息比较微妙, 例如图 1(a) 所示的“看报纸”(左图) 和“看书”(右图) 两个老年人行为类别, 背景环境和行为轨迹都非常相似. 在图 1(b) 中, “吃药”(左图) 和“喝水”(右图) 的行为特征相似度也很高.


图 1 老年人日常行为类别示例

另一方面, 当前主流的针对行为识别的基准方法大致分为 3 类: (1) 基于循环神经网络 (recurrent neural network, RNN)[8]的方法; (2) 基于卷积神经网络 (convolutional neural network, CNN)[4]的方法; (3) 基于图卷积网络 (graph convolutional network, GCN)[9,10]的方法. 其中, 基于 RNN 的方法可以利用时序关系处理序列数据, 但是会产生梯度消失问题; 基于 CNN 的方法可以处理高维数据并自动进行特征提取, 但是无法处理非欧式空间数据. 而基

© 中国科学院软件研究所 http://www.jos.org.cn

于 GCN 的方法可以完整地学习非欧式空间的数据, 聚合空间和时间信息, 相比前两种深度网络模型更有优势.

基于以上分析, 针对老年人的日常行为识别任务, 本文提出了一种多模态多粒度图卷积网络 (multimodal multi-granularity graph convolutional networks, MM-GCN), 通过注意力[11]图卷积网络对 2 种模态、4 种颗粒度的数据进行联合建模来全方位揭示老年人行为的时空演变规律. 其中, 2 种模态指的是: (1) 骨骼序列用于捕捉个体行为的结构信息; (2) RGB 视频用于捕捉个体行为的视觉信息. 4 种颗粒度指的是人体骨骼点 (“点”)、人体骨架 (“线”)、关键帧 (“面”)、提名段 (“段”). 在这 4 种颗粒度的数据上, 针对特定的数据类型设计合适的注意力图卷积网络结构, 从不同模态、多种粒度刻画发生重点区域的关注程度, 从而捕捉细粒度级别下的类判别信息. 最后, 通过在标准数据集上的实验评测, 本文所提出的方法达到了最高的识别性能.

综上所述, 本文提出了一种新的基于 GCN 的行为识别模型, 通过融合不同模态多种粒度的高层语义特征来捕捉细粒度的老年人行为信息, 在大规模老年人行为数据集和行业基准数据集上均取得了优秀结果. 本文的贡献主要体现在以下 3 个方面.

- 提出了一种新的多模态多粒度图卷积网络, 通过注意力图卷积网络对 2 类模态、4 种颗粒度的数据同时建模来捕捉人体行为, 解决实际场景下的老年人行为识别问题.
- 设计了一种“点-线-面-段” 4 种颗粒度的数据表示策略, 利用多粒度数据的信息互补与整合来精细刻画视频中的细粒度人体行为.
- 所提出的方法在业界标准的老年人行为识别数据集 ETRI-Activity3D 上进行性能评测, 取得了最好的性能, 其识别精度领先现有的所有方法.

本文第 1 节主要介绍了相关工作. 第 2 节详细介绍了新模型 MM-GCN 的构建. 第 3 节验证实验及结果分析. 最后. 第 4 节对本文的工作进行了总结与展望.

## 1 相关工作

基于机器视觉的人体行为识别是从一个视频或者图像序列中自动分析其中正在进行的行为[5]. 早期的行为识别方法主要是通过手工设计特征的方式来表征行为, 例如方向梯度直方图 (histogram of oriented gradient, HOG)[12]、尺度不变特征转换 (scale-invariant feature transform, SIFT)[13]等. 但是手工设计特征的方式不仅表征能力有限, 还需要耗费大量的时间与精力. 得益于深度学习的发展与普及, 各种行为识别任务的性能相比于传统的浅层方法, 都得到了巨大的提升. 本节以下内容主要是对图卷积网络、多模态学习和注意力机制的调研与介绍.

### 1.1 图卷积网络

卷积神经网络 (convolutional neural network, CNN)[4]是一种经典的行为识别模型, 能够高效地处理欧式空间的特征数据, 因为欧式空间的数据具有平移不变性, 可以共享全局卷积核. 然而, CNN 并不适合非欧式空间数据的表征学习, 因为传统的离散卷积在非欧式空间的数据上无法保持平移不变性. 对于人体骨骼序列这种非欧式空间数据, 传统的 CNN 方法通常是将骨骼点坐标转换为规则的特征向量, 但是没有考虑人体骨骼的自然连接关系.

图卷积网络 (graph convolutional networks, GCN)[9,14]能够学习数据中带有关联信息的特征, 对结构化或时序化数据具有强大的表征能力. 在此基础上, 作为 GCN 的改进模型, 时空图卷积网络 (spatial temporal GCN, ST-GCN)[15]是第 1 个将 GCN 运用到人体行为识别任务上的工作. 针对骨骼序列, Yan 等人[15]将 GCN 拓展到时空图模型上, 从数据中自动地学习时间特征和空间特征, 从而提出了 ST-GCN 模型. 其中, 时空图从两个角度构造: (1) 空间角度. 在每一帧中, 骨骼点作为空间图的节点, 骨骼点的物理连接 (骨架) 作为空间图的边. (2) 时间角度. 将相邻两帧中相同的空间图节点连接, 构成时序边. 类似地, Li 等人[16]提出了另一种时空图卷积 (spatio-temporal graph convolution, STGC) 方法, 通过构建多尺度局部图卷积滤波器和递归学习对动态图进行编码, 并且该方法还可以推广到其他的动态模型中. SlowFast-GCN[17]框架结合了 ST-GCN 和 SlowFastNet[18]的优势: 利用 ST-GCN 对人体骨骼的时空信息进行建模, 同时引入了 Slow-Fast 双流框架, 其中 Slow 流捕获静态语义, Fast 流捕获细粒度的运动变化. 此外, Gao 等人[19]以 ST-GCN 为主干网络, 将人体骨骼分为 5 个区域来识别单人和双人运动并分析涉及动作

© 中国科学院软件研究所 http://www.jos.org.cn

的人数.

由于骨骼点之间不仅有显式的物理连接关系, 还存在隐式的高阶连通性, 因此, Li 等人[20]引入了动作连接推理模块 (a-link inference module, AIM) 来捕获特定动作中存在潜在依赖关系的动作连接, 并且利用动作连接和结构连接构造图结构, 从而提出了动作-结构图卷积网络 (actional-structural graph convolution network, AS-GCN). Li 等人[21]提出了时空图路由 (spatio-temporal graph routing, STGR) 方案来自适应地学习骨骼点之间的高阶依赖关系. 此外, 双流自适应图卷积网络 (two-stream adaptive graph convolutional networks, 2s-AGCN)[22]针对不同的图卷积层设计了自适应的 GCN 结构, 融合了一阶信息和二阶信息来强化学习能力. Zhang 等人[23]提出了时序推理图 (temporal reasoning graph, TRG), 可以在多个时间尺度上同时捕获视频序列之间的外观特征和时间关系, 利用 GCN 提取特征中的语义信息. Shi 等人[24]利用双流 GCN 分别对坐标特征和方向特征建模, 将双流的结果融合来提升识别性能. Shift-GCN[25]摒弃了普通的图卷积操作, 采用新的移位图卷积, 使得空间图和时间图有更灵活的感受野.

### 1.2 多模态融合

多模态学习 (multimodal machine learning, MML)[26]是一种利用多个模态信息进行联合学习的机器学习机制, 通过挖掘模态间的互补性和一致性来提升模型的泛化能力. 其中, 多模态融合是当前应用最多的方向, 针对不同的下游任务选择不同的融合方式, 以达到最优性能.

用于人体行为识别任务的数据主要有 3 种模态: RGB、深度图和骨骼点. 现有的研究一般选择一种或将多种模态融合. Wang 等人[27]将 RGB 视觉特征和深度特征相结合, 协同训练了一个卷积神经网络. Liu 等人[28]基于 RGB 模态和骨骼模态生成了姿态估计图和热图, 利用两种图的互补性来生成分类标签. Hu 等人[29]提出了异质特征学习模型, 将从 RGB、深度图和骨骼数据中提取出的特征进行融合, 挖掘不同模态间的异质性. 此后, Hu 等人[30]又提出了深层双线性学习框架, 将 RGB 特征、深度特征和骨骼特征组合成一种新的特征, 用于行为识别. 多模态关联表示学习 (multimodal correlative representation learning, MCRL)[31]利用不同模态特征来捕获骨骼点周围的局部动态模式, 从而挖掘出多种模态之间的共享特征. SGM-Net (skeleton-guided multimodal network)[32]提出了一个指导模块, 利用骨骼特征来引导 RGB 特征, 在语义特征级别上实现互补.

### 1.3 多粒度融合

多粒度融合通过融合不同粒度数据的侧重信息进行学习和训练, 能对模型进行更全面的指导. 在基于骨骼的行为识别任务中, ST-GCN[15]首先利用 GCN 对骨骼信息进行建模, 但是仅使用了骨骼点信息和固定的图结构, 缺乏对多级语义信息进行学习的能力. 因此 2s-AGCN 利用双流网络对多粒度的骨骼信息进行融合, 即对骨骼点和骨架信息建立双流的自适应图卷积网络, 这样增加了模型的通用性以适应训练数据. DGNN[33]为了更好地利用骨骼点和骨架数据, 基于自然人体骨骼点和骨架之间的运动学相关性, 将数据作为有向无环图, 并设计了一种双流的定向图神经网络 (directed graph neural networks, DGNN), 用于提取两种粒度数据及其关系的信息. GR-GCN[34]提出了一种基于图回归的图卷积神经网络 (graph regression based GCN, GR-GCN) 来表示底层图的稀疏性, 并且对连续帧上的图结构进行了优化, 此外还提供了对骨架的时空建模, 有效地表示了两种粒度的数据. MS-G3D[35]提出了一种简单的多尺度卷积聚合方法和一个统一时空图卷积算子 G3D, 通过结合两者开发了一个强大的特征提取器 (multi-scale G3D, MS-G3D), 学习了骨骼点和骨架信息的时空信息传播.

### 1.4 注意力机制

注意力机制[36]的灵感来源可以归结到人对环境的生理感知上来. 比方说, 人类的视觉系统更倾向于去挑选影像中的部分信息进行集中分析而忽略掉图像中的无关信息. 目前, 注意力机制已经成为深度神经网络中的一个非常重要的嵌入化模块, 被广泛应用到行为识别任务中. Du 等人[37]提出了一个端到端的循环姿态注意力网络 (recurrent pose-attention network, RPAN), 向包含语义信息的关节点共享注意力. Baradel 等人[38]基于注意力模型使用 Glimpse Clouds 从视频帧中提取局部特征, 指导完成行为识别任务. 最近几年, 注意力与图卷积网络进行结合的工作也有一些. 比较有代表性的方法有: 2s-AGCN[22]引入了一种具有注意力的图适应性模块, 能够灵活地对没有物

© 中国科学院软件研究所 http://www.jos.org.cn

理连通性却有隐式连接的关节进行建模; STGR[21]引入了一种挤压-激发注意力机制 (squeeze-and-excitation attention) 来选择信息量最大的图作为代表; 全局上下文感知注意力长短期记忆网络 (global context-aware attention long short-term memory, GCA-LSTM))[39]引入了一种循环注意力机制 (recurrent attention), 能够在全局上下文信息的帮助下选择性地专注于动作序列中信息丰富的关节; 动态聚合图卷积网络 (dynamic aggregate graph convolution neural network, DAG-GCN)[40]引入了一个双重注意力引导模块来完善高层语义特征并强调不同节点之间的语义相关性; 时空注意力网络 (spatio-temporal attention networks, STAN)[41]引入了一种注意力神经元, 不仅能估计每个空间位置的注意力概率, 而且能估计每个视频片段的注意力概率.

## 2 多模态多粒度图卷积网络 (MM-GCN)

### 2.1 细粒度划分

由于老年人的日常行为在大部分时间内具有高重合度, 因此传统模型难以捕捉真正区分类别的语义信息并学习到准确的表示. 为了精准刻画细粒度的人体行为, 本文引入了多模态学习, 并在此基础上对 RGB 和骨骼序列进行了进一步地细粒度划分.

对于给定的 3D 骨骼序列数据, 如图 2(a) 所示, 本文将其细分为两个细粒度.


图 2 细粒度的定义与划分

(1) “点” (骨骼点). 将骨骼点作为图的节点, 人体中骨骼点的物理连通作为图的边. 如果两个骨骼点在物理上是连通的, 则生成一条边; 否则不生成边.

(2) “线” (骨架). 将骨架 (骨骼点的差分) 作为图的节点, 人体中骨架的物理连通性作为图的边. 如果两根骨骼在物理上是连接的, 则生成一条边, 否则不生成边. 值得注意的是, 对于骨架的设置, 本文将靠近人体重心的骨骼点作为源节点, 将远离中心的骨骼点作为目标节点, 因此可以将每根骨骼看作是由源节点指向目标节点的向量, 该向量不仅包含长度信息, 还包含方向信息. 每根骨骼会被分配给唯一对应的目标节点, 但是中心骨骼点未分配给任何骨骼, 因此中间骨骼点添加了值为 0 的骨骼. 上述操作使得基于骨架的图和网络的设计可以和骨骼点保持同步.

RGB 模态数据内容丰富, 包含了大量的视觉语义信息, 挖掘其内在联系对行为识别任务有重要意义. 如图 2(b) 所示, 本文将 RGB 细分为两种粒度.

(1) “面” (关键帧). 对于一个包含完整动作的视频, 从中抽取能体现该动作类别的数帧关键帧. 使得模型在提升判断能力的同时, 降低冗余特征带来的无关干扰和计算消耗.

(2) “段” (提名段, 动作实例). 对于一个包含完整动作的视频, 提名中仅包含发生动作的帧, 即从动作开始到结束的所有帧. 通常可以将同一个动作实例中的提名分为 2 类: 包含相同动作实例的不同阶段的提名和不包含动作示例、只包含背景的提名. 在本文中, 每个动作实例都被提取出多个提名, 并通过建模图结构来显式表示提名间的时序上下文关系.

### 2.2 模型概述

本文提出了一种新的多模态多粒度图卷积神经网络 (MM-GCN), 在多模态融合的双流基础上将进一步细化成为四流图卷积学习框架, 模型的整体结构如图 3 所示.

© 中国科学院软件研究所 http://www.jos.org.cn

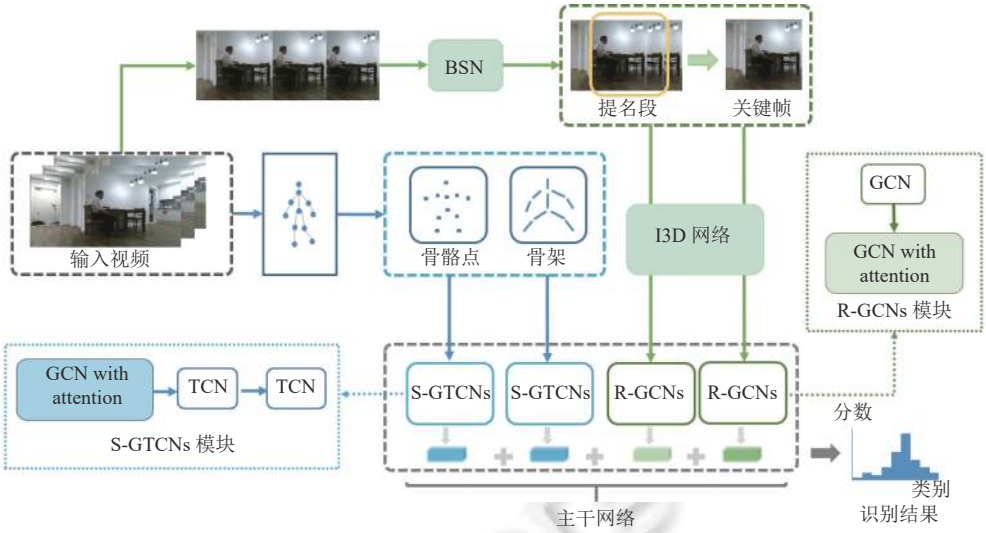


图 3 MM-GCN 的框架示意图

原始输入视频的骨骼模态 (蓝色路径) 和 RGB 模态 (绿色路径) 分别被细化为骨骼点+骨架和提名+关键帧, 并被送入到主干网络中. 具体地, 骨骼点与骨架特征被输入到两个结构相同的 S-GTCNs 模块, 提名和关键帧特征通过 I3D (inflated 3D ConvNet)[42]网络提取底层特征后被输送到 R-GCNs 模块, 最后将四流的输出加权平均得到类别得分. 其中, 模块 R-GCNs 和 S-GTCNs 将在第 2.2.1 节和第 2.2.2 节中进行详细的阐述.

2.2.1 基于骨骼模态的图卷积网络模块 (S-GTCNs)

S-GTCNs 模块通过对大范围的空间和时间信息进行聚合从而学习到基于骨骼数据的行为特征表示.

对于单帧骨骼点数据, 定义图结构 $G_S=(V_S,E_S)$, 其中, $V_S$ 代表拥有 $N$ 个骨骼点的节点集合, $E_S$ 是由邻接矩阵 $A_S\in R^{N\times N}$ 定义的骨骼边集. 如果节点 $v_i$ $(i=1,\ldots,N)$ 和 $v_j$ $(i=1,\ldots,N)$ 存在连接边, 则初始 $A_{i,j}=1$, 否则为 0. 一个完整的动作可以用图序列 $G_S$ 表示, $X_S=\left\{x_{t,n}^S\in R^C\,|t,\ n\in Z,\ 1\leqslant t\leqslant T,\ 1\leqslant n\leqslant N\right\}$ 表示 $G_S$ 的节点特征集合, 其中 $x$ 是在 $t$ 时刻节点 $v_n$ 的特征向量 ($C$ 维), $T$ 表示总帧数, $N$ 表示总节点数.

对于单帧骨架数据, 定义图结构 $G_B=(V_B,\ E_B)$, 其中, $V_B$ 代表拥有 $N$ 根骨架的节点集合, $E_B$ 是由邻接矩阵 $A_B\in R^{N\times N}$ 定义的边集. 如果节点 $v_i(i=1,\ldots,N)$ 和 $v_j(i=1,\ldots,N)$ 存在连接边, 则初始 $A_{i,j}=1$ , 否则为 0. 图序列 $G_B$ 表示一个完整的动作, 节点特征集合为 $X_B=\left\{x_{t,n}^B\in R^C\,|t,\ n\in Z,\ 1\leqslant t\leqslant T,\ 1\leqslant n\leqslant N\right\}$, 其中 $x$ 表示 $t$ 时刻节点 $v_n$ 的特征向量 ($C$ 维), $T$ 表示总帧数, $N$ 表示总节点数.

基于骨骼模态的图卷积网络模型如图 4 所示. 原始输入首先被细分为骨骼点和骨架两种粒度, 输入尺寸为 $(N,M,V,C_1,T)$, 其中 $N$ 表示样本数, $M$ 表示视频中的人数, $C_1$ 表示通道数, $T$ 表示总帧数, $V$ 表示节点数, 经过批量标准化 (batch normalization, BN) 操作归一化后分别被送入两个结构相同、参数独立的网络单元 GTCN 模块中. GTCN 模块由一层图卷积和两层时序卷积[43]组成. 在执行图卷积的过程中, 每个节点都会从其邻域聚合信息, 因此每个节点的特征都会被其他节点增强. 而 TCN 的时序特性能够捕获扩展的上下文信息, 在多种任务上的性能都达到甚至超过了 RNNs 模型. 本文主要采用了 TCN 的膨胀卷积策略, 相比于传统卷积操作, 膨胀卷积允许输入特征被间隔采样, 步长受膨胀率 $d$ 控制. 如图 5(a) 所示, 初始层 $d=1$, 表示输入时每个点都被采样, 图 5(b) 中间层 $d=2$, 则表示输入时进行间隔采样. 一般来说, 越高的网络层使用的 $d$ 越大, 这样网络设置相对较少的卷积层, 就可以获得很大的感受野.

GTCN 单元的输出被送入到 A-GTCN 单元, 与前者不同的是, A-GTCN 单元引入了注意力机制, 其作用是为了使得模型能够专注于关键特征, 从而使得训练过程更加快速、鲁棒. 具体实现为: 通过添加一个可学习、无约束的掩码矩阵来生成注意力, 并且和邻接矩阵 $A$ 相加. 完整的 S-GTCNs 模块由一个 GTCN 单元和两个 A-GTCN 组

© 中国科学院软件研究所 http://www.jos.org.cn

成, 共 9 层, 每层的输出通道数为 96、96、96、96、192、192、192、384 和 384. 网络的最后设置了一层全连接层, 以得到每个动作类别的分数 *score*, 如下所示:

$$\begin{cases} score_S = FC\left(\text{S-GTCNs}\left(\left(\left(x_{t,n}^S\right)_{n=1}^N\right)_{t=1}^T, G_S(V_S, E_S)\right)\right) \\ score_B = FC\left(\text{S-GTCNs}\left(\left(\left(x_{t,n}^B\right)_{n=1}^N\right)_{t=1}^T, G_B(V_B, E_B)\right)\right) \end{cases} \tag{1}$$

其中, *FC* 表示全连接层 (full-connect layer), $score_S$ 和 $score_B$ 分别代表基于骨骼点和基于骨架的输出得分.

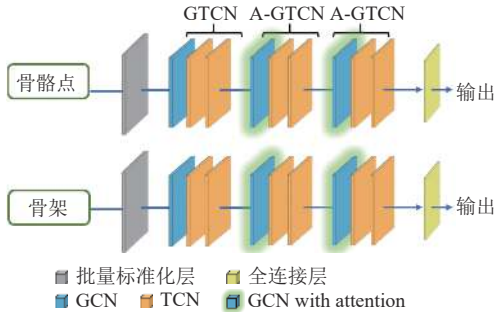


图 4 基于骨骼模态的图卷积网络 (S-GTCNs)

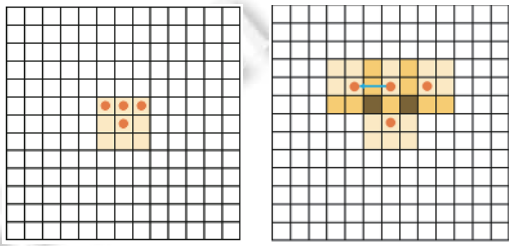


图 5 膨胀卷积对应不同 *d* 值的情况

2.2.2 基于 RGB 模态的图卷积网络模块 (R-GCNs)

类似于 S-GTCNs 模块的处理方式, 本文对 RGB 采用了两种处理策略, 第 1 种是从视频中抽取少量关键帧, 第 2 种则是提取视频的提名 (proposal) 作为输入.

一个动作实例包含了动作的始末时间和动作类别, 在未修剪的视频中常常会有很多动作实例. 一般地, 用提名集合 $P = \left\{p_i \middle| p_i = (x_i, (t_{i,s}, t_{i,e}))\right\}_{i=1}^N$ 来表示视频中的动作实例, 其中 $x_i$ 表示特征, $t_S$ 表示动作开始时间, $t_e$ 表示动作结束时间.

利用边界敏感网络 (boundary sensitive network, BSN)[44]可以提取一个视频的多个提名段, 定义图 $G_P = (V_P, E_P)$. 从整个视频来说, 可以将该视频包含的每个提名作为一个节点 $v_i \in V_P$ $(i = 1, \ldots, N)$. 为了对提名段之间的关联关系进行建模, 本文利用提名段之间的时序重叠程度 $O_{overlap(P_i,P_j)}$ 来构造上下文边. 具体来说, 当某两个提名 $P_i = (t_{i,s}, t_{i,e})$ 和 $P_j = \left(t_{j,s}, t_{j,e}\right)$ 的时序重叠度 $O_{overlap(P_i,P_j)}$ 大于阈值 $\tau$ 时就形成一条上下文有关边:

$$O_{overlap(p_i,p_j)} = \frac{\min(t_{i,e}, t_{j,e}) - \max(t_{i,s}, t_{j,s})}{\min(t_{i,e} - t_{i,s}, t_{j,e} - t_{j,s})} \tag{2}$$

显然, 高度重合的邻域能够提供丰富的上下文信息.

实际上, 虽然提名段中包含了动作开始到结束的所有帧, 但是当时长较长时其信息比较冗余. 为了高效地利用关键信息, 本文从这些提名段采用抽样策略来进一步提取了关键帧. 定义图 $G_F = (V_F, E_F)$ 来表示关键帧构成的图, 其中每个提名中的关键帧作为一个节点 $v_i \in V_F$, 而边集定义和上述提名段的边集定义相同.

$$\begin{cases} score_F = FC\left(\text{R-GCNs}\left(\left(\left(x_{t,n}^F\right)_{n=1}^N\right)_{t=1}^T, G_F(V_F, E_F)\right)\right) \\ score_P = FC\left(\text{R-GCNs}\left(\left(\left(x_{t,n}^P\right)_{n=1}^N\right)_{t=1}^T, G_P(V_P, E_P)\right)\right) \end{cases} \tag{3}$$

本文拟建的基于 RGB 模态的 R-GCNS 模型如图 6 所示. 首先, 利用边界敏感网络从原始视频中提取多个提名, 并从提名中抽取若干关键帧. 其次, 利用 I3D 网络分别提取提名段和关键帧的视觉特征并送入到两个结构相同、参数独立的 GCNs 模块中. GCNs 由两层图卷积组成, 其中第 2 层图卷积层加入了注意力机制. 然而与 S-GTCNs 模块中注意力的添加方式不同, GCNs 模块中是将无约束的掩码矩阵和邻接矩阵相乘, 从而使得重要特征和次要特征的区分度更大, 能够增加网络的鲁棒性. 提名段与关键帧经由 R-GCNs 模块得到各自的输出, 如公式 (3) 所示. 其中, *FC* 表示全连接层, $score_F$ 和 $score_P$ 分别代表基于关键帧和基于提名的输出得分.

© 中国科学院软件研究所 http://www.jos.org.cn

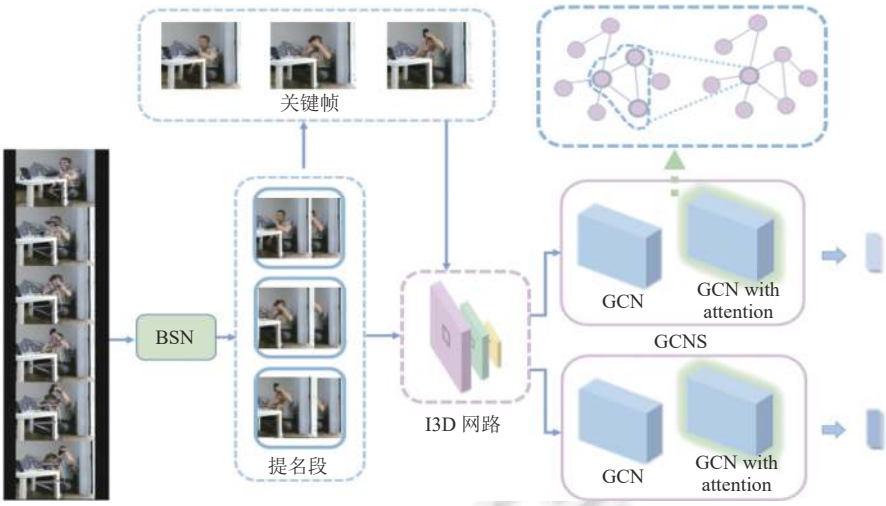


图 6 基于 RGB 模态的图卷积模块 (R-GCNs)

最后, 通过公式 (1) 与公式 (3), 对双流 R-GCNs 与双流 S-GTCNs 进行整合得到最终的分类结果, 如下所示:

$$score=\alpha\cdot score_S+\beta\cdot score_B+\delta\cdot score_F+\sigma\cdot score_P \tag{4}$$

其中, $\alpha$、$\beta$、$\delta$ 和 $\sigma$ 代表各个流输出的权重.

## 3 实 验

为了验证所提出的 MM-GCN 方法的有效性, 本文首先在两个大型基准数据集 ETRI-Activity3D[7]和 NTU RGB+D[45]上分别进行老年人行为识别与正常行为识别实验, 与当前最先进的方法进行了对比分析. 此外, 为了验证 MM-GCN 在老年人行为识别任务中的鲁棒性能, 本文还在 ETRI-Activity3D 数据集上对 MM-GCN 进行了跨年龄的人体行为识别实验与分析.

### 3.1 数据集介绍

ETRI-Activity3D[7]是第 1 个大规模的老年人日常活动视觉数据集, 由 Jang 等人收集并发布. 该数据集由 3 个同步数据模态组成: RGB (如图 7(a) 与图 7(c) 所示)、深度图和骨骼序列 (如图 7(b) 与图 7(d) 所示), 其中 RGB 图像分辨率为 1920×1080, 深度图分辨率为 512×424, 骨骼序列包含被跟踪人体的 25 个身体关节的 3D 位置. 该数据集总共有 112 620 组数据, 采集于 50 位老年人志愿者 (设置 ID 为{1, 2, 3,…, 50}) 和 50 位年轻人志愿者 (设置 ID 为{51, 52, 53, …, 100}), 包括 55 个行为类别, 其中 52 种是从观察老年人的日常活动得出的, 其余 3 种是人机交互特定的行为.

(a) (b) (c) (d)

图 7 ETRI-Activity3D 数据集示例

NTU RGB+D 行为识别数据集由 56 880 个样本组成, 采集于 40 名志愿者, 包含 60 个类别, 每个样本由 RGB、深度图序列、3D 骨架数据和红外视频组成. 其中, RGB 的分辨率为 1920×1080, 深度图和红外视频均为 512×424, 3D 骨架数据包含每帧 25 个主要身体关节的三维位置.

© 中国科学院软件研究所 http://www.jos.org.cn

### 3.2 实验设置

本实验遵循两种实验设置: (1) 交叉主体 (cross subject, CS), 即将不同行为主体混合进行数据集划分; (2) 交叉视角 (cross view, CV), 即训练集和测试集的观察视角不同, 因为动作在不同视角下类间差异不同, 所以 CV 动作识别具有一定挑战性.

在数据集 ETRI-Activity3D 上, 本实验遵循 CS 设置, 将原始数据集划分为训练集和测试集, 其中训练集由 67 位志愿者行为数据组成, 对应的 ID 为{1, 2, 4, 5, 7, 8,…, 97, 98, 100}, 测试集由 33 位志愿者行为数据组成, 对应的 ID 为{3, 6, 9, 12,…, 99}. 在 NTU RGB + D 数据集上, 本文遵循 CS 和 CV 设置. 在 CS 设置下, 划分训练集的 ID 为{1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38}, 其余 ID 为测试集. 同时, 由于该数据集共用 3 个相机捕获动作实例, 因此在 CV 设置下本文将相机 2 和 3 采集的数据作为训练集, 相机 1 采集的数据作为测试集.

对于 RGB 模态数据, 首先利用边界敏感网络为每个动作实例生成 100 个包含 64 帧的提名, 并且同步地在提名中利用间隔采样抽取 20 帧的关键帧, 然后将提名和关键帧分别送入到 I3D 网络提取关键帧和提名的语义特征, 经过一个最大池化层, 输出特征维度为 1 024 维. 将 I3D 网络学习到的特征送入到 R-GCNs 模块中进行训练, 采用交叉熵损失, 阈值$\tau$置为 0.7 (其超参数实验分析见第 3.8 节), 初始学习率 (learning rate) 为 0.001, 学习率调整步长为 50. 对于骨骼模态数据, 初始学习率设置为 0.01, 学习率调整步长为 10.

在最后的决策阶段, 将基于骨骼点、骨架序列、关键帧、动作提名的预测得分融合. 在本实验中, $\alpha$、$\beta$、$\delta$和$\sigma$最佳权重值可以分别设为 0.4、0.3、0.2 和 0.1 (其超参数实验分析见第 3.8 节).

本文实验使用的 GPU 型号是 TITAN RTX, PyTorch 版本是 1.8.0, CUDA 版本是 11.0. 将所提出的 MM-GCN 在 ETRI-Activity3D 数据集全集上进行训练和测试, 在实验过程中将批大小 (batch size) 设置为 32, 模型迭代次数 (epoch) 设置为 120. 基于骨骼模态数据训练, 1 个 epoch 的运行时间开销为 26 min, 测试时间开销为 3 min. 基于 RGB 模态数据训练时, 1 个 epoch 的运行时间开销为 20 min, 测试时间开销为 2 min. 由于 ETRI-Activity3D 是大型数据集, 因此 MM-GCN 的时间开销相对较大, 但仍处于相对可接受的范围内.

### 3.3 实验结果

表 1 展示了在 CS 设置下, MM-GCN 和其他相关先进方法在 ETRI-Activity3D 数据集上的识别结果. 进一步地, 表 2 记录了 MM-GCN 在 ETRI-Activity3D 上 Top-K 的结果.

从表 1 可以看出, 在 CS 设置下, MM-GCN 达到了 94.9% 的识别正确率, 相比于当前性能最好的 FSA-CNN 提高了 1.2%. 相比于单模态方法 Deep Bilinear Learning、Evolution Pose Map 等以及未划分细粒度的 c-ConvNet、HCN 等模型在性能上有很大幅度的提升. 此外, 表 2 中 Top-1 到 Top-5 的结果也验证了 MM-GCN 的有效性, 其 Top-5 的识别精度达到了 99.7%.

为了验证 MM-GCN 的在常规行为识别任务上的鲁棒性能, 本文同时也在 NTU RGB+D 数据集上进行了行为识别实验, 结果如表 3 所示. 在 CS 和 CV 设置下, MM-GCN 分别取得了 90.2% 和 95.5% 的正确率, 其性能处于一个可以接受的水平, 表明其具有良好的鲁棒性能. 但是相比于在 ETRI-Activity3D 数据集上的表现有一定程度的降低, 说明 MM-GCN 对老年人识别任务有一定的针对性. 为了验证此猜想, 本文开展了进一步的对比实验来测试 MM-GCN 在跨年龄段人体行为识别任务上的性能, 其实验结果与分析参见第 3.4 节.

### 3.4 扩展实验: 跨年龄人体行为识别分析

为了进一步验证新模型对于老年人行为识别的针对性与泛化性, 本节探讨了 MM-GCN 对跨年龄段人体行为的识别表现.

基于上述目的, 遵循 Jang 等人[7]的工作, 本文首先将 ETRI-Activity3D 数据集划分为两个子集: 老年人集合 (elderly) 和年轻人集合 (adults). 在此基础上, 将老年人集合划分为训练集 (train set, ID 为{1, 2, 4, 5,…, 49, 50}) 和测试集 (test set, ID 为{3, 6, 9,…, 48}), 将年轻人集合划分为训练集 (ID 为{52, 53, 55, 56,…, 98, 100}) 和测试集 (ID 为{51, 54, 57,…, 99}). 为评估模型跨年龄段的识别性能, 本实验观察 MM-GCN 和当前最先进的方法 FSA-CNN 在 6 种训练测试条件下的表现, 结果如后文表 4 所示, 其中 Mixed 表示训练集同时包含老年人集合和年轻人集合.

© 中国科学院软件研究所 http://www.jos.org.cn

表 1 在 ETRI-Activity3D 数据集上的动作识别结果

| 方法 | 基础模型种类 | CS (%) |
|---|---|---|
| IndRNN[46] | RNN | 73.9 |
| Beyond Joint[47] | | 79.1 |
| MANs[48] | CNN | 82.4 |
| Ensem-NN[49] | | 83.0 |
| SK-CNN[50] | | 83.6 |
| HCN[51] | | 88.0 |
| Deep Bilinear Learning[30] | | 88.4 |
| c-ConvNet[27] | | 91.3 |
| Evolution Pose Map[28] | | 93.6 |
| FSA-CNN[7] | | 93.7 |
| ST-GCN[15] | GCN | 86.8 |
| Motif ST-GCN[52] | | 89.9 |
| **MM-GCN** | | **94.9** |

表 2 在 ETRI-Activity3D 数据集上的 Top-K 识别结果 (%)

| Top-K | Top-1 | Top-2 | Top-3 | Top-4 | Top-5 |
|---|---|---|---|---|---|
| Accuracy | 94.9 | 98.6 | 99.3 | 99.5 | 99.7 |

表 3 在 NTU RGB + D 数据集上的识别结果 (%)

| 方法 | CS | CV |
|---|---|---|
| IndRNN[46] | 81.8 | 88.0 |
| Beyond Joint[47] | 79.5 | 87.6 |
| SK-CNN[50] | 83.2 | 89.3 |
| ST-GCN[15] | 81.5 | 88.3 |
| Motif ST-GCN[52] | 84.2 | 90.2 |
| Ensem-NN[49] | 85.1 | 91.3 |
| MANs[48] | 83.0 | 90.7 |
| HCN[51] | 86.5 | 91.1 |
| Deep Bilinear Learning[30] | 85.4 | — |
| Evolution Pose Map[28] | **91.7** | — |
| c-ConvNet[27] | 82.6 | — |
| FSA-CNN[7] | 91.5 | — |
| STGR-GCN[21] | 86.9 | 92.3 |
| 2s-AGCN[22] | 88.5 | 95.1 |
| DGNN[33] | 89.9 | 96.1 |
| Shift-GCN[25] | 90.7 | **96.5** |
| MS-G3D Net[35] | 91.5 | 96.2 |
| **MM-GCN** | 90.2 | 95.5 |

从表 4 可以看出, 当训练集和测试集属于不同年龄段集合时, 两种方法的识别性能都较低, 说明老年人行为特征和年轻人行为特征有较大的差异; 当训练集和测试集都为老年人集合时, MM-GCN 相对于当前最好的方法 FSA-CNN 提高了 2.9% 的正确率; 当训练集和测试集都为年轻人集合时, MM-GCN 的结果也优于 FSA-CNN; 当训练集为 Mixed 时, 无论测试集是老年人集合还是年轻人集合, MM-GCN 都有较大的优势. 总体来说, MM-GCN 比 FSA-CNN 表现更加优秀. 基于上述分析, MM-GCN 在老年人行为识别任务上表现出了良好的针对性与泛化性.

### 3.5 扩展实验: 消融实验分析

本文提出的 MM-GCN 方法中, S-GTCNs 和 R-GCNs 模块的特点是分别提取骨骼模态数据和 RGB 模态数据的高层语义信息, 并对此进行融合, 利用两种模态信息的互补性完成老年人行为识别任务. 为了分别展示 S-GTCNs、R-GCNs 以及融合后的性能, 本节将通过消融实验展示两个模块的实验结果来验证 MM-GCN 模型的合理性. 如第 2 节所述, MM-GCN 将两个模块应用于骨骼模态数据和 RGB 模态数据, 同时骨骼模态数据被细分为骨骼点和骨架两种颗粒度, RGB 模态被细分为关键帧和提名段两种颗粒度. 在 ETRI-Activity3D 数据集上每个模块的行为识别结果如表 5 所示, 显然, 两种模态融合的结果优于单模态的识别结果.

表 4 在 3 种训练集上的识别结果

| Train set | Test set | FSA-CNN (acc%) | MM-GCN (acc%) |
|---|---|---|---|
| Elderly | Elderly | 87.7 | **90.6** |
| Elderly | Adults | **69.0** | 63.5 |
| Adults | Elderly | **74.9** | 70.0 |
| Adults | Adults | 85.0 | **90.4** |
| Mixed | Elderly | 84.8 | **94.6** |
| Mixed | Adults | 82.1 | **94.1** |

表 5 每个模块的识别结果

| S-GTCNs | R-GCNs | CS (%) |
|---|---|---|
| √ | × | 93.9 |
| × | √ | 76.6 |
| √ | √ | 94.9 |

© 中国科学院软件研究所 http://www.jos.org.cn

从表 5 中可以看出, S-GTCNs 模块的识别正确率为 93.9%, R-GCNs 模块的识别正确率为 76.6%, 两个模块获得的动作识别性能相差较大. 在 ETRI-Activity3D 数据集中, RGB 视频包含杂乱的背景信息 (受拍摄与光照环境影响)、以及人的衣着纹理信息, 这些都与人体的动作行为语义无关. 另一方面, 需要识别的许多行为都属于细粒度行为, 其行为差异往往只体现在人体的局部区域, 而骨骼点更能精确地反映出局部行为的语义信息. 因此, 基于骨骼模态数据的 S-GTCNs 比基于视频模态数据的 R-GCNs 在 ETRI-Activity3D 数据集上的动作识别性能要表现得好很多. 同时, 表 5 的结果也显示两个模块结合后获得了更好的性能. 这是由于骨骼模态数据专注于动作本身, 无法对动作中涉及的交互物体进行准确表示, 而 RGB 视频模态数据中包含物体信息, 这使得在涉及人与物体的交互行为时 RGB 模态数据显得十分重要, 因此将两种模块结合后得到的识别性能比 S-GTCNs 的识别性能更高.

### 3.6 扩展实验: 融合实验分析

为了验证所提出的 MM-GCN 对多种特征融合策略的泛化性能, 本文选取特征平均融合、特征加权融合、得分平均融合和得分加权融合 4 种融合策略进行对比实验. 实验结果如后文表 6 所示, 在 4 种不同的融合策略下, 所提出的 MM-GCN 均取得了不错的性能, 这其中, 在得分加权融合下取得的性能最佳.

### 3.7 扩展实验: 与基准方法的对比实验

为了验证 MM-GCN 方法中 S-GTCNs 模块处理骨骼模态数据和 R-GCNs 模块处理 RGB 模态数据的有效性, 本文也采用代表性的 2s-AGCN 来替换 S-GTCNs, 以及 ResNet18 替换 R-GCNs 作为基准方法来进行对比实验分析, 如表 7 所示. 实验结果显示, 基准方法会降低识别的性能, 从而验证了本文所提出的 MM-GCN 中 S-GTCNs 与 R-GCNs 模块在性能表现方面具有一定的优势.

表 6 4 种融合方法的识别结果

| 方法 | Accuracy |
|---|---|
| 特征平均融合 | 92.3 |
| 特征加权融合 | 92.4 |
| 得分平均融合 | 94.2 |
| 得分加权融合 | 94.9 |

表 7 MM-GCN 与 2s-AGCN+ResNet18 的结果对比

| 方法 | Skeleton (acc%) | Bone (acc%) | Frame (acc%) | Proposal (acc%) | CS (%) |
|---|---|---|---|---|---|
| 2s-AGCN+ ResNet18 | 92.5 | 91.5 | 70.2 | 70.6 | 93.7 |
| MM-GCN | 93.6 | 92.7 | 75.8 | 76.3 | 94.9 |

### 3.8 扩展实验: 超参数实验分析

为了探索形成上下文有关边的阈值$\tau$对识别性能的影响, 本文将$\tau$值分别设置为 0.1、0.3、0.5、0.7 和 0.9, 并进行实验验证分析. 实验结果如表 8 所示, 从该表中可以看出, 当阈值$\tau$=0.7 时, 所提出的方法取得了最好的识别结果.

此外, 在最后的决策阶段, 本文将基于骨骼点、骨架序列、关键帧、动作提名的预测得分融合, 如公式 (4) 所示. 这里, 融合比例分别表示为$\alpha$、$\beta$、$\delta$和$\sigma$ ($\alpha+\beta+\delta+\sigma=1$). 由于骨骼模态数据能够准确地表示动作内容, 而 RGB 视频在部分涉及人-物交互的动作上有重要作用, 因此骨骼模态数据的权重应较大, 而 RGB 模态的权重应较小, 即本文将$\alpha$和$\beta$的值设定为大于$\delta$和$\sigma$的值. 这其中, 骨骼点 (关键帧) 比骨架序列 (动作提名) 的粒度更小, 人体的行为信息更丰富, 因此本文将$\alpha$的值设定为不能小于$\beta$的值, 将$\delta$的值设定为不能小于$\sigma$的值. 基于这些经验分析, 即$\alpha\geqslant\beta\geqslant\delta\geqslant\sigma$, 本文将$\{\alpha, \beta, \delta, \sigma\}$的权重值分别设为以下组: {0.3 : 0.3 : 0.2 : 0.2}、{0.4 : 0.3 : 0.2 : 0.1}、{0.4 : 0.4 : 0.1 : 0.1}、{0.5 : 0.2 : 0.2 : 0.1}、{0.5 : 0.3 : 0.1 : 0.1}、{0.6 : 0.2 : 0.1 : 0.1}. 在这些不同权重下, MM-GCN 的识别结果如表 9 所示. 从表中可以看出当权重$\{\alpha, \beta, \delta, \sigma\}$的值设为为{0.4 : 0.3 : 0.2 : 0.1}时, 所提出的方法取得了最好的识别结果.

表 8 在不同阈值$\tau$下 R-GCNs 的识别结果 (%)

| 阈值 | Accuracy |
|---|---|
| 0.1 | 75.1 |
| 0.3 | 75.8 |
| 0.5 | 76.1 |
| 0.7 | **76.6** |
| 0.9 | 76.3 |

表 9 在不同权重下 MM-GCN 的识别结果对比

| 权重比例 | Accuracy |
|---|---|
| {0.3 : 0.3 : 0.2 : 0.2} | 94.7 |
| {0.4 : 0.3 : 0.2 : 0.1} | **94.9** |
| {0.4 : 0.4 : 0.1 : 0.1} | 94.8 |
| {0.5 : 0.2 : 0.2 : 0.1} | 94.7 |
| {0.5 : 0.3 : 0.1 : 0.1} | 94.7 |
| {0.6 : 0.2 : 0.1 : 0.1} | 94.6 |

© 中国科学院软件研究所 http://www.jos.org.cn

## 4 总 结

为了解决人口老龄化带来的空巢老人等社会问题, 国内外的学者们积极开展了针对老年人日常智能化看护的研究. 其中, 老年人日常行为智能分析与理解是一类关键性问题, 如老年人日常行为识别、预测等问题. 为了识别老年人日常行为, 本文提出了一种新的基于多模态多粒度图卷积神经网络的行为识别方法. 该方法基于多模态学习框架设计了一种“点-线-面-段”数据表示策略, 利用注意力图卷积对这 2 类模态、4 种颗粒度的数据同时进行建模, 学习细粒度的人体行为并获取高层语义特征完成老年人行为识别任务. 为了验证所提方法的有效性和鲁棒性, 在 ETRI-Activity3D 数据集和 NTU RGB+D 数据集上进行了对比实验. 实验结果表明本文所提出的方法在老年人行为识别与常规行为识别任务上均表现出了不错的性能. 但是该方法也存在一定的局限性, 在年龄段跨度较大的人体行为识别任务中的表现并不理想, 这将是未来工作中研究与探索的问题和方向.

**References:**


[1] López-Otín C, Blasco MA, Serrano M, Kroemer G. The hallmarks of aging. Cell, 2013, 153(6): 1194–1217. [doi: 10.1016/j.cell.2013.05.039]

[2] Sun ZJ, Xue L, Xu YM, Wang Z. Overview of deep learning. Application Research of Computers, 2012, 29(8): 2806–2810 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-3695.2012.08.002]

[3] Xi XF, Zhou GD. A survey on deep learning for natural language processing. Acta Automatica Sinica, 2016, 42(10): 1445–1465 (in Chinese with English abstract). [doi: 10.16383/j.aas.2016.c150682]

[4] Zhang S, Gong YH, Wang JJ. The development of deep convolution neural network and its applications on computer vision. Chinese Journal of Computers, 2019, 42(3): 453–482 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2019.00453]

[5] Zhu Y, Zhao JK, Wang YN, Zheng BB. A review of human action recognition based on deep learning. Acta Automatica Sinica, 2016, 42(6): 848–857 (in Chinese with English abstract). [doi: 10.16383/j.aas.2016.c150710]

[6] Kidd CD, Orr R, Abowd GD, Atkeson CG, Essa IA, Macintyre B, Mynatt E, Starner TE, Newstetter W. The aware home: A living laboratory for ubiquitous computing research. In: Proc. of the 1999 Int'l Workshop on Cooperative Buildings. Integrating Information, Organizations, and Architecture. Pittsburgh: Springer, 1999. 191–198. [doi: 10.1007/10705432_17]

[7] Jang J, Kim D, Park C, Jang M, Lee J, Kim J. ETRI-Activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly. In: Proc. of the 2020 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Las Vegas: IEEE, 2020. 10990–10997. [doi: 10.1109/IROS45743.2020.9341160]

[8] Veeriah V, Zhuang NF, Qi GJ. Differential recurrent neural networks for action recognition. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 4041–4049. [doi: 10.1109/ICCV.2015.460]

[9] Bruna J, Zaremba W, Szlam A, LeCun Y. Spectral networks and locally connected networks on graphs. arXiv:1312.6203, 2014.

[10] Yan R, Xie LX, Tang JH, Shu XB, Tian Q. HiGCIN: Hierarchical graph-based cross inference network for group activity recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020. [doi: 10.1109/TPAMI.2020.3034233]

[11] Wang WG, Shen JB, Jia YD. Review of visual attention detection. Ruan Jian Xue Bao/Journal of Software, 2019, 30(2): 416−439 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5636.htm [doi: 10.13328/j.cnki.jos.005636]

[12] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886−893. [doi: 10.1109/CVPR.2005.177]

[13] Lowe DG. Object recognition from local scale-invariant features. In: Proc. of the 7th IEEE Int'l Conf. on Computer Vision. Kerkyra: IEEE, 1999. 1150−1157. [doi: 10.1109/ICCV.1999.790410]

[14] Yan R, Xie LX, Tang JH, Shu XB, Tian Q. Social adaptive module for weakly-supervised group activity recognition. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 208–224. [doi: 10.1007/978-3-030-58598-3_13]

[15] Yan SJ, Xiong YJ, Lin DH. Spatial temporal graph convolutional networks for skeleton-based action recognition. arXiv:1801.07455, 2018.

[16] Li CL, Cui Z, Zheng WM, Xu CY, Yang J. Spatio-temporal graph convolution for skeleton based action recognition. arXiv:1802.09834, 2018.

[17] Lin CH, Chou PY, Lin CH, Tsai MY. SlowFast-GCN: A novel skeleton-based action recognition framework. In: Proc. of the 2020 IEEE Int'l Conf. on Pervasive Artificial Intelligence. Taipei: IEEE, 2020. 170–174. [doi: 10.1109/ICPAI51961.2020.00039]

[18] Feichtenhofer C, Fan HQ, Malik J, He KM. Slowfast networks for video recognition. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on


© 中国科学院软件研究所 http://www.jos.org.cn

Computer Vision. Seoul: IEEE, 2019. 6201–6210. [doi: 10.1109/ICCV.2019.00630]

[19] Gao XS, Li KQ, Zhang Y, Miao QG, Sheng LJ. Xie J, Xu JF. 3D skeleton-based video action recognition by graph convolution network. In: Proc. of the 2019 IEEE Int'l Conf. on Smart Internet of Things. Tianjin: IEEE, 2019. 500–501. [doi: 10.1109/SmartIoT.2019.00093]

[20] Li MS, Chen SH, Chen X, Zhang Y, Wang YF, Tian Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3590–3598. [doi: 10.1109/CVPR.2019.00371]

[21] Li B, Li X, Zhang ZF, Wu F. Spatio-temporal graph routing for skeleton-based action recognition. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conf. and the 9th AAAI Symp. on Educational Advances in Artificial Intelligence. Honolulu: AAAI, 2019. 8561–8568. [doi: 10.1609/aaai.v33i01.33018561]

[22] Shi L, Zhang YF, Cheng J, Lu HQ. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12018–12027. [doi: 10.1109/CVPR.2019.01230]

[23] Zhang JR, Shen FM, Xu X, Shen HT. Temporal reasoning graph for activity recognition. IEEE Trans. on Image Processing, 2020, 29: 5491–5506. [doi: 10.1109/TIP.2020.2985219]

[24] Shi XB, Li HW, Liu F, Zhang DY, Bi J, Li KZ. Graph convolutional networks with objects for skeleton-based action recognition. In: Proc. of the 2019 IEEE Int'l Conf. on Ubiquitous Computing & Communications and Data Science and Computational Intelligence and Smart Computing, Networking and Services. Shenyang: IEEE, 2019. 280–285. [doi: 10.1109/IUCC/DSCI/SmartCNS.2019.00074]

[25] Cheng K, Zhang YF, He XY, Chen WH, Cheng J, Lu HQ. Skeleton-based action recognition with shift graph convolutional network. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 180–189. [doi: 10.1109/CVPR42600.2020.00026]

[26] Du PF, Li XY, Gao YL. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 327–348 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[27] Wang PC, Li WQ, Wan J, Ogunbona P, Liu XW. Cooperative training of deep aggregation networks for RGB-D action recognition. In: Proc.of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 7404–7411.

[28] Liu MY, Yuan JS. Recognizing human actions as the evolution of pose estimation maps. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1159–1168. [doi: 10.1109/CVPR.2018.00127]

[29] Hu JF, Zheng WS, Lai JH, Zhang JG. Jointly learning heterogeneous features for RGB-D activity recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(11): 2186–2200. [doi: 10.1109/TPAMI.2016.2640292]

[30] Hu JF, Zheng WS, Pan JH, Lai JH, Zhang JG. Deep bilinear learning for RGB-D action recognition. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 346–362. [doi: 10.1007/978-3-030-01234-2_21]

[31] Liu TS, Kong J, Jiang M. RGB-D action recognition using multimodal correlative representation learning model. IEEE Sensors Journal, 2019, 19(5): 1862–1872. [doi: 10.1109/JSEN.2018.2884443]

[32] Li JN, Xie XM, Pan QZ, Cao YH, Zhao ZF, Shi GM. SGM-Net: Skeleton-guided multimodal network for action recognition. Pattern Recognition, 2020, 104: 107356. [doi: 10.1016/j.patcog.2020.107356]

[33] Shi L, Zhang YF, Cheng J, Lu HQ. Skeleton-based action recognition with directed graph neural networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7904–7913. [doi: 10.1109/CVPR.2019.00810]

[34] Gao X, Hu W, Tang JX, Liu JY, Guo ZM. Optimized skeleton-based action recognition via sparsified graph regression. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 601–610. [doi: 10.1145/3343031.3351170]

[35] Liu ZY, Zhang HW, Chen ZH, Wang ZY, Ouyang WL. Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 140–149. [doi: 10.1109/CVPR42600.2020.00022]

[36] Bartolomeo P. The attention systems of the human brain. In: Bartolomeo P, ed. Attention Disorders After Right Brain Damage. London: Springer, 2014. 1–19. [doi: 10.1007/978-1-4471-5649-9_1]

[37] Du WB, Wang YL, Qiao Y. RPAN: An end-to-end recurrent pose-attention network for action recognition in videos. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 3745–3754. [doi: 10.1109/ICCV.2017.402]

[38] Baradel F, Wolf C, Mille J, Taylor G W. Glimpse clouds: Human activity recognition from unstructured feature points. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 469–478. [doi: 10.1109/CVPR.2018.00056]

[39] Liu J, Wang G, Hu P, Duan LY, Kot AC. Global context-aware attention LSTM networks for 3D action recognition. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3671–3680. [doi: 10.1109/CVPR.2017.391]

© 中国科学院软件研究所 http://www.jos.org.cn

[40] Hu ZY, Lee EJ. Dual attention-guided multiscale dynamic aggregate graph convolutional networks for skeleton-based human action recognition. Symmetry, 2020, 12(10): 1589. [doi: 10.3390/sym12101589]

[41] Li D, Yao T, Duan LY, Mei T, Rui Y. Unified spatio-temporal attention networks for action recognition in videos. IEEE Trans. on Multimedia, 2019, 21(2): 416–428. [doi: 10.1109/TMM.2018.2862341]

[42] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4724–4733. [doi: 10.1109/CVPR.2017.502]

[43] Kim TS, Reiter A. Interpretable 3D human action analysis with temporal convolutional networks. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1623–1631. [doi: 10.1109/CVPRW.2017.207]

[44] Lin TW, Zhao X, Su HS, Wang CJ, Yang M. BSN: Boundary sensitive network for temporal action proposal generation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 3–21. [doi: 10.1007/978-3-030-01225-0_1]

[45] Shahroudy A, Liu J, Ng TT, Wang G. NTU RGB+D: A large scale dataset for 3D human activity analysis. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1010–1019. [doi: 10.1109/CVPR.2016.115]

[46] Li S, Li WQ, Cook C, Zhu C, Gao YB. Independently recurrent neural network (indrnn): Building a longer and deeper RNN. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5457–5466. [doi: 10.1109/CVPR.2018.00572]

[47] Wang HS, Wang L. Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. IEEE Trans. on Image Processing, 2018, 27(9): 4382–4394. [doi: 10.1109/TIP.2018.2837386]

[48] Li C, Xie CY, Zhang BC, Han JG, Zhen XT, Chen J. Memory attention networks for skeleton-based action recognition. IEEE Trans. on Neural Networks and Learning Systems, 2021. [doi: 10.1109/TNNLS.2021.3061115]

[49] Xu YY, Cheng J, Wang L, Xia HY, Liu F, Tao DP. Ensemble one-dimensional convolution neural networks for skeleton-based action recognition. IEEE Signal Processing Letters, 2018, 25(7): 1044–1048. [doi: 10.1109/LSP.2018.2841649]

[50] Li C, Zhong QY, Xie D, Pu SL. Skeleton-based action recognition with convolutional neural networks. In: Proc. of the 2017 IEEE Int'l Conf. on Multimedia & Expo Workshops. Hong Kong: IEEE, 2017. 597–600. [doi: 10.1109/ICMEW.2017.8026285]

[51] Li C, Zhong QY, Xie D, Pu SL. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: ACM, 2018. 786–792. [doi: 10.5555/3304415.3304527]

[52] Wen YH, Gao L, Fu HB, Zhang FL, Xia SH. Graph CNNs with motif and variable temporal block for skeleton-based action recognition. In: Proc. of the 2019 AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 8989–8996. [doi: 10.1609/aaai.v33i01.33018989]

**附中文参考文献**:

[2] 孙志军, 薛磊, 许阳明, 王正. 深度学习研究综述. 计算机应用研究, 2012, 29(8): 2806–2810. [doi: 10.3969/j.issn.1001-3695.2012.08.002]

[3] 奚雪峰, 周国栋. 面向自然语言处理的深度学习研究. 自动化学报, 2016, 42(10): 1445–1465. [doi: 10.16383/j.aas.2016.c150682]

[4] 张顺, 龚怡宏, 王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用. 计算机学报, 2019, 42(3): 453–482. [doi: 10.11897/SP.J.1016.2019.00453]

[5] 朱煜, 赵江坤, 王逸宁, 郑兵兵. 基于深度学习的人体行为识别算法综述. 自动化学报, 2016, 42(6): 848–857. [doi: 10.16383/j.aas.2016.c150710]

[11] 王文冠, 沈建冰, 贾云得. 视觉注意力检测综述. 软件学报, 2019, 30(2): 416–439. http://www.jos.org.cn/1000-9825/5636.htm [doi: 10.13328/j.cnki.jos.005636]

[26] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述. 软件学报, 2021, 32(2): 327–348. http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

© 中国科学院软件研究所 http://www.jos.org.cn

**丁静**(1997－), 女, 硕士生, 主要研究领域为视频行为识别.

**舒祥波**(1986－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为图像视频内容分析, 多媒体分析, 计算机视觉.

**黄捧**(1996－), 男, 博士生, 主要研究领域为视频行为识别.

**姚亚洲**(1987－), 男, 博士, 教授, CCF 专业会员, 主要研究领域为多媒体技术, 计算机视觉, 机器学习.

**宋砚**(1983－), 女, 博士, 副教授, 主要研究领域为多媒体内容分析, 视频内容理解, 计算机视觉.

© 中国科学院软件研究所 http://www.jos.org.cn