基于视觉的人体动作质量评价研究综述

沈媛媛 [1] 张燕明 [2] 沈燕飞 [1]

摘 要 基于视觉的人体动作质量评价利用计算机视觉相关技术自动分析个体运动完成情况, 并为其提供相应的动作质量评价结果. 这已成为运动科学和人工智能交叉领域的一个热点研究问题, 在竞技体育、运动员选材、健身锻炼、运动康复等领域具有深远的理论研究意义和很强的实用价值. 本文将从数据获取及标注、动作特征表示、动作质量评价 3 个方面对涉及到的技术进行回顾分析, 对相关方法进行分类, 并比较分析不同方法在 AQA-7、JIGSAWS、EPIC-Skills 2018 三个数据集上的性能. 最后讨论未来可能的研究方向.

关键词 动作质量, 评价, 计算机视觉, 信息获取, 特征表示, 损失函数

引用格式 沈媛媛, 张燕明, 沈燕飞. 基于视觉的人体动作质量评价研究综述. 自动化学报, 2025, **51**(2): 404−426

DOI 10.16383/j.aas.c230551 **CSTR** 32138.14.j.aas.c230551

A Survey of Vision-based Motion Quality Assessment

SHEN Yuan-Yuan[1] ZHANG Yan-Ming[2] SHEN Yan-Fei[1]

Abstract Vision-based motion quality assessment utilizes computer vision techniques to analyze the quality of individual movement behavior automatically and provide the corresponding assessments of movement quality. It has gradually become the hot issue at the intersection of the sport science and artificial intelligence, and has widely used in the fields of sporting events, athlete selection, fitness and rehabilitation. This article conducts a retrospective analysis of the involved technologies from three aspects: Data acquisition and annotation, motion feature representation, and motion quality assessment. It categorizes and compares various mainstream methods on three datasets: AQA-7, JIGSAWS, and EPIC-Skills 2018. Finally, potential future research directions are discussed.

Key words Motion quality, assessment, computer vision, data acquisition, feature representation, loss function

Citation Shen Yuan-Yuan, Zhang Yan-Ming, Shen Yan-Fei. A survey of vision-based motion quality assessment. *Acta Automatica Sinica*, 2025, **51**(2): 404−426

随着视觉数据采集设备日渐普及和计算机技术的持续发展, 基于视觉的人体行为理解与分析任务已经取得了长足的发展. 早期研究工作多数局限在动作识别[1] 任务上, 即对视觉数据中包含的人体动作类别进行识别. 该类任务接收一个视觉动作数据 (如视频片段) 作为输入, 输出一个动作类别, 表示主体正在进行的动作. 然而, 在实际应用场景中, 常常还需要模型具有对人体执行的动作质量进行客观量化的评价和分析能力, 进而帮助提升运动技能、保障比赛公平、促进改善健康等. 因此, 基于视觉的人体动作质量评价任务被提出并受到研究关注.

基于视觉的动作质量评价[2−3] 首先使用运动相机、深度相机等视觉数据采集设备记录人体的动作过程, 形成 RGB 视频、骨架序列等运动数据; 然后借助计算机视觉[4−5] 与深度学习[6−7] 相关技术对运动数据进行处理分析, 学习并理解其中人的动作和行为; 最终获得动作评分、评级或者排序结果, 使得计算机能够像人类专家一样去“评价”动作的完成质量.

人体动作质量评价在竞技体育、运动员选材、健身锻炼、运动康复等领域都有着广泛应用. 传统的动作质量评价过程依靠领域专家观察运动者的完整动作, 并结合先验知识对运动表现进行人工判断. 然而, 有限的专家资源难以满足不断增长的运动需求, 且专家评价也难以做到绝对的公正客观. 因此, 研究基于视觉的动作质量评价方法有助于改进传统的人体动作质量评价手段, 在提升评价效率和准确性等方面具有重要的应用价值.

人体兼具刚性和柔性物体的特性, 其运动过程也极其复杂. 动作特征表示易受多种因素的影响,

收稿日期 2023-09-05 录用日期 2024-08-07
Manuscript received September 5, 2023; accepted August 7, 2024
北京市自然科学基金 (9234029), 国家自然科学基金 (72071018), 中央高校基本科研业务费专项资金 (2024JCYJ004) 资助
Supported by Natural Science Foundation of Beijing (9234029), National Natural Science Foundation of China (72071018), and Fundamental Research Funds for the Central Universities (2024JCYJ004)
本文责任编委 郑伟诗
Recommended by Associate Editor ZHENG Wei-Shi
1. 北京体育大学体育工程学院 北京 100084 2. 中国科学院自动化研究所模式识别国家重点实验室 北京 100190
1. School of Sport Engineering, Beijng Sport University, Beijing 100084 2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190

包括光照、遮挡、环境条件、着装、姿态、视角变化、运动速度、人体形态变化以及复杂互动等. 此外, 人体动作质量评价属于典型的细粒度学习任务: 不同运动者在执行相同的动作时通常呈现出相似的姿态和运动轨迹, 即运动模式之间存在极大的相似性, 评价结果的差异主要体现在局部动作模式上, 比如在跳水运动中双腿是否保持笔直, 或者在康复训练中运动角度是否符合标准等方面. 因此, 在人体动作质量评价任务中, 建立稳健的动作特征表示并设计适用于细粒度评价任务的模型对于构建高性能的动作质量评价方法至关重要. 基于视觉的人体动作质量评价研究具有重要的理论价值.

近年来, 国内外众多的研究机构, 如美国的麻省理工学院[8]、约翰斯霍普金斯大学[9]、英国的布里斯托大学[10]、中国科学院[11]、清华大学[12]、北京大学[13]等都对基于视觉的人体动作质量评价进行了广泛而深入的研究. 鉴于基于视觉的人体动作质量评价应用广泛, 国内外相关综述文献较少且主要针对特定的应用领域, 因此有必要对相关研究进行全面和系统的梳理. 本文拟对已有研究成果进行调研综述, 对比分析相关方法的优缺点, 并给出不同方法在同一数据集上的性能比较, 为后续的研究工作提供参考.

基于视觉的人体动作质量评价大体上可以划分为动作数据获取[14]、动作特征表示[15]和动作质量评价[16]三个阶段:

1) 动作数据获取. 动作数据的获取为基于视觉的动作质量评价研究提供了数据基础. 随着视觉传感技术的不断发展, 获取基于视觉的动作数据变得更加容易. 然而, 收集大规模的动作质量评价数据仍然面临挑战. 为了建立鲁棒的模型, 通常需要构建多样性的动作质量数据集, 其中包括各种类型的动作、不同水平的运动者以及不同环境条件下的动作数据. 这增加了数据采集的复杂性和耗时性. 此外, 由于动作质量的评价通常需要多名领域专家的共同参与, 不同的专业人员可能在评价同一动作时产生不同的结果. 因此, 需要通过质量控制方法来确保评分过程的准确性和一致性.

2) 动作特征表示. 动作特征表示是研究该任务的关键环节, 它综合了图像和动作等多方面的信息, 用于设计特征向量以描述人体的运动过程. 通过这一过程, 原始数据被转换为具有区分性的向量表示形式, 从而有助于后续的评价任务. 其主要的技术挑战在于: 动作质量不仅取决于动作的静止姿态是否符合技术执行标准, 还依赖于动作的协调性、流畅性等动态特征. 这要求动作特征表示能够描述身体各部位间复杂的时间、空间关系. 同时, 由于动作评价任务的细粒度特点, 动作特征必须具备很强的鉴别性, 能够有效地抽取和表示细粒度差异.

3) 动作质量评价. 在经过特征提取和处理后, 需要构建动作质量评价模型, 用于将提取的特征与相应的评价目标关联起来, 使得动作数据映射到正确的分数、类别或排序信息上. 由于不同的动作质量评价任务具有各自不同的评价标准, 因此可能需要设计不同的损失函数, 以确保模型能够达到所期望的目标. 同时, 动作质量评价任务中存在着评分专家主观因素引起的不确定性, 以及需要考虑动作之间的细粒度差异等挑战因素, 这些问题在动作质量评价过程中都需要进行合理的建模来解决.

上述各不同阶段的主要任务及存在的问题如表 1 所示. 图 1 列举了动作特征表示和动作质量评价两个阶段所介绍的不同方法及其解决的主要问题.

本文结构如下: 第 1 节至第 3 节分别围绕视觉动作质量评价任务的关键步骤展开介绍, 包括动作数据获取、动作特征表示以及动作质量评价; 第 4 节详细阐述了现有方法在典型动作质量评价数据集上的性能评估结果, 并对实验结果进行了详细分析; 第 5 节作为总结, 概括了全文的主要内容, 并提出了未来的研究展望.

1　动作数据获取

基于视觉的人体动作质量评价技术需要利用高质量的动作数据来保障训练和提供性能评估, 因此如何获取动作数据为后续的动作质量评价任务研究提供了基础支撑. 根据使用者是否自主采集数据,

表 1　基于视觉的动作质量评价方法不同阶段的主要任务及存在的问题

Table 1　Main tasks and existing challenges in different stages of vision-based motion quality assessment

阶段	主要任务	存在的问题
动作数据获取	通过视觉传感器来收集和记录与动作相关的数据 (RGB、深度图、骨架序列)	如何根据不同的应用场景选择适用的数据模态? 如何确保专家的评分质量?
动作特征表示	综合利用静态图像和人体动作等多方面信息, 设计具有区分性的特征向量以描述人体的运动过程	如何根据动作质量评价任务本身的特性学习具有强鉴别性的动作特征, 以有效地抽取和表示不同运动者在执行相同动作时的细微差异?
动作质量评价	设计特征映射方式, 将提取的特征与相应的评分、评级或排序评价目标关联起来	如何在设计损失函数时考虑标注不确定性 (如不同专家的评分差异)、同一动作之间的评分差异等问题?

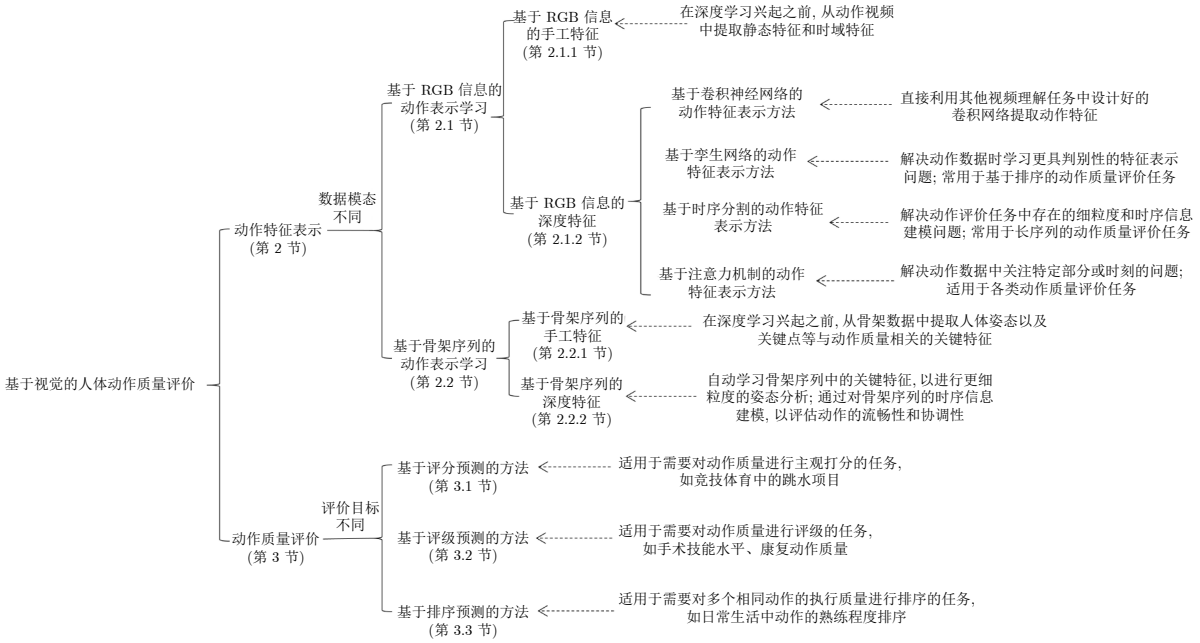


图 1 文中总结的不同方法及其解决的主要问题

Fig. 1 Different methods summarized in this article and the main issues they address

常见的动作数据获取方法可以被划分为基于数据采集的方法和从公开数据集获取的方法. 本节分别对上述两类动作数据获取方法进行介绍, 并对两类方法的优缺点进行分析和讨论.

1.1 基于数据采集的方法

随着视觉传感技术的不断进步, 基于视觉的数据采集方法在人体动作信息获取方面得到了广泛应用. 这类方法使用无接触的视觉传感器设备, 如运动相机、双目相机、Kinect 相机等, 能够实时捕获包含人体运动过程的二维或三维数据.

二维动作数据是通过普通的数码相机、摄像机或者 RGB 摄像头等 2D 相机获取的. 这类数据用于记录人体在拍摄场景中的运动姿态和动作过程, 形成 RGB 动作图像或 RGB 图像序列构成的视频动作数据. 在 RGB 图像或视频中, 人体动作会被投影在二维平面上. 由于收集数据方便且处理较为容易, 二维动作数据已广泛应用于基于视觉的动作质量评价任务中, 如体育赛事[8]、技能训练[17]、健身锻炼[18] 等.

三维动作数据 (如带有深度信息的 RGB-D 图像、3D 骨架数据等) 的获取通常依赖深度相机、立体摄像头、相机阵列等传感器设备. 在三维动作数据中, 除 RGB 图像所包含的颜色空间信息外, 还能够提供与目标对象表面距离有关的深度信息, 因此能够更全面地反映人体在三维空间中的运动轨迹和姿态. 根据是否需要在被测对象身体上放置特定位置的标记点或传感器, 通常将三维动作数据的采集方法进一步划分为有标记点的动作捕捉方法和无标记点的动作捕捉方法. 其中, 基于标记点的动作捕捉方法利用传感器系统跟踪标记点的运动, 并利用计算机算法分析位置, 从而确定对象的运动轨迹. 该类方法通常能够提供相对较高的精确度, 尤其适合于需要准确测量关节角度及身体位置的研究应用, 如运动员的训练动作评价[19]、康复动作评价[20] 等. 基于无标记点的动作捕捉方法从传感器数据中直接提取人体的三维运动信息. 由于不需要在人体上放置标记点, 因此这类方法在数据获取时更加灵活, 能够更好地适应不同的运动场景, 包括居家运动中的动作质量评价[21−22]. 尽管三维数据在描述运动的空间结构和深度信息上具有天然优势, 但其在算法处理过程中往往需要依靠更高的算力及更加复杂的模型. 此外, 由于人体动作的复杂性、多变性及存在遮挡与自遮挡等问题, 基于视觉的动作信息获取通常采用多相机从不同角度捕获人体的运动行为, 并依据拍摄目标调整相机位置.

为了训练高效的动作质量评价模型, 数据样本中通常还应包括对动作数据的标注信息. 常见的标注信息包括:

1) 评分标注[8]. 针对执行者的具体动作, 专家对其表现进行打分. 标注结果为最低分到最高分之间的某个连续数值, 常应用于对竞技体育中的主观打

分项目进行标注.

2) 评级标注[23]. 针对执行者的具体动作, 专家对其等级进行评估. 标注结果为若干个离散等级中的某一个, 可应用于对参与者技能水平 (如舞蹈动作的评级) 或执行动作的等级 (如康复动作的标准程度) 进行标注.

3) 排序标注[24]. 针对多个同类型的动作, 排序标注给出不同动作质量的排序信息.

4) 反馈标注[8]. 在运动执行过程中, 专家给予运动者更加详细的反馈信息, 如动作中存在的问题以及改进动作的方向.

在动作质量评价任务中, 确保标注信息的准确性至关重要. 因此, 通常需要参与标注的人员是特定运动领域的专家, 他们能够深刻理解任务的上下文和领域特定的要求. 在进行多专家标注任务时, 同样需要确保所有参与的专家在进行动作标注时遵循一致的评价标准.

1.2 从公开数据集获取的方法

在评估不同人体动作质量评价方法的性能方面, 现在已经有部分公开的动作质量评价数据集供研究者使用. 按照应用场景不同, 本文对常用的公开数据集进行分类梳理, 相关结果见表 2 所示.

表 2 统计的公开数据集共有 34 项, 几乎都集中于 2010 年之后发布. 数据集中包括滑雪、跳水、艺术体操、深蹲、坐立、抓取、打结、杠铃等多种动作, 涉及体育赛事、运动康复、技能训练、健身锻炼四个不同应用场景. 其中, 与体育赛事相关的数据集数量最多. 这些数据集多数基于 RGB 视频构建, 覆盖了从几秒钟的跳水动作到几分钟的滑雪动作等不同的动作时长. 由于在之前的奥林匹克等国际赛事中已经积累了丰富的历史比赛视频素材和专家评分信息, 因此此类数据集通常是基于已有的比赛素材构建而来. 其评价目标主要集中在对视频中运动员的动作质量进行评分. 在运动康复领域, 大多数数据集基于 3D 骨架数据构建. 3D 骨架数据具备提供关节在三维空间中的坐标信息以及姿态深度信息的能力, 这对于准确理解和评估康复动作的执行情况至关重要. 动作质量评价的目标通常是对康复动作的质量等级进行评估. 技能训练的数据集大多基于网络收集的 RGB 视频构建, 常用的评价目标包括对技能水平的评级以及成对视频中技能动作水平高低的排序. 现有的健身锻炼数据集既包括了通过 3D 相机采集的骨架数据, 又包括了来自网络收集的 RGB 视频, 而评价目标也更加多样化, 包含了评级、评分和排序三种不同的方式.

下面将介绍目前研究中使用最为广泛的三个公开数据集:

1) AQA-7 数据集[30] 于 2019 年发布, 共包含 1 189 段比赛视频. 涵盖了竞技体育中的七项运动项目, 包括单人 10 米跳水 (Diving)、体操跳马 (Vault)、越野滑雪 (Skiing)、单板滑雪 (Snowboard)、双人 3 米跳水 (Sync. 3 m)、双人 10 米跳水 (Sync. 10 m) 和蹦床 (Trampoline). 该数据集提供了运动员比赛的视频记录以及相应的得分数据, 是目前应用广泛的主观评分数据集之一.

2) JIGSAWS 数据集[9, 46] 是一个收集了外科医生手术数据并进行详细标注的重要资源, 于 2014 年发布. 在动作质量评价领域的文献中, 这个数据集已经得到了广泛的评估和应用. 该数据集包括了 8 名外科医生参与的三项基本外科技能任务, 分别为缝合 (Suturing, SU)、打结 (Knot-tying, KT) 以及穿针 (Needle-passing, NP), 每项任务都进行了 5 次独立重复. 经过去除损坏的数据后, 最终获得了 39 次 SU 试验, 36 次 KT 试验和 28 次 NP 试验数据. 所获得的数据由两部分组成: 一是由达芬奇外科手术系统记录的手术数据, 包括运动学数据和视频数据两种模态; 二是外科专家对手术数据的标注信息, 包括动作分割标注和技能水平标注两部分. 按照外科医生自身手术经验情况 (Self-claimed) 或全球评级得分 (Global rating score, GRS), 技能水平被划分为初级、高级两类等级或初级、中级和高级三类等级.

3) EPIC-Skills 2018[24] 数据集发布于 2018 年, 由 216 个动作视频对组成. 对于每个视频对, 标注信息指示哪个视频中包含的动作质量更高. 每对视频由四位标注者标记, 仅选取获得一致性标注的样本构成最终的数据集. 该数据集包括了 Chopstick-using、Surgery、Drawing 和 Dough-rolling 四类动作任务.

总体而言, 人体动作质量评价是一个较新的研究方向, 数据集的构建也正处于逐步完善的阶段. 相比人类庞大的动作库, 现有的动作质量评价数据集也仅覆盖了有限的几种动作类别. 此外, 现公开的数据集多数规模较小, 单个数据集通常仅包括有限的几类动作, 并且每类动作的样本数较为有限. 这主要是因为动作质量评价数据集需要特定领域的专家进行标注, 而专家资源又相当稀缺.

1.3 小结

本节对不同的动作数据获取方法进行归纳总结. 根据使用者是否自主采集数据, 常见的动作数据

表 2 主流的动作质量评价数据集总览

Table 2 Brief overview of mainstream motion quality assessment dataset

数据集	动作类别	样本数 (受试者人数)	标注类别	应用场景	数据模态	发表年份
Heian Shodan[25]	1	14	评级标注	健身锻炼	3D 骨架	2003
FINA09 Dive[26]	1	68	评分标注	体育赛事	RGB 视频	2010
MIT-Dive[8]	1	159	评分标注、反馈标注	体育赛事	RGB 视频	2014
MIT-Skate[8]	1	150	评分标注	体育赛事	RGB 视频	2014
SPHERE-Staircase2014[10]	1	48	评级标注	运动康复	3D 骨架	2014
JIGSAWS[9]	3	103	评级标注	技能训练	RGB 视频、运动学数据	2014
SPHERE-Walking2015[16]	1	40	评级标注	运动康复	3D 骨架	2016
SPHERE-SitStand2015[16]	1	109	评级标注	运动康复	3D 骨架	2016
LAM Exercise Dataset[23]	5	125	评级标注	运动康复	3D 骨架	2016
First-Person Basketball[27]	1	48	排序标注	健身锻炼	RGB 视频	2017
UNLV-Dive[28]	1	370	评分标注	体育赛事	RGB 视频	2017
UNLV-Vault[28]	1	176	评分标注	体育赛事	RGB 视频	2017
UI-PRMD[20]	10	100	评级标注	运动康复	3D 骨架	2018
EPIC-Skills 2018[24]	4	216	排序标注	技能训练	RGB 视频	2018
Infant Grasp[29]	1	94	排序标注	技能训练	RGB 视频	2019
AQA-7[30]	7	1 189	评分标注	体育赛事	RGB 视频	2019
MTL-AQA[31]	1	1 412	评分标注	体育赛事	RGB 视频	2019
FSD-10[32]	10	1 484	评分标注	体育赛事	RGB 视频	2019
BEST 2019[32]	5	500	排序标注	技能训练	RGB 视频	2019
KIMORE[22]	5	78	评分标注	运动康复	RGB、深度视频、3D 骨架	2019
Fis-V[33]	1	500	评分标注	体育赛事	RGB 视频	2020
TASD-2(SyncDiving-3m)[34]	1	238	评分标注	体育赛事	RGB 视频	2020
TASD-2(SyncDiving-10m)[34]	1	368	评分标注	体育赛事	RGB 视频	2020
RG[35]	4	1 000	评分标注	体育赛事	RGB 视频	2020
QMAR[36]	6	38	评级标注	运动康复	RGB 视频	2020
PISA[37]	1	992	评级标注	技能训练	RGB 视频、音频	2021
FR-FS[38]	1	417	评分标注	体育赛事	RGB 视频	2021
SMART[39]	8	640	评分标注	体育赛事、健身锻炼	RGB 视频	2021
Fitness-AQA[40]	3	1 000	反馈标注	健身锻炼	RGB 视频	2022
Finediving[41]	1	3 000	评分标注	体育赛事	RGB 视频	2022
LOGO[42]	1	200	评分标注	体育赛事	RGB 视频	2023
RFSJ[43]	23	1 304	评分标注	体育赛事	RGB 视频	2023
FineFS[44]	2	1 167	评分标注	体育赛事	RGB 视频、骨架数据	2023
AGF-Olympics[45]	1	500	评分标注	体育赛事	RGB 视频、骨架数据	2024

获取方法可以被划分为基于数据采集的方法和从公开数据集获取的方法. 总体而言, 基于数据采集的方法能够根据应用需求灵活采集动作数据, 包括自主地选择数据采集设备、设计动作方案及选用数据格式. 然而, 收集并构建一个足够大的数据集往往也耗费大量的人力、物力和时间成本. 数据捕获过程需要专业的设施以及足够的受试者, 而精确的标签信息则需要领域专家对采集的数据进行逐条标注. 因此, 多数学者仍然依赖通用的公开数据集资源展开相关研究工作. 公开数据集主要涉及体育赛事、运动康复、技能训练、健身锻炼等应用场景. 近年来, 虽然已经发布了不少公开数据集用于动作质量评价任务的研究, 但多数数据集规模仍然较小, 且每类动作的样本数较为有限.

2 动作特征表示

动作特征表示旨在从底层的人体动作数据中抽取部分具有代表性的特征信息, 以对运动过程进行表征. 在人体动作质量评价任务中, 由于面临着动作复杂性、视觉差异细微、时序信息捕获、背景噪声

等多方面的困难, 需要研究人员针对这些挑战提出有效鲁棒的特征表示方法. 根据采用的数据模态不同, 动作特征表示大体上可以划分为基于 RGB 信息的方法[11, 29, 47] 和基于骨架序列[48−50] 的方法. 前者将原始视频数据看作序列化的图像集, 尝试从已有图像特征技术出发提取形状、颜色、纹理、姿态等静态特征, 并结合视频的时域特性学习更具区分性和表达能力的动作特征表示; 后者从人体动作序列中提取包含局部关节点的骨架序列, 并融合运动过程中的时空信息学习动作特征表示. 此外, 也有研究工作尝试先从 RGB 视频数据中提取骨架序列作为特征表示[50], 进而采用基于骨架序列的动作质量评价方法. 接下来分别对基于 RGB 信息的动作表示学习方法和基于骨架序列的动作表示学习方法进行介绍.

2.1 基于 RGB 信息的动作表示学习

基于 RGB 信息的动作表示学习将视频视为一系列 RGB 图像帧组成的序列, 通过从 RGB 图像序列中抽取静态特征和运动信息来表示动作. 按照特征的学习方法不同来划分, 基于 RGB 信息的动作表示学习又可以进一步划分为手工特征表示和深度特征表示.

2.1.1 基于 RGB 信息的手工特征

在早期的计算机视觉动作质量评价研究中, 由于计算机视觉和深度学习技术尚未成熟, 研究人员主要根据评价任务特性手动设计和选择合适的特征来表示和描述动作. 这些特征通常包括静态特征和时域特征, 分别用于捕捉动作中逐帧的空间结构和时序变化情况. 该类方法从原始视频数据中提取人体关键点、时空兴趣点等轨迹时序数据, 然后对这些时序数据进行特征建模, 以捕获运动估计信息. Gordon[51] 提出了一种利用运动跟踪算法提取人体的运动轨迹信息作为特征表示的方法, 从而捕捉视频中目标的运动模式. 通过将提取的特征与体操动作的规则进行匹配, 他们实现了自动化地分析和评价运动员的体操动作质量. Pirsiavash 等[8] 结合时间维度, 将 Gabor 滤波器扩展到时空维度, 进而同时捕获视频动作的空间结构 (梯度) 和时间变化 (速度) 信息. 为了进一步获得动作改进的反馈信息, 文中同时采用了另一种高级人工特征——先提取姿态轨迹表征人体的运动行为, 再使用离散余弦变换 (DCT) 提取低频特征进行去噪处理. 基于运动者的姿态信息, Venkataraman 等[52] 提出采用近似熵对人体运动进行建模, 该方法能够同时编码各关节以及关节之间的动态信息. Zia 等[53] 从视频中提取时空兴趣点 (STIP) 特征形成动作的时序表示, 并采用离散建模、纹理建模、频率建模三种不同的方式对形成的时序数据进行动作特征建模.

总而言之, 这些研究工作多数是通过跟踪运动轨迹来进一步获取局部信息, 从而表示动作的细节特征. 虽然人工设计和提取的特征通常适用于简单的动作质量评价任务, 并具有良好的适应性, 但难以有效地表示多样变化的复杂动作特征. 在实际应用中, 这类特征表示方法往往面临着不具有良好泛化性的问题.

2.1.2 基于 RGB 信息的深度特征

随着深度学习等技术的发展, 动作质量评价任务中越来越多的研究工作倾向于使用自动学习特征的方法, 尤其是基于深度神经网络的方法, 来更好地表示和描述复杂动作的特征. 这些自动学习的方法可以更好地从原始数据中学习特征, 并捕捉不同执行者之间的细微差异和个体特征. 根据动作质量评价文献中使用的动作特征表示主体网络架构不同, 本文将基于 RGB 信息的深度特征归纳为如下 4 种: 基于卷积神经网络的动作特征表示方法, 基于孪生网络的动作特征表示方法, 基于时序分割的动作特征表示方法, 基于注意力机制的动作特征表示方法.

1) 基于卷积神经网络的动作特征表示方法

由于卷积神经网络 (CNN) 在处理图像序列和时间序列数据方面的有效性和明显优势, 基于 CNN 的动作特征表示方法已经被广泛用于动作质量评价任务中[54]. 通过卷积层、池化层和全连接层等组件有效捕捉图像或时间序列数据中的空间和时间信息, CNN 能够用来提取特定动作的判别特征.

图 2 给出一个典型的基于 CNN 的动作质量评价方法框架. 该类方法首先将整段视频动作切分成若干等长的视频片段, 其次利用卷积神经网络对每段视频片段从局部到全局逐层提取深度特征, 最后利用特征聚合方法获取整段动作视频的特征表示.

其中, CNN 结构主要包括以下四种类型:

a) 双流 CNN (2S-CNN)[24, 55]

双流 CNN 由 Simonyan 和 Zisserman 提出, 最早应用于动作识别任务中的特征表示方法. 该方法由两个并行的 CNN 组成: 一个处理视频的空间信息, 即静态图像帧的内容; 另一个处理视频的时间信息, 即连续视频帧之间的运动信息. 通过融合这两个网络的特征, 双流 CNN 能够更全面地表示动作的时空关系.

b) 三维卷积神经网络 (3D convolutional networks, C3D)[30, 33, 56]

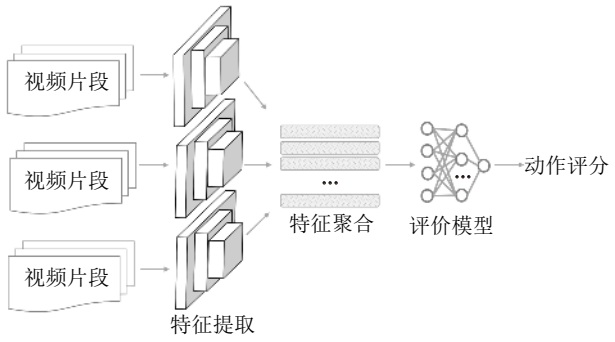


图 2 卷积神经网络的动作质量评价方法框架

Fig. 2 A CNN framework for motion quality assessment

C3D 是 Tran 等[56] 提出的一种能够同时处理时域和空域信息的深度学习网络. 其采用多个 3D 卷积层、3D 池化层对连续视频帧的各个通道数据进行处理, 并且组合最终的通道信息得到时空特征描述子. 利用 3D 卷积和 3D 池化操作, C3D 能够直接针对连续视频帧组成的视频块进行特征提取, 同时捕获到时间维度和空间维度的特征信息. 具体地, 假设第 l 层网络的第 m 个特征映射图上 (x, y, z) 点处的特征值记作 a_{lm}^{xyz}, 如下式所示:

$$a_{lm}^{xyz} = f\left(b_{lm} + \sum_{n=1}^{N}\sum_{i=0}^{h_l-1}\sum_{j=0}^{w_l-1}\sum_{k=0}^{t_l-1} s_{lm}^{ijk} a_{(l-1)n}^{(x+i)(y+j)(z+k)}\right) \tag{1}$$

其中, b_{lm} 是特征映射的偏置值, N 为第 $l-1$ 层网络的特征映射图个数, h_l、w_l 和 t_l 分别是第 l 层三维卷积核的高度、宽度和时域维度大小, s_{lm} 表示连接上层特征映射的三维卷积核权重, $f(\cdot)$ 表示激活函数.

c) 双流膨胀三维卷积网络 (Two-stream inflated 3D ConvNets, I3D)[12, 32, 57]

尽管 C3D 网络适合对视频中的时序信息建模, 然而相比 2D 网络, 其参数量和计算量均有大幅度增加. 为了能够充分利用已有的 2D 网络结构和图像预训练模型, I3D 网络结合 3D 卷积和双流模型学习高效的时空特征表示. 具体地, I3D 网络采用 3D 卷积分别对 RGB 图和光流图进行训练, 并融合双流网络的两个分支输出作为最终结果. 该网络从已有 2D 网络结构出发, 在滤波器和池化层增加时间维度, 进而将其转换为 3D 卷积网络. 此外, I3D 提出了一种新的 3D 滤波器初始化策略, 该策略首先将 2D 网络中预训练的滤波器权重沿着时间维度重复多次, 其次对滤波器权重进行归一化, 最后将归一化后的权重作为 3D 滤波器的初始权重.

d) 伪三维卷积网络 (Pseudo-3D, P3D)[58–59]

P3D 网络在残差学习框架下设计了一系列经济有效的三维卷积基础构建块, 进而缓解了 3D 网络结构计算复杂度高、参数量多产生的问题. 具体而言, 该网络将大小为 $h\times w\times t$ ($h\times w$ 为卷积核的空间大小, t 为卷积核的时序长度) 的三维卷积核分解为一个 $1\times h\times w$ 的二维空间卷积核和一个 $t\times 1\times 1$ 的一维时间卷积核, 并基于两类卷积核之间的影响关系设计三种不同的基本块结构. 相比同样深度的二维卷积网络, P3D 仅仅增加了一定数量的一维卷积, 在参数数量、运行耗时小幅度增长的前提下将二维卷积核扩展到三维的应用场景下使用. 此外, P3D 能够利用已存在的图像预训练模型.

2S-CNN、C3D、I3D 和 P3D 网络均是从视频片段中提取特征表示, 接下来还需要对来自不同视频片段的特征进行聚合进而获取整个运动视频的特征表示, 常用的方法包括:

a) 均值化[28, 30–31]. 该方法直接对多个不同视频片段的特征求平均值作为最终的视频特征表示, 因简单高效被应用的较为广泛.

b) LSTM 层[30–31]. 由于能够对不同视频片段特征的时序关系进行建模, LSTM 结构能够发现时序数据中距离较远的长期依赖关系.

c) 其他网络结构. 如 Li 等[48] 专门设计了一个由两个卷积层组成的网络学习视频特征表示. Parmar 等[31] 采用了基于空洞卷积的多尺度上下文特征聚合方法[60].

2) 基于孪生网络的动作特征表示方法

在动作质量评价任务中, 每个样本代表一个执行者完成的某个动作. 不同执行者之间可能在动作的细节、姿态、速度等方面存在细微差异, 因此需要学习对细粒度任务更具判别性的特征表示. 孪生网络[61] 能够比较不同执行者动作之间的微小差异, 从而有助于提高动作质量评价任务的准确性和可靠性.

孪生网络是一种特殊的神经网络架构, 它由两个结构相同且共享权重的子网络组成. 在动作质量评价任务中, 这两个子网络同时接收不同执行者执行同一种动作的视频数据作为输入. 通过共享权重的子网络, 不同执行者的同一动作样本被映射到同一个特征空间, 并使用某种距离函数计算特征向量之间的距离. 这种设计使得孪生网络能够对比和度量不同执行者执行的同一动作的相似性或差异性, 从而能够评价动作的质量.

为了对视频中出现的动作技能水平进行排序, Doughty 等[24] 提出了一种基于孪生网络架构的动作特征表示方法. 该方法使用时序段网络 (Temporal segment network, TSN) 从每个视频中提取动作特征. 它将每个视频均匀划分成 3 个视频段, 并从

每个视频段中随机抽取帧, 分别采用单帧图像和连续 5 帧的光流信息作为 2S-CNN 的输入, 提取空间和时序特征. 融合空间和时序特征, 2S-CNN 能够同时捕获视频中的静态表观信息和动态运动信息, 从而更好地表示视频中的动作特征. 在文献 [62] 中, Jain 等将孪生网络结构应用于动作序列的相似度计算, 基于学习的相似度网络度量模板动作序列和测试动作序列之间的距离. 具体地, 该网络接收成对动作序列样本作为输入, 将其转换为一个二分类问题进行求解, 最终预测输入的动作序列是否相似. 由于原训练集没有成对样本的标注信息, 论文中依据动作样本的真实得分构造训练样本: 首先预定义阈值 θ, 如两个输入样本的得分差值小于 θ, 则认为这一对样本是相似的, 将其标记为正类; 反之将其标记为负类. 此外, 通过和专家动作序列进行相似对计算, 该方法能够为测试动作提供相应的反馈信息.

在动作质量评价任务中, 常见的解决方案是将此问题视为回归任务, 将输入视频映射到裁判提供的最终分数. 然而, 这忽视了视频之间微妙且关键的差异. 为了解决这个问题, Yu 等[63] 将动作质量评分问题重新定义为两个视频中包含的动作对之间相对得分的回归问题. 这里的相对分数指的是相对于另一个具有共同属性 (例如类别和难度) 的动作视频来进行评分. 他们采用了孪生网络结构, 使用 I3D 模型分别从评价视频和示例视频中提取动作特征, 并设计了一个组感知回归树来将传统的分数回归转换为从粗分类到细粒度的小区间回归问题.

基于孪生网络结构, Li 等[64] 提出了一种新的成对对比学习网络, 用于对视频对中的动作之间的差异进行建模. 在网络训练过程中, 他们同时采用基本回归网络与成对对比学习网络, 并通过定义一种新的一致性约束, 使得基本回归网络对两个动作视频的预测评分结果的差异尽可能接近成对对比学习网络预测的相对差异.

3) 基于时序分割的动作特征表示方法

一个复杂的动作往往是由多个子动作构成, 每个子动作的表现都可能对最终的动作质量产生影响. 基于子动作分割的特征表示学习方法依据动作的执行阶段将视频中的动作划分成若干个更小的、有意义的子动作单元, 并通过学习每个子动作的特征表示来表示整个动作视频. 这种方法通过更细粒度的划分和建模, 充分利用动作的局部细节和时序信息, 从而更好地表示和理解复杂动作.

具体而言, 基于时序分割的动作特征表示学习方法通常包含四个阶段: 子动作定义、时序分割、子动作特征表示、整体动作特征表示. 在子动作定义阶段, 整个动作被分解为一系列有意义的子动作单元, 这需要依赖领域专业知识来确定. 子动作是整个动作的基本组成单元, 通常表现为动作的独立运动片段, 例如跳水动作中的准备片段动作、跳跃片段动作等. 在时序分割阶段, 整个动作序列被划分为多个时序连续的子阶段, 每个子阶段对应一个子动作. 时序分割可以通过手工标注、时序分割算法来实现, 这一步骤应确保每个子动作单元在时间上是连续的. 一旦子动作被定义并通过时序分割确定, 接下来需要对每个子动作进行特征表示. 这可能涉及提取子动作中的关键帧、关键点、运动轨迹等特征提取, 以便更好地捕捉子动作的表现. 最后, 从所有子动作的特征中构建整体动作的表示. 通过对各个子动作特征的聚合、组合或序列建模, 以获得对整体动作的全面特征表示.

针对视频中的跳水动作, Xiang 等[59] 和 Dong 等[65] 采用一个两阶段的动作质量评价模型. 在第一阶段, 他们使用 Encoder-decoder temporal convolutional network (ED-TCN)[66] 将整个跳水动作划分为开始、跳跃、下降、入水和结束 5 个连续子动作. 在第二阶段, 他们以每个子动作的中间帧作为关键帧向外扩展为 16 帧, 每个子动作采用一个 16 帧的序列数据表示, 并使用 P3D 网络从连续 16 帧中提取子动作特征. 不同的是, Xiang 等将 5 个子动作特征直接拼接, 并采用全连接层、线性回归层、支持向量回归层对动作质量的评分结果进行预测. 为了获取更优的动作质量评价结果, Dong 等设计了一个由两分支构成的评价模型: 一分支通过拼接子动作的特征形成最终的特征表示, 并设计一个四层神经元逐层递减的回归模型预测评分结果; 另一分支将每个子动作作为单独的特征进行回归, 分别对总分、执行分和难度系数得分进行预测. 损失函数由两个分支的预测损失共同组成, 对所有参数进行联合训练. 与先前的研究不同, 前者通常对整个动作进行整体评分, Liu 等[67] 提出了一种多阶段的动作质量评价方法. 他们首先对动作进行子动作划分, 然后对每个子动作进行独立的评分预测. 最后, 通过加权求和这些子动作的分数, 得到最终的整体得分. 为了更好地捕获动作质量评价任务中同类动作样本之间的细微差异, Gedamu 等[68] 提出一种细粒度的时空解析网络. 该方法一方面采用无监督的对比学习损失, 以挖掘子动作的高级语义表示; 另一方面设计了多尺度时空变换器模块, 用于建模子动作之间的长程依赖关系.

为了对长视频中的动作质量进行评价, Li 等[48] 设计了一个包含关键片段分割 (Key fragment seg-

mentation, KFS) 和得分预测 (Score prediction, SP) 两个部分的评价网络. 该网络先用 3D 卷积网络和双向 LSTM 搭建动作分割子网络, 通过语义分割的方法, 从视频中挑选出最能反映动作质量的关键片段, 去除冗余的视频片段信息. 接着, 该网络再用 3D 卷积网络从关键片段中提取特征, 并根据这些特征进行得分预测, 以达到评价动作质量的目的. 在对花样滑冰的运动员动作进行评价时, Ji 等[44] 将评分划分为节目内容分和技术动作分两个部分进行独立评估. 在技术动作评分部分, 他们采用了基于弱监督的时序子动作定位方法, 确定技术子动作的开始时间和结束时间, 并分别对每个技术子动作进行评估, 以获取最终的积累技术得分.

4) 基于注意力机制的动作特征表示方法

动作视频是由二维空间序列随时间变化而形成的, 每个时间步对应一个空间位置上的图像帧. 空间位置可以反映人体的不同部分或场景中的不同元素, 而时间位置可以反映动作的不同阶段. 因此, 空间位置和时间位置所包含的信息是丰富而多样的, 对动作特征的表达有着不同的影响.

基于注意力机制的动作特征表示方法采用注意力机制来强调或抑制输入序列中不同时空位置的信息, 从而更有效地捕获关键动作特征. 在动作质量评价任务中, 注意力机制的使用又大体分成三种不同的方式: 第一种方式是将注意力机制与其他神经模型结合, 用于对输入数据中的不同部分进行加权处理, 从而让模型更关注最相关和重要的信息, 同时减少不相关的信息对模型的影响. 第二种方式是利用强大的注意力机制对各种不同类型的特征进行有效融合. 它可以自适应地学习并挖掘特征之间的深层相关性和重要性, 提高特征的表达能力. 最后一种方式是利用自注意力机制作为主要的特征提取模型, 用于直接从输入数据中提取关键动作特征.

注意力机制通常与其他神经网络模型结合使用, 以提高模型的表现. 为有效地学习动作视频中的时序特征, Xu 等[33] 设计了一个由自注意力 LSTM 和多尺度卷积跳跃 LSTM 两个互补组件构成的深度架构. 自注意力 LSTM 通过注意力模块对视频特征进行加权, 然后用 LSTM 对加权特征进行建模, 从而突出重要的视频片段, 抑制不重要的视频片段. 多尺度卷积跳跃 LSTM 通过多尺度卷积核提取多层次的特征, 并利用跳跃连接来捕捉时序数据中的长期依赖关系. 基于孪生网络框架, Doughty 等[32] 利用可学习的时序注意力模块评估长视频中动作的相对技能水平. 他们设计出两类时序注意力来关注视频中与技能评价有关的部分, 两类注意力分别关注表示出较高 (优点) 和较低 (缺点) 技能水平的视频部分. Lei 等[69] 提出一种端到端的时间注意力学习方法来改进动作质量评价的性能. 他们构建了一个注意力学习模块来模拟人类对动作质量评价的注意机制和判断偏好, 依据分段预测误差的损失学习不同阶段的权重, 进而平衡视频中不同分段特征的重要性. 人类在评价视频中动作质量时会使用注意力机制, 仅关注视频中的一小部分区域, 而忽略其他无关区域. 为更好地聚焦于关键重要的视频区域, Li 等[29] 考虑累积注意力状态以及任务相关的高层信息提出一种基于循环神经网络的空间注意力模型, 用于从冗余的背景中捕获视频序列中对评价技能水平有关键作用的区域.

在多个特征来源或多个模态的数据融合时, 注意力机制根据输入数据的重要性或相关性, 将不同来源的信息加权结合, 提高模型的表示能力. 为了捕获运动过程中不同主体之间以及主体与环境之间的交互关系, Gao 等[34] 提出了一个非对称交互模块 (AIM), 该模块能够显式地建模动作中存在的非对称交互, 并利用注意力机制融合 I3D 提取的整体场景特征和 AIM 特征. Zeng 等[35] 提出一种混合动态–静态上下文感知注意力模型, 该模型不仅能够从动作视频中学习动态特征, 还能够捕捉运动员在特定帧中的静态姿势. 此外, 他们设计了一个上下文感知注意力模块, 用于学习视频段 (或视频帧) 之间的关系, 并通过聚合所有段/帧生成更加有效的动态–静态特征.

注意力机制本身可作为特征提取器, 有助于从输入的动作数据中学习关键特征. 相关研究已广泛探讨采用自注意力机制[38, 68] 或交叉注意力机制[44, 70] 在动作质量评价任务中的应用. 基于注意力的模型引入动态权重分配机制, 更有效地关注动作视频中关键的时空信息, 从而提升对动作特征的建模能力. 其中, 基于自注意力机制的动作时序数据处理模型通过将每个动作片段与其他动作片段的数据相互关联, 使得模型能够动态关注动作序列不同位置之间的关系, 从而更灵活地捕捉长动作序列中的重要动作片段, 并建模时序关系. 而交叉注意力机制则允许模型在处理多个相关的动作序列时, 动态关注不同序列之间的关系, 更好地建模整体动作的特征.

Wang 等[38] 提出一种基于稀疏特征交互的管道自注意力网络 (TSA-Net), 用于有效地生成丰富的时空上下文信息. 研究中, 他们首先将动作特征的提取分为两个分支: 一支利用视频跟踪算法逐帧提取出与运动对象相关的管道特征; 另一支采用 I3D 进行视频片段的特征提取. 通过融合这两个分支的

输出, 得到稀疏的时空管道特征. 随后, 利用自注意力机制对时空管道特征进行特征提取, 以保留时间维度上的重要上下文信息, 并减弱冗余的空间信息影响. Bai 等[70] 基于 Transformer 的解码器结构提出“时序解析变换器”, 定义一组可学习的查询表示特定动作的原子动作模式, 利用交叉注意力机制将整个视频特征解码为采用一系列原子动作进行表示. 此外, 他们利用组感知回归树预测输入样本和示例样本之间的相对得分, 进而学习更加有效的细节特征. 受心理学中李克特量表启发, Xu 等[71] 提出等级解耦李克特变换器. 该方法首先利用 I3D 从均匀划分的视频片段中提取特征; 再维护一组可学习的查询向量, 同样利用 Transformer 的解码器结构学习等级感知的特征表示; 最后, 基于等级感知的特征估计不同等级的响应强度, 并将其与等级对应的定量值进行线性组合预测最终的模型得分. 为了更好地学习同类动作不同样本之间的类内细微差异, Gedamu 等[68] 提出一种基于子动作的特征表示方法, 并利用时空多尺度的 Transformer 模块学习不同子动作之间的时序依赖关系. Ji 等[44] 通过为花样滑冰中的节目内容和执行技术分别定义可学习的原型表示, 利用交叉注意力机制建模动作特征与可学习原型之间的关联关系, 从而构建了动作特征表示模型. Xu 等[41] 和 Liu 等[43] 选择示例动作, 并利用交叉注意力机制来捕捉评价动作和示例动作之间的时空关联关系, 以学习关于评价动作的有效特征表示. 与以往仅关注将视觉信息作为动作质量评价任务输入的研究不同, Du 等[72] 在他们的工作中同时考虑了文本评论和动作视频两种互补的模态信息. 他们提出了一种基于注意力机制的教师–学生网络, 实现了从语义领域到视觉领域的知识传递. 该方法首先通过交叉注意力机制在语义评论和动作视频之间聚合语义描述和视觉特征, 形成语义感知的动作特征表示. 接着, 将语义感知表示作为教师信号, 为视觉特征的学习提供监督信号. 在学生分支中, 他们使用一组可学习的原子查询建模动作视频中的关键原子动作, 并进一步学习视觉模态中的动作特征表示. 最后, 通过约束不同分支之间的注意力分布以及特征输出差异来优化整个模型.

2.2　基于骨架序列的动作表示学习

人体可以看成由骨骼关节点连接的刚体构成的复杂系统, 人的行为则是刚体的空间位置随时间的演变过程. 如图 3 所示, 骨架序列能够进一步简化为由点和边所构成的图, 点对应骨架中的关节点, 边对应骨架中的骨骼信息, 则人体运动过程可以采用骨架序列 (骨架关节点坐标的时间序列) 进行表示. 骨架序列能够同时包含空间坐标信息和时域信息, 且相比 RGB 图像序列, 具有更少的冗余信息和较低的计算消耗. 因此, 人体骨架模型[73−76] 适用于人体运动行为表示.

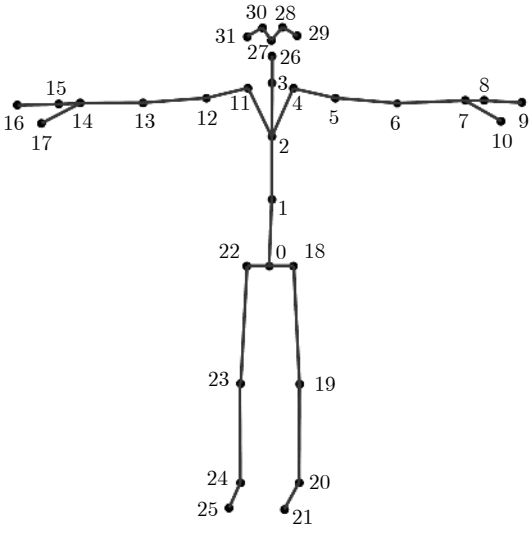


图 3　人体骨架示意图

Fig. 3　The schematic diagram of human skeleton

基于骨架序列的动作表示学习通常包括三个步骤: 1) 提取原始骨架序列特征; 2) 对获取的骨架序列进行预处理; 3) 基于骨架序列的动作特征表示学习.

人体骨架序列提取的方法可以分为 2D 骨架序列提取和 3D 骨架序列提取, 具体取决于提取的骨架序列是在二维图像空间还是三维空间中. 2D 人体骨架序列提取方法又包括传统的人体姿态估计方法和基于深度学习的姿态估计方法. 传统的人体姿态估计方法充分利用人体解剖学结构和运动学信息, 通过图模型和形变部件模型进行建模, 以推断人体的姿态. 典型的方法如 FPM (Flexible parts model)[8, 77]、DPM (Deformable parts model)[78−79] 等; 基于深度学习的人体姿态估计方法充分发挥深度学习模型卓越的表示学习能力, 使得模型能够在原始图像中提取更为丰富的语义信息. 通过设计回归网络, 这类方法将提取的特征映射到人体关节点的坐标空间, 实现了对人体姿态的更精准估计. 典型的方法如 OpenPose[50, 80]、AlphaPose[49, 81] 以及 Mask RCNN[49, 82] 等. 3D 人体骨架序列提取通常需要利用深度传感器 (例如 Microsoft Kinect) 或者多个视角的图像进行三维重建. 深度传感器能够提供图像中每个像素点的深度信息, 因此通常用于实时捕捉人体的三维姿态[76, 83]. 另外, 通过使用多个视觉

摄像头从不同角度捕捉人体的动作数据, 可以借助多视角重建的技术进一步获取人体的三维姿态估计信息[84−85].

受光照和背景等因素影响, 从原始数据中提取的骨架序列往往包含一定的噪声信息, 无法直接用于后续任务中. 为了获得更加鲁棒的特征表示, 噪声滤波、正则化和对齐等预处理过程经常被用来减少类内差异. 假设已经提取了骨架序列, 下面进一步就如何基于骨架序列学习动作特征表示进行介绍, 主要分为手工特征和深度特征.

2.2.1 基于骨架序列的手工特征

基于骨架序列的手工特征依据先验知识从骨架序列中提取能够表征人体运动的特征描述子, 该类方法主要致力于挖掘关键骨架点位置、骨架点之间相对位置和骨架点组合的动态信息. 其基本思想为: 对骨架关节点坐标进行组合或变换, 获取局部的人体姿态特征后, 再基于局部姿态特征进一步提取全局运动特征向量. Ilg 等[25] 于 2003 年提出一种分层动作特征表示学习方法, 并将其应用于 Kata 运动评价问题. 该方法假设人体动作序列由若干个原子动作组成, 并采用运动轨迹在零速度处的特征作为原子动作描述符. 对于任意的人体动作序列, 其首先识别动作序列中的原子动作, 其次采用时空变形模型 (Spatio-temporal morphable models, STMMs) 学习原子动作模板, 最后整个运动过程被表征为原子动作模板的线性组合. Celiktutan 等[86] 于 2013 年提出一种基于图结构的运动分析方法, 并将其应用于体育锻炼运动分析中. 该方法首先计算关节点相对躯干的倾斜角、方位角及关节点间的相对距离等局部姿态特征, 其次组合局部姿态特征获得人体的运动特征表示, 最后基于图匹配算法对不同运动序列进行对比, 进而评价人体动作质量. Pirsiavash 等[8] 于 2014 年提出了基于离散余弦变换 (DCT) 的运动特征表示方法, 并将其应用于体育赛事评分中. 该方法首先计算人体各关节点相对头部的坐标信息, 其次采用 DCT 对各个关节点坐标分别进行滤波, 最后保留低频分量作为动作特征表示. Lei 等[50] 于 2020 年提出了基于关节点位置关系的自相似运动特征表示, 并将其应用于体育赛事评分中. 该方法依次计算关节点在时间序列上的变化和不同关节点对之间的相对距离来表示运动轨迹特征和运动位移特征, 其次计算各个关节点的自相似矩阵, 最后提取自相似矩阵的 HOG 特征作为人体运动特征表示.

2.2.2 基于骨架序列的深度特征

近年来, 随着 GPU 计算能力的提升, 涌现出了大量基于骨架序列的深度特征表示方法. 这类方法利用卷积神经网络[87−88]、循环神经网络[89−90] 和时空图卷积神经网络 (Spatial-temporal graph convolutional network, ST-GCN)[91−92] 等神经网络结构从人体骨架序列中提取有效的运动特征表示, 从而用于描述和理解人体运动的时空变化. 具体而言, 卷积神经网络能够从骨架序列中提取人体关节的位置和运动等空间特征, 从而捕捉动作的空间结构. 循环神经网络主要用于建模动作序列的时序信息. 一旦通过卷积神经网络获得骨架节点的空间表示, 循环神经网络模型可以进一步处理序列中的时序信息, 捕捉骨架节点之间的时序依赖关系. 时空图卷积神经网络将人体骨架数据看作由多个骨架关节点相互连接而成的空间图结构. 通过在时空相关的节点之间进行反复的消息传递和消息融合, 探索基于人体骨架序列数据构造图结构, 进而提取具有辨别力的空间和时间特征. 这些方法在动作质量评价任务中发挥了重要作用, 使得人体动作的时空特征表达更加丰富和准确.

在文献 [93] 中, Yu 等采用时空图卷积神经网络结构从骨架序列中提取动作特征, 并将其应用于康复训练的动作质量评价任务. ST-GCN 利用人体关节点之间的自然连接关系构造空间边, 在不同帧的相同关节点间形成时间边, 基于上述图结构邻域划分策略和分层表示思想构建多层时空图卷积操作, 并沿着时间和空间两个维度不断进行特征聚合. 为进一步捕捉骨架节点之间的时序依赖关系, Chowdhury 等[94] 和 Deb 等[95] 在康复训练的动作质量评价任务中先利用 ST-GCN 从人体骨架序列中提取空间特征和时空特征, 再利用循环神经网络对 ST-GCN 提取的特征序列进行时序建模, 从而增强对动作序列时序信息的建模能力. 为了评估长时间体育视频中的复杂动作, Li 等[96] 提出了一种基于骨架的深度姿态特征学习方法, 并将其应用于花样滑冰中. 该方法先基于 OpenPose 姿态估计方法从原视频中提取骨架特征, 再类似于之前的工作, 结合 ST-GCN 和 LSTM 网络对人体骨架序列的动态时间结构进行建模, 进而捕获人体运动的细微变化. 尽管 ST-GCN 能够基于骨架序列数据较好地学习动作特征表示, Pan 等[49, 97] 认为该网络结构仅沿着时间轴连接相同关节点, 这种构图方式难以刻画动作的短时特性、流畅程度和熟悉程度等细节特征, 因此不足以用于细粒度的动作质量评价任务中的动作表征学习. 他们采用一种新的时空图卷积神经网络学习基于骨架序列的人体动作特征表示. 该方法基于相邻关节点之间的连接定义了可学习的空间关系

图 (表示某一时刻的关节点连接关系) 和时间关系图 (表示两个连续时刻的关节点连接关系), 并基于上述图结构建立关节共性模块和关节差异模块, 其中, 共性模块和差异模块分别通过聚合空间图中不同关节点的运动特征和分析时空图中相邻关节点的运动差异提取运动特征. 此外, 利用可训练的关系图模型有利于分析相邻关节点间是如何相互影响的. 同时考虑运动中的外观信息和姿态特征, Nekoui 等[98] 基于双流网络结构提出一种新的动作特征学习方法. 在这一方法中, 双流网络的一支专门负责利用姿态信息来评估关节和身体部位之间的协调性; 而另一支则致力于捕捉动作外观变化的信息, 并通过引入注意力机制更加有效地关注动作中的时序依赖关系.

2.3 小结

本节对动作质量评价任务中出现的不同动作特征表示学习方法进行总结. 根据采用的数据模态不同, 大致将这些方法划分为基于 RGB 信息的动作表示学习方法和基于骨架序列的动作表示学习方法. 这两类方法分别从动作视频和骨架序列数据中提取具有时空信息的动作特征表示. 前者能够更直接地捕捉视觉运动和外观信息, 但相对计算量较大; 而后者只捕捉人体结构和运动信息, 因此更注重运动的本质, 能够更好地表示人体姿态变化且冗余噪声信息小. 按照特征学习方法不同, 每类动作特征表示学习又可以进一步划分为手工特征表示和深度特征表示.

在基于 RGB 信息的动作表示学习方法中, 早期的研究工作根据动作数据本身的特性手动设计静态特征和时域特征, 分别用于捕捉动作中的空间结构和时序变化特征. 随着深度学习的快速发展, 基于卷积神经网络的动作特征表示方法开始复用已有的卷积神经网络结构从视频中自动提取人体动作特征. 然而, 该类方法存在一些缺点, 包括对 RGB 视频进行处理导致计算复杂度高, 易受拍摄环境、光照和行为人衣着纹理等与行为无关的外界因素影响, 以及在人体尺寸变化、速度变化以及动作差异微小时存在鲁棒性差、精度低等问题. 为获取更加有效的动作特征表示, 研究者们采用了各类不同的解决方案对该类卷积神经网络结构进行改进, 常见的方法包括基于孪生网络的动作特征表示方法, 基于时序分割的动作特征表示方法和基于注意力机制的动作特征表示方法. 基于孪生网络的动作特征表示方法能够同时接收不同执行者执行同一动作的数据, 通过更好地捕获不同执行者动作之间的细微动作差异学习更具判别性的动作特征表示. 基于时序分割的动作特征表示方法将视频中的完整动作划分成多个子动作, 通过更细粒度的建模充分利用动作的细节和时序信息. 基于注意力机制的动作特征表示方法通过引入动态权重分配机制, 使得模型能够自动地关注对动作表示最重要的部分, 从而提高动作表示的表达能力和鲁棒性.

在基于骨架序列的动作表示学习方法中, 早期的研究工作主要集中在如何根据先验知识, 通过对人体骨架关节点信息进行组合、变换等方式, 从局部和全局角度综合设计动作特征. 近年来, 随着技术的发展, 深度学习方法在基于骨架序列的动作表示学习中取得了显著进展. 通过深度学习方法设计神经网络结构, 使其能够自动学习骨架序列中的关键特征, 避免了手动设计特征的繁琐过程, 提高了对复杂动作模式的建模能力. 由于人体骨架可以被看作是天然的图结构, 因此 GCN 在基于骨架的深度学习方法中得到了广泛的应用. 将人体骨架数据视为由多个骨架关节点相互连接而成的空间图结构, ST-GCN 通过在时空相关的节点之间进行反复的消息传递和消息融合, 进而提取具有辨别力的时空运动特征. 然而, 直接利用 ST-GCN 进行骨架特征提取也存在一些缺点, 包括在处理长序列的消息传递时容易发生信息丢失, 缺乏对动作质量评价任务中细节特征的针对性考虑, 以及由于运动过程中的自遮挡、互遮挡导致的骨架序列可能存在的关键信息丢失问题. 针对以上问题, 研究者们提出了不同的解决方法对 ST-GCN 进行改进, 常见的方法包括三种: 利用 LSTM 增强 ST-GCN 特征的时序建模能力; 设计改进的图卷积神经网络; 构建双流网络增加 RGB 视频特征以提供补充的动作外观变换信息.

表 3 总结了上述两类动作表示学习方法的优缺点. 针对目前广泛采用的深度学习方法, 本文在表 4 和表 5 中进一步对其优缺点进行了详细分析.

3 动作质量评价

动作特征表示方法旨在从原始数据中提取出对于动作质量评价任务有用的信息, 进而将原始的输入视频或骨架序列数据转换为更为紧凑、更有意义的特征向量或特征集合表示方式. 为了获得最终的评分结果、评级结果或排序结果, 通常还需要设计动作质量评价模型, 以进一步将动作特征表示与相应的评价目标相关联. 根据评价目标的不同, 本文大体上可以将动作质量评价划分为基于评分预测的方法、基于评级预测的方法和基于排序预测的方法.

表 3 两类动作特征表示方法优缺点对比

Table 3 Advantage and disadvantage comparison for two types of motion feature methods

方法分类	优点	缺点
基于 RGB 信息的动作表示学习[11, 29, 47]	数据易获取, 包含关于动作的丰富视觉信息, 对环境要求较低, 适用性广	数据量高, 存储和处理成本高, 易受光照、复杂背景等无关环境因素影响
基于骨架序列的动作表示学习[48–50]	冗余数据少、计算开销小, 对外部干扰的抗性较强	对骨架序列的准确度要求高, 无法捕捉运动者与环境的交互信息

表 4 基于 RGB 信息的深度动作特征方法优缺点对比

Table 4 Advantage and disadvantage comparison for RGB-based deep motion feature methods

方法分类	优点	缺点
基于卷积神经网络的动作特征表示方法[12, 24, 28, 30–33, 48, 54, 59]	简单易实现	无法充分捕捉动作特征的复杂性
基于孪生网络的动作特征表示方法[24, 62–64]	便于建模动作之间的细微差异	计算复杂度较高, 需要构建有效的样本对
基于时序分割的动作特征表示方法[44, 48, 59, 65–68]	降低噪声干扰, 更好地捕获动作的细节和变化	额外的分割标注信息, 片段划分不准确对性能影响较大
基于注意力机制的动作特征表示方法[29, 32–35, 38, 41, 43–44, 68–72]	自适应性好, 对重要特征的捕获能力强, 可解释性较好	计算复杂度高、内存消耗大

表 5 基于骨架序列的深度动作特征方法优缺点对比

Table 5 Advantage and disadvantage comparison for skeleton-based deep motion feature methods

方法分类	优点	缺点
ST-GCN[93]	模型结构简单, 易实现	长期依赖关系建模困难, 对细节特征的建模能力有限
ST-GCN + LSTM[94–95]	相比 ST-GCN, 具有更优的时序建模能力	计算复杂度增加, 需要对 LSTM 的超参数精调
改进的时空图卷积神经网络[49, 97]	能够对细节特征进行针对性建模	模型泛化性能不佳
基于多模态的双流网络[98]	具有更加丰富的特征表示, 模型的整体鲁棒性更优	数据获取难度增加, 计算复杂度增加, 需要有效的模态特征融合策略

3.1 基于评分预测的方法

基于评分预测的方法直接从动作特征中预测出一个连续值作为最终评分结果或者相对评分结果, 表示动作的质量. 这类方法将动作质量评价问题转化为回归任务, 利用线性回归、支持向量回归、神经网络回归、回归树等回归器预测连续的评分值, 并基于预测评分和真实得分之间的差异定义损失函数, 适用于那些动作质量具有连续性的评价任务, 例如竞技体育中主观打分项目的评分预测、难度预测等.

在传统的两阶段学习方法中, 回归器通常是一个单独的模型, 负责将第一阶段中提取的动作特征映射到连续的评分空间; 而在端到端的方法中, 回归器通常作为整个网络的最后一层, 与动作特征提取过程联合训练获取更加精确的评分预测结果. 此外, 如何选择或设计损失函数是评分预测的核心问题. 为了获得预测评分, 常用的损失函数包括:

1) ϵ 不敏感损失函数

Pirsiavash 等[8] 于 2014 年首次提出将动作质量评价任务看成回归问题, 并使用支持向量回归 (SVR) 模型求解. 该方法的基本思想是: 寻找一个超平面, 使得所有样本点到超平面的距离应小于最大偏差 ϵ. 为了使数据样本"更容易"线性可分, 通常考虑采用核函数将原空间的输入样本映射到高维特征空间, 最终得到的目标函数如下:

$$L = \frac{1}{2}w^{\mathrm{T}}w + C\sum_{i=1}^{N}\|g_i - (w^{\mathrm{T}}\phi(x_i) + b)\|_{\epsilon} \quad (2)$$

其中, w 和 b 是要学习的超平面参数, C 表示正则化系数, x_i 和 g_i 分别表示第 i 个样本的输入数据和真实得分, N 表示样本点个数, $\phi(\cdot)$ 表示采用的映射函数. 上述损失函数也被称为 ϵ 不敏感损失函数: 即当预测值和实际值之间的差别小于 ϵ 时, 损失值等于 0; 否则损失为正. 由于性能良好且使用简单, 该损失函数也被广泛应用于后续的动作质量评价方法[12, 27, 50] 中.

2) 均方误差损失函数 (Mean square error, MSE)

MSE 是动作质量评价任务中一类最常用的优化损失函数[11, 28, 30, 59, 96, 99], 该损失函数计算网络预测

值和真实值误差的平方和均值, 计算公式如下:

$$L = \frac{1}{2N}\sum_{i=1}^{N}(s_i - g_i)^2 \tag{3}$$

其中, s_i 和 g_i 分别为网络预测得分和专家真实打分.

竞技比赛中动作质量的评分往往受到动作类型、难度系数以及运动表现多方面因素影响. 在已知执行者的动作类型、难度系数时, 为了更加关注运动者执行的动作细节, Yu 等[63] 提出一种基于相对分数的回归方法. 将具有共同属性的不同运动者之间相对分数划分为 R 个组, 该方法为学习组区间的相对分数定义回归损失函数:

$$L = \sum_{r=0}^{R}\mathbb{1}(l_r = 1)(\hat{\sigma}_r - \sigma_r)^2 \tag{4}$$

其中, $\mathbb{1}(\cdot)$ 表示指示函数, 而 $l_r = 1$ 表示相对分数位于第 r 个组的区间范围内. 当 $l_r = 1$ 成立时, $\mathbb{1}(l_r = 1)$ 的取值为 1; 否则为 0. $\hat{\sigma}_r$ 和 σ_r 分别是第 r 个组的预测相对分数和真实相对分数. 回归相对分数的思想也被应用在后续研究[41, 70] 中.

3) 基于概率的损失函数

考虑比赛打分过程具有一定的随机性, Tang 等[12] 定义了不确定感知评分分布学习损失 (Uncertainty-aware score distribution learning, USDL). 该方法提出使用概率分布函数表示动作评分的预测结果, 并采用 KL 散度度量预测分布和真实分布之间的差异程度. 目标函数如下:

$$\mathrm{KL}\{p_c||s_{pre}\} = \sum_{i=1}^{m} p(c_i)\ln\frac{p(c_i)}{s_{pre}(c_i)} \tag{5}$$

其中, p_c 和 s_{pre} 分别表示真实分布和预测分布, 前者由真实分布离散化得到, 后者则直接从网络中预测. 此外, 该工作还模拟多专家打分系统, 进一步提出多路径的不确定性得分分布学习方法. USDL 的优秀预测性能也在后续动作质量评价研究工作[38, 100] 中得到了广泛验证.

将运动员的得分看成符合正态分布的随机变量, Li 等[101] 定义高斯损失函数计算预测得分和真实得分之间的误差:

$$L = \frac{1}{N}\sum_{n=1}^{N}\left(1 - \mathrm{e}^{-\frac{(S_n-\mu_n)^2}{2\sigma^2}}\right) \tag{6}$$

其中, N 为样本个数, S_n 和 μ_n 分别为第 n 个样本的预测得分和真实得分. σ 是根据经验设置的超参数.

4) 排序损失函数

为了学习更具判别能力的动作特征表示, Li 等[11] 提出在目标函数中进一步引入排序损失, 总体目标函数如下式:

$$L = L_1 + \alpha L_2 + \beta\|w\|^2 \tag{7}$$

其中, L_1 和 L_2 分别表示 MSE 损失函数和排序损失函数, 前者使得同一动作的预测得分与真实分数尽可能相近 (详见式 (3)), 后者确保不同动作的预测得分与真实分数具有一致的排序结果. α 和 β 是预定义的均衡参数, w 是网络要学习的参数. 具体地, 排序损失又可以进一步写成:

$$L_2 = \sum_{i=1}^{N}\sum_{j=1,\ j>i}^{N}\mathrm{ReLU}(-(s_j - s_i)\mathrm{sign}(g_j - g_i) + \theta) \tag{8}$$

其中, $\mathrm{ReLU}(\cdot)$ 是神经元的非线性激活函数, 计算公式为: $\mathrm{ReLU}(x) = \max(x,\ 0)$. $\mathrm{sign}(\cdot)$ 是符号函数. θ 是依据经验提前预定义的间隔参数.

基于评分预测的方法将动作质量评价任务转化为回归问题进行建模求解. 在这种方法中, 模型的输入和输出分别为动作数据和预测得分, 而真实的专家打分被用作实际的观测值. 为了评估不同模型的预测性能, 现有的研究通常计算测试样本集上真实值与模型预测值之间的斯皮尔曼等级相关系数 (Spearman′s rank correlation, SRC). SRC 值用于评估模型的预测评分与真实评分之间的一致性: 较大的 SRC 值表明预测评分与真实评分之间的一致性较高, 反之则一致性较低.

具体来说, SRC 的计算公式如下:

$$\rho = 1 - \frac{6\sum_{i=0}^{N} h_i^2}{N(N^2-1)} \tag{9}$$

其中 N 表示数据量, h_i 表示对某 i 个样本观察时真实值和预测值的排序位置差. SRC 主要用于确保预测结果的排序准确性, 但并不能保证预测结果与实际得分的一致性.

3.2 基于评级预测的方法

基于评级预测的方法[23, 102−104] 通常将动作质量评价问题转化为分类任务, 其中预测的评级结果是一个具有离散值的等级, 用于表示动作的质量等级. 这种方法适用于那些动作质量具有离散性的评价任务, 例如运动康复中的动作评级、手术技能的等级评估等.

在这类方法中, 可以利用支持向量机、贝叶斯分类器、提升树等传统分类器模型, 或者神经网络中的 softmax 函数, 将动作特征表示转化为概率分布, 从而使得每个等级类别都有一个对应的概率值. 对于训练过程, 常用的损失函数包括 Hinge 损失函

数、Logistic 损失函数、交叉熵损失 (Cross-entropy loss) 等, 用于衡量预测评级和真实等级之间的差异. 这些损失函数在训练过程中帮助优化分类器的参数, 使得预测结果更接近真实等级.

基于评级预测的方法采用通用的分类评估指标衡量模型性能. 举例来说, JIGSAWS 数据集[9] 采用机器人辅助外科手术系统记录动作数据, 根据动作的熟练程度将其划分为专家 (Expert)、中等 (Intermediate) 和新手 (Novice) 三种技能水平; LAM 动作质量数据集[23] 利用 Kinect 采集患者的运动数据, 将运动者的动作质量划分为好 (Good) 和差 (Bad) 两种等级. 为评估基于评级预测的方法性能, 常使用的评测指标包括正确率 (Accuracy)、精确率 (Precision)、召回率 (Recall) 及 F_1 值.

3.3 基于排序预测的方法

基于排序预测的方法专注于对同一任务中多个执行动作的质量水平进行排序, 其核心目标是解决如何对成对样本中出现的同一动作质量高低进行比较. 为实现这一目标, 相关研究工作首先接收包含同一动作的成对视频样本作为输入, 其次采用孪生网络结构从视频中提取动作特征表示, 最后利用输出的特征表示构建排序损失函数, 以实现对成对动作中的相似性和差异性进行建模. 图 4 展示了一个典型的基于排序预测的方法框架.

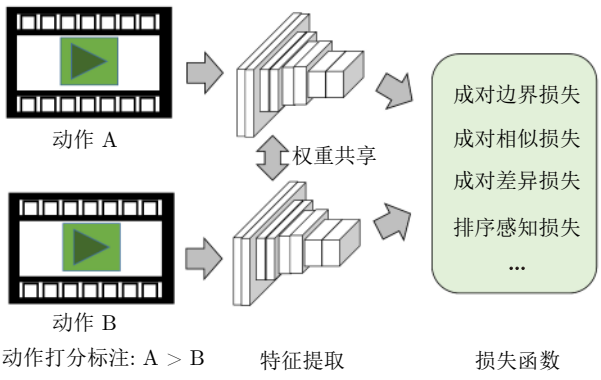


图 4 基于排序预测的方法

Fig. 4 The method based on sorting prediction

其中, 如何从视频中提取动作特征已经在前文中有较详尽的介绍. 接下来, 本文将重点讨论基于排序预测的方法中常用的损失函数. 在文献 [24] 中, Doughty 等提出了成对边界损失和成对相似损失. 成对边界损失函数的设计旨在约束技能水平不同的样本之间的评分预测值尽可能大, 从而使得模型能够更准确地反映样本之间的技能水平差异; 而成对相似损失函数则用于确保对于技能水平相似的样本对, 它们的评分预测之间的差异尽可能小, 从而使得模型能够更好地捕捉样本之间的相似性. 具体而言, 假设有样本对 $(p_i,\ p_j)$, $f(p_i)$ 和 $f(p_j)$ 分别是网络对第 i 个样本和第 j 个样本预测的评分值, 则成对边界损失定义如下:

$$\sum_{(p_i,\ p_j)\in\psi} \max(0,\ m - f(p_i) + f(p_j)) \tag{10}$$

其中, $(p_i,\ p_j) \in \psi$ 表示第 i 个样本的技能水平高于第 j 个样本. 在该损失函数中, 若第 i 个样本与第 j 个样本通过网络预测的评分值大于阈值 m, 则损失为 0; 否则损失为正. 成对相似损失被定义为:

$$\sum_{(p_i,\ p_j)\in\phi}\sum_{k=1}^{N}\max(0,\ |f(p_i) - f(p_j)| - m) \tag{11}$$

其中, $(p_i,\ p_j) \in \phi$ 表示第 i 个样本与第 j 个样本的技能水平相当. 在该损失函数中, 若第 i 个样本与第 j 个样本通过网络预测的评分值之差小于阈值 m, 则损失为 0; 否则损失为正.

引入注意力机制分别聚焦于动作视频中的高技能水平部分和低技能水平部分, Doughty 等[32] 进一步提出“注意力感知的排序损失”方法. 在这项研究中, 他们基于高技能水平注意力表示、低技能水平注意力表示以及平均加权表示这三个分支对视频中出现的动作技能进行评分预测. 通过优化边界损失约束函数, 他们确保在不同技能水平上, 不同分支预测的评分值之间的差异尽可能大, 以便更好地区分不同技能水平下的样本. 此外, 他们还在研究中引入成对差异损失和排序感知损失. 这两类损失函数主要用于优化与技能水平相关的注意力分配, 其中前者用于确保两个注意力分支的评分预测结果均优于均匀分支的评分预测结果; 而后者则用于约束学习到的注意力机制具有排名感知能力, 即能够分别关注到视频中与高技能水平和低技能水平相关的部分. 最终, 通过组合以上三类不同的损失函数, 他们完成了模型的整体训练过程.

在性能评估过程中, 基于排序预测的方法通常使用排序准确率作为评价指标. 该指标反映了模型在比较不同样本对时正确预测样本中动作质量顺序的能力, 其定义为正确排序的样本对数量与总样本对数量的比例. 通过比较模型预测的样本对动作之间的质量排序与样本对之间的实际质量顺序, 可以评估模型在动作质量排序方面的性能.

3.4 小结

本节对动作质量评价的相关技术进行总结. 按照评价目标不同, 大体上可以将运动评价划分为基

于评分预测的评价方法, 基于评级预测的评价方法和基于排序预测的评价方法. 基于评分 (评级) 预测的评价方法利用精准的评分 (评级) 标签数据作为监督信号指导模型训练, 因此能够直接对测试动作进行评分 (评级), 预测结果通常是 0 到 100 之间的某个连续值或者若干个离散的类别之一. 这些方法在竞技体育的主观打分、运动康复的康复进程监测以及姿态矫正与运动训练等领域具有广泛的应用潜力. 然而, 这类方法需要收集领域专家对不同动作的评分 (评级) 情况. 为了确保结果的客观准确, 通常需要多位专家同时进行评分 (评级). 与此不同, 基于排序预测的评价方法采用成对动作序列作为训练数据, 通过比较成对动作之间的差异进行动作质量评价. 该方法仅需要给出弱监督的标注信号, 而无需针对每个动作样本提供精确的评分 (评级) 信息. 因此, 该类方法更适用于难以获得准确评分 (评级) 的情况, 同时能够有效地衡量模型在动作质量排序上的表现.

综上所述, 基于评分预测、基于评级预测和基于排序预测的方法各具特点, 在不同领域都具备广泛的应用前景, 为动作质量评价提供了多样化的解决途径.

4　典型方法性能介绍

为了更深入地追踪动作质量评价方法的最新研究进展, 本节在应用最广泛的三个动作质量评价数据集上比较了现有动作质量评价方法的性能, 并对结果进行了分析和讨论.

4.1　AQA-7

表 6 列出了不同动作质量评价方法在 AQA-7 数据集上的评分性能, 评测指标采用斯皮尔曼等级相关系数. 由于蹦床运动在前期的研究工作中没有相关结果汇报, 表 6 中仅包含其他 6 项运动 (对应第 1 列到第 6 列) 的评测结果, 最后一列为平均值.

由表 6 可知: 1) 目前在 AQA-7 数据集上, 主要的动作质量评价方法都采用了深度学习来学习动作特征表示. 在早期, 一些方法直接采用通用的视频特征提取技术 (如 C3D[11, 28] 等). 随后, 研究者们探索了多种改进模型性能的方法, 包括基于孪生网络的动作表示方法[62−64]、基于时序分割的动作表示方法[59] 以及基于注意力机制的动作表示方法[38, 69−70], 这些方法都被用来改善模型的性能. 此外, 基于人体姿态信息的方法[8, 36, 49, 97] 能够分析人体关节的位置、角度以及运动轨迹, 进而捕捉到动作的细微变化, 也被广泛应用在 AQA-7 数据集的动作特征提取中. 2) AQA-7 数据集上的最优性能由 TPT 方法取得, 获得斯皮尔曼等级相关系数为 0.8715. 该方法利用最新的自注意力机制捕捉视频数据中的时空关系, 能够实现更有效的动作特征表示. 3) 在 AQA-

表 6　在体育评分数据集 AQA-7 上的不同方法性能对比

Table 6　Performance comparison of different methods on sports scoring dataset AQA-7

方法	Diving	Gym Vault	Skiing	Snowboard	Sync. 3m	Sync. 10m	AQA-7	传统/深度	发表年份
Pose+DCT+SVR[8]	0.5300	0.1000	—	—	—	—	—	传统	2014
C3D+SVR[28]	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937	深度	2017
C3D+LSTM[28]	0.6047	0.5636	0.4593	0.5029	0.7912	0.6927	0.6165	深度	2017
Li 等[11]	0.8009	0.7028	—	—	—	—	—	深度	2018
S3D[59]	—	0.8600	—	—	—	—	—	深度	2018
All-action C3D+LSTM[30]	0.6177	0.6746	0.4955	0.3648	0.8410	0.7343	0.6478	深度	2019
C3D-AVG-MTL[30]	0.8808	—	—	—	—	—	—	深度	2019
JRG[49]	0.7630	0.7358	0.6006	0.5405	0.9013	0.9254	0.7849	深度	2019
USDL[12]	0.8099	0.7570	0.6538	0.7109	0.9166	0.8878	0.8102	深度	2020
AIM[36]	0.7419	0.7296	0.5890	0.4960	0.9298	0.9043	0.7789	深度	2020
DML[62]	0.6900	0.4400	—	—	—	—	—	深度	2021
CoRe[63]	0.8824	0.7746	0.7115	0.6624	0.9442	0.9078	0.8401	深度	2021
Lei 等[69]	0.8649	0.7858	—	—	—	—	—	深度	2021
EAGLE-EYE[98]	0.8331	0.7411	0.6635	0.6447	0.9143	0.9158	0.8140	深度	2021
TSA-Net[38]	0.8379	0.8004	0.6657	0.6962	0.9493	0.9334	0.8476	深度	2021
Adaptive[97]	0.8306	0.7593	0.7208	0.6940	0.9588	0.9298	0.8500	深度	2022
PCLN[64]	0.8697	0.8759	0.7754	0.5778	0.9629	0.9541	0.8795	深度	2022
TPT[70]	0.8969	0.8043	0.7336	0.6965	0.9456	0.9545	0.8715	深度	2022

7 数据集的评分任务中, 与基于姿态的特征表示方法[49, 97] 相比, 最新的基于视觉表观特征的表达方法[64, 70] 展现出更卓越的性能, 同时将这两种方法融合起来, 有望进一步提取更为强健的特征. 4) 知识迁移对于提升 AQA-7 数据集性能具有明显作用. All-action C3D+LSTM 定义源数据集 (AQA-7) 和目标数据集 (AQA-7 中单个动作组成的数据集), 并利用微调的方法将从源数据集学习的知识 (动作特征表示) 迁移到目标数据集上. C3D-AVG-MTL 采用两个相关学习任务 (动作分类、评论生成) 的有用信息来提升动作质量的评估性能.

4.2 JIGSAWS

为了比较不同方法在 JIGSAWS 数据集上的表现, Gao 等[9] 引入了两种留一交叉验证方法, 分别是 LOSO (Leave-one-supertrial-out) 和 LOUO (Leave-one-user-out). 其中, LOSO 采用五折交叉验证, 即在每个受试者的五次实验数据中留出一次, 以评估不同方法的性能; 而 LOUO 将八位受试者的样本分别储存在八个文件夹中, 每次系统会将一位受试者的数据保留用于评估各种方法的性能. 此外, Doughty 等[24] 提出了四折交叉验证 (4-fold cross validation) 的方法评估不同方法的性能. 该方法将数据集随机划分为四份, 每次留出其中的一份作为测试集, 其他数据用于测试. 表 7 对不同方法在 JIGSAWS 数据集上的表现进行综合比较.

由表 7 可知: 1) 早期的研究工作主要集中在基于动作特征的方法, 因为这些特征具有维度较低、易于处理的特点. 然而, 这些方法往往难以达到令人满意的结果. 近年来, 随着计算机视觉领域的不断发展, 越来越多的研究者开始考虑采用原始的视频数据来评价手术动作的质量. 这种方法能够以非接触的方式收集数据, 并且通过应用深度学习等最新技术, 取得了更为卓越的性能. 2) LOSO、LOUO 和 4-fold cross validation 是目前最常用的交叉验证法, 在相同条件下 4-fold cross validation 的 SRC 值最高; LOSO 次之; 而 LOUO 表现最低. 也就是说, 模型在不同样本上的泛化能力通常优于在不同受试者上的泛化能力. 3) 将手术技能水平的评价视为分类任务, 在早期阶段主要借用准确率、F1 指标等常见的分类度量指标. 然而, 鉴于这一任务更加强调模型在动作等级排序方面的一致性, 近年来, 越来越多的研究开始将 SRC 作为首选的评测指标. 4) 在手术动作技能评级任务中, 手术操作者在执行过程中呈现出的各种细节特征尤为关键. 近期的研究工作逐渐从学习通用的动作特征[59, 99, 102−103, 105] 转向着重关注更具区分性的细粒度特征表达[34, 106−109], 旨在进一步提升模型的性能水平. 为了捕获运动过程中主体与环境之间的交互关系, Gao 等[34, 108] 提出非对称交互模块对手术动作中存在的复杂交互过程进行建模. 考虑手术工具使用、现场事件以及其他的技能指标等多个组成方面, Liu 等[106] 采用一种综合性的框架, 用于建模手术不同方面之间的依赖关系. 为了充分利用不同语义概念之间的差异, Li 等[107] 提出一种创新的视频语义聚合框架. 该框架旨在通过在时空维度上对不同语义部分的特征进行聚合, 从而更好地学习动作特征表示. Anastasiou 等[109] 提出一种基于注意力网络的回归框架. 该框架通过对比学习捕捉测试视频与代表最佳外科操作的参考视频之间的表现差异.

4.3 EPIC-Skills 2018

表 8 列举了不同方法在 EPIC-Skills 2018 数据集上的性能. 由表 8 可知, 为了取得更好的动作特征表示, 目前该数据集上的研究工作都基于深度网络展开. 相关研究工作主要借鉴了计算机视觉中的卷积神经网络、循环神经网络、注意力机制等最新的深度学习技术. Doughty 等[24] 使用时序分割网络 (TSN) 从视频中提取动作特征, 并联合 0-1 排序损失和相似度损失作为目标函数, 最终在四个任务中获取的排序准确率分别为 71.5% 、70.2% 、83.2% 、79.4%. 观察到一段视频中的不同片段对评价结果影响程度不同, Doughty 等[32] 又提出了排序感知的注意力网络 (Rank-aware attention), 利用注意力机制赋予不同视频片段不同权重, 将结果进一步提升到 84.7% 、68.5% 、82.3% 和 86.9%. Li 等[29] 提出基于注意力的时序神经网络从视频中提取时空特征, 并采用成对排序损失函数作为目标函数, 在四个任务上分别获得了 85.5% 、73.1% 、85.3% 和 82.7% 的排序正确率. 值得注意的是, 尽管网络众包的方式能够减少对专家标注的需求, 但在性能评估时往往难以获得针对某个动作优劣的直观结果. 因此, 相比于前两种数据集的广泛研究, 针对 EPIC-Skills 2018 数据集的研究工作相对较为有限. 此外, 将特征提取网络扩展到孪生网络结构, 并相应地修改损失函数为对比损失, 可以将最初用于动作质量评分或评级的方法进一步扩展应用于动作质量排序任务[97].

5 总结与展望

基于视觉的人体动作质量评价是一个涉及体育科学、计算机视觉、机器学习、人工智能等多个学科的研究领域, 正受到广泛的关注和应用. 本文对基

表 7　JIGSAWS 数据集上的不同方法性能对比

Table 7　Performance comparison of different methods on JIGSAWS

方法	数据模态	评价方法	技能水平划分	交叉验证方法	评测指标	SU	KT	NP	发表年份
k-NN[110]	动作特征	GRS	两类	LOSO	Accuracy	0.897	—	0.821	2018
				LOUO	Accuracy	0.719	—	0.729	2018
LR[110]	动作特征	GRS	两类	LOSO	Accuracy	0.899	—	0.823	2018
				LOUO	Accuracy	0.744	—	0.702	2018
SVM[110]	动作特征	GRS	两类	LOSO	Accuracy	0.754	—	0.754	2018
				LOUO	Accuracy	0.798	—	0.779	2018
SMT[111]	动作特征	Self-proclaimed	三类	LOSO	Accuracy	0.990	0.996	0.999	2018
				LOUO	Accuracy	0.353	0.323	0.571	2018
DCT[111]	动作特征	Self-proclaimed	三类	LOSO	Accuracy	1.000	0.997	0.999	2018
				LOUO	Accuracy	0.647	0.548	0.357	2018
DFT[111]	动作特征	Self-proclaimed	三类	LOSO	Accuracy	1.000	0.999	0.999	2018
				LOUO	Accuracy	0.647	0.516	0.464	2018
ApEn[111]	动作特征	Self-proclaimed	三类	LOSO	Accuracy	1.000	0.999	1.000	2018
				LOUO	Accuracy	0.882	0.774	0.857	2018
CNN[102]	动作特征	Self-proclaimed	三类	LOSO	Accuracy	0.934	0.898	0.849	2018
CNN[102]	动作特征	GRS	三类	LOSO	Accuracy	0.925	0.954	0.913	2018
CNN[105]	动作特征	Self-proclaimed	三类	LOSO	Micro F1	1.000	0.921	1.000	2018
					Macro F1	1.000	0.932	1.000	2018
Forestier 等[112]	动作特征	GRS	三类	LOSO	Micro F1	0.897	0.611	0.963	2018
					Macro F1	0.867	0.533	0.958	2018
S3D[59]	视频数据	GRS	三类	LOSO	SRC	0.680	0.640	0.570	2018
				LOUO	SRC	0.030	0.140	0.350	2018
FCN[99]	动作特征	Self-proclaimed	三类	LOSO	Micro F1	1.000	0.921	1.000	2019
					Macro F1	1.000	0.932	1.000	2019
3D ConvNet (RGB)[103]	视频数据	Self-proclaimed	三类	LOSO	Accuracy	1.000	0.958	0.964	2019
3D ConvNet (OF)[103]	视频数据	Self-proclaimed	三类	LOSO	Accuracy	1.000	0.951	1.000	2019
JRG[49]	视频数据	GRS	三类	LOUO	SRC	0.350	0.190	0.670	2019
USDL[12]	视频数据	GRS	三类	4-fold cross validation	SRC	0.710	0.710	0.690	2020
AIM[34]	视频数据 动作特征	GRS	三类	LOUO	SRC	0.450	0.610	0.340	2020
MTL-VF (ResNet)[113]	视频数据	GRS	三类	LOSO	SRC	0.790	0.630	0.730	2020
				LOUO	SRC	0.680	0.720	0.480	2020
MTL-VF (C3D)[113]	视频数据	GRS	三类	LOSO	SRC	0.770	0.890	0.750	2020
				LOUO	SRC	0.690	0.830	0.860	2020
CoRe[63]	视频数据	GRS	三类	4-fold cross validation	SRC	0.840	0.860	0.860	2021
VTPE[106]	视频数据 动作特征	GRS	三类	LOUO	SRC	0.450	0.590	0.650	2021
				4-fold cross validation	SRC	0.830	0.820	0.760	2021
ViSA[107]	视频数据	GRS	三类	LOSO	SRC	0.840	0.920	0.930	2022
				LOUO	SRC	0.720	0.760	0.900	2022
				4-fold cross validation	SRC	0.790	0.840	0.860	2022
Gao 等[108]	视频数据 动作特征	GRS	三类	LOUO	SRC	0.600	0.690	0.660	2023
				4-fold cross validation	SRC	0.830	0.950	0.830	2023
Contra-Sformer[109]	视频数据	GRS	三类	LOSO	SRC	0.860	0.890	0.710	2023
				LOUO	SRC	0.650	0.690	0.710	2023

表 8 在 EPIC-Skills 2018 上的不同方法性能对比

Table 8 Performance comparison of different methods on EPIC-Skills 2018

方法	Chopstick-Using	Surgery	Drawing	Rough-Rolling	发表年份
Siamese TSN with L_{rank3} [24]	71.5%	70.2%	83.2%	79.4%	2018
Rank-aware Attention[32]	84.7%	68.5%	82.3%	86.9%	2019
RNN-based Spatial Attention[29]	85.5%	73.1%	85.3%	82.7%	2019
Adaptive[97]	87.7%	71.9%	88.2%	88.5%	2021

于视觉的动作质量评价方法流程进行了全面阐述, 详细探究了动作数据获取、动作特征表示以及动作质量评价三个步骤所采用的具体方法, 并在 AQA-7、JIGSAWS、EPIC-Skills 2018 三个数据集上详细分析不同方法的性能. 通过前述总结和回顾, 可以看出科研工作者在动作质量评价技术方面已经取得了一些显著成果. 然而, 当前的应用研究仍有不足之处, 尚存在许多值得深入探索和研究的内容. 通过本次综述, 本文也列举未来可能的研究重点和方向, 以期启发该领域学者对动作质量评价的进一步研究.

1) 更丰富的开放数据集

随着深度学习技术的不断发展, 基于数据驱动的方法已经在动作质量评价任务中展现出巨大潜力, 数据集的规模、多样性和代表性对于模型性能至关重要. 近年来, 虽然已经发布了不少公开数据集用于动作质量评价任务的研究, 但相比人类庞大的动作库, 仍需要构建更丰富的公开数据集, 以促进该领域的进一步发展. 此外, 一个包含更多变化的大规模动作质量评价数据集能够更好地模拟现实情况, 为研究人员提供更具挑战性的问题. 然而, 现有的公开动作质量评价数据集仅仅针对有限的应用场景, 且数据规模较小. 因此, 针对特定的动作质量评价应用领域 (如竞技体育或者运动康复领域) 构建一个大规模动作质量评价数据集具有重要意义. 这样的数据集应拥有更多动作类别、涵盖丰富运动场景与环境, 并且包括各类表现水平的运动个体数据, 能够用于测试、比较和改进不同的动作质量评价方法, 从而推动该领域的发展. 在数据集构建过程中, 专家标注的准确性和一致性也至关重要. 通常情况下, 构建这样的动作质量评价数据库需要跨学科的协作, 融合计算机视觉、运动科学等领域的专业知识.

2) 领域专家指导的表示学习

未来的研究应该进一步探索如何更好地利用领域知识, 从动作质量评价任务本身的特点出发, 实现更精细、准确的动作质量评价. 在动作特征表示时, 加强与运动学、运动控制学、运动康复等领域专家合作交流, 将运动相关领域的知识融入到动作特征提取过程中, 以更好地捕捉相关的特征信息. 此外, 考虑让多领域专家同时参与模型的设计和优化目标的定义过程, 以确保构建的模型能够更好地理解与动作质量相关的信息.

3) 实时性与移动应用

随着移动设备的普及和计算能力的提高, 实现在线动作质量评价正变得切实可行. 在线实时系统能够为运动者提供即时反馈, 帮助他们在动作过程中纠正错误、调整姿势, 从而提高运动技能水平. 此外, 在运动过程中及时识别异常或错误的动作能够有效降低运动者受伤的风险, 特别是在高强度运动项目中显得尤为重要. 当前的研究工作多数建立在离线动作数据基础上, 主要对提前收集到的动作数据进行分析和评估. 这些方法通常更注重动作质量评价的准确性, 却忽略了评价的时效性. 因此, 在未来的研究中, 可以考虑致力于开发实时且便携式的动作质量评价系统, 以便及时地为运动者提供反馈和指导, 满足他们在动作执行过程中的质量评价需求.

4) 可解释的动作质量评价

动作质量评价的可解释性[114] 指的是对评价结果和决策的解释能力. 在动作质量评价任务中, 模型的可解释性能够清晰地解释为什么模型对于某个动作给出了特定的质量评价结果, 以及哪些因素影响了该结果的形成. 这种可解释性对于帮助运动者理解模型的决策、提高模型的可信度以及应用领域的可接受性都非常重要. 只有当运动者能够理解模型的评价基础和建议依据时, 他们才能更好地理解如何改善自己的动作、调整姿势或纠正问题. 此外, 这种可解释性有助于建立运动者的信任, 使得他们更愿意接受和遵循模型提供的建议, 从而提高运动技能、降低受伤风险或提高康复效果. 尽管当前各类动作质量评价模型层出不穷, 但多数模型仍然被当做黑盒模型使用, 理解和解释其内部机理仍然面临诸多困难和挑战. 因此, 在未来的工作中, 应更加重视相关模型的解释理论和方法的研究.

References

1 Zhu Yu, Zhao Jiang-Kun, Wang Yi-Ning, Zheng Bing-Bing. A review of human action recognition based on deep learning. *Acta Automatica Sinica*, 2016, **42**(6): 848−857
(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)

2 Lei Q, Du J X, Zhang H B, Ye S, Chen D S. A survey of vision-based human action evaluation methods. *Sensors*, 2019, **19**(19): Article No. 4129

3 Ahad M A R, Antar A D, Shahid O. Vision-based action understanding for assistive healthcare: A short review. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. Long Beach, USA: IEEE, 2019. 1–11

4 Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018, **2018**(1): Article No. 7068349

5 Zheng Tai-Xiong, Huang Shuai, Li Yong-Fu, Feng Ming-Chi. Key techniques for vision based 3D reconstruction: A review. *Acta Automatica Sinica*, 2020, **46**(4): 631–652
(郑太雄, 黄帅, 李永福, 冯明驰. 基于视觉的三维重建关键技术研究综述. 自动化学报, 2020, **46**(4): 631–652)

6 Lin Jing-Dong, Wu Xin-Yi, Chai Yi, Yin Hong-Peng. Structure optimization of convolutional neural networks: A survey. *Acta Automatica Sinica*, 2020, **46**(1): 24–37
(林景栋, 吴欣怡, 柴毅, 尹宏鹏. 卷积神经网络结构优化综述. 自动化学报, 2020, **46**(1): 24–37)

7 Zhang Chong-Sheng, Chen Jie, Li Qi-Long, Deng Bin-Quan, Wang Jie, Chen Cheng-Gong. Deep contrastive learning: A survey. *Acta Automatica Sinica*, 2023, **49**(1): 15–39
(张重生, 陈杰, 李岐龙, 邓斌权, 王杰, 陈承功. 深度对比学习综述. 自动化学报, 2023, **49**(1): 15–39)

8 Pirsiavash H, Vondrick C, Torralba A. Assessing the quality of actions. In: Proceedings of the 13th European Conference on Computer Vision (ECCV 2014). Zurich, Switzerland: Springer, 2014. 556–571

9 Gao Y, Vedula S S, Reiley C E, Ahmidi N, Varadarajan B, Lin H C, et al. JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling. In: Proceedings of the Modeling and Monitoring of Computer Assisted Interventions (M2CAI)-MICCAI Workshop. 2014.

10 Paiement A, Tao L L, Hannuna S, Camplani M, Damen D, Mirmehdi M. Online quality assessment of human movement from skeleton data. In: Proceedings of the British Machine Vision Conference. Nottingham, UK: 2014. 153–166

11 Li Y J, Chai X J, Chen X L. End-to-end learning for action quality assessment. In: Proceedings of the 19th Pacific-Rim Conference on Multimedia, Advances in Multimedia Information Processing (PCM 2018). Hefei, China: Springer, 2018. 125–134

12 Tang Y S, Ni Z L, Zhou J H, Zhang D Y, Lu J W, Wu Y, et al. Uncertainty-aware score distribution learning for action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 9839–9848

13 Xu J L, Yin S B, Zhao G H, Wang Z S, Peng Y X. FineParser: A fine-grained spatio-temporal action parser for human-centric action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 14628–14637

14 Morgulev E, Azar O H, Lidor R. Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, 2018, **5**(4): 213–222

15 Butepage J, Black M J, Kragic D, Kjellström H. Deep representation learning for human motion prediction and classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 1591–1599

16 Tao L L, Paiement A, Damen D, Mirmehdi M, Hannuna S, Camplani M, et al. A comparative study of pose representation and dynamics modelling for online motion quality assessment. *Computer Vision and Image Understanding*, 2016, **148**: 136–152

17 Khalid S, Goldenberg M, Grantcharov T, Taati B, Rudzicz F. Evaluation of deep learning models for identifying surgical actions and measuring performance. *JAMA Network Open*, 2020, **3**(3): Article No. e201664

18 Qiu Y H, Wang J P, Jin Z, Chen H H, Zhang M L, Guo L Q. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. *Biomedical Signal Processing and Control*, 2022, **72**: Article No. 103323

19 Niewiadomski R, Kolykhalova K, Piana S, Alborno P, Volpe G, Camurri A. Analysis of movement quality in full-body physical activities. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2019, **9**(1): Article No. 1

20 Vakanski A, Jun H P, Paul D, Baker R. A data set of human body movements for physical rehabilitation exercises. *Data*, 2018, **3**(1): Article No. 2

21 Alexiadis D S, Kelly P, Daras P, O′Connor N E, Boubekeur T, Moussa M B. Evaluating a dancer′s performance using Kinect-based skeleton tracking. In: Proceedings of the 19th ACM International Conference on Multimedia. Scottsdale, USA: ACM, 2011. 659–662

22 Capecci M, Ceravolo M G, Ferracuti F, Iarlori S, MonteriùA, Romeo L, et al. The KIMORE dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2019, **27**(7): 1436–1448

23 Parmar P, Morris B T. Measuring the quality of exercises. In: Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Orlando, USA: IEEE, 2016. 2241–2244

24 Doughty H, Damen D, Mayol-Cuevas W. Who′s better? Who′s best? Pairwise deep ranking for skill determination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 6057–6066

25 Ilg W, Mezger J, Giese M. Estimation of skill levels in sports based on hierarchical spatio-temporal correspondences. In: Proceedings of the 25th DAGM Symposium on Pattern Recognition. Springer, 2003. 523–531

26 Wnuk K, Soatto S. Analyzing diving: A dataset for judging action quality. In: Proceedings of the Asian 2010 International Workshops on Computer Vision (ACCV 2010 Workshops). Queenstown, New Zealand: Springer, 2010. 266–276

27 Bertasius G, Park H S, Yu S X, Shi J B. Am I a baller? Basketball performance assessment from first-person videos. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2196–2204

28 Parmar P, Morris B T. Learning to score Olympic events. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, USA: IEEE, 2017. 76–84

29 Li Z Q, Huang Y F, Cai M J, Sato Y. Manipulation-skill assessment from videos with spatial attention network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea: IEEE, 2019. 4385–4395

30 Parmar P, Morris B. Action quality assessment across multiple actions. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa Village, USA: IEEE, 2019. 1468–1476

31 Parmar P, Morris B T. What and how well you performed? A multitask learning approach to action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 304–313

32 Doughty H, Mayol-Cuevas W, Damen D. The pros and cons: Rank-aware temporal attention for skill determination in long

videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 7854−7863

33 Xu C M, Fu Y W, Zhang B, Chen Z T, Jiang Y G, Xue X Y. Learning to score figure skating sport videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, **30**(12): 4578−4590

34 Gao J B, Zheng W S, Pan J H, Gao C Y, Wang Y W, Zeng W, et al. An asymmetric modeling for action assessment. In: Proceedings of the 16th European Conference on Computer Vision (ECCV 2020). Glasgow, UK: Springer, 2020. 222−238

35 Zeng L A, Hong F T, Zheng W S, Yu Q Z, Zeng W, Wang Y W, et al. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In: Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM, 2020. 2526−2534

36 Sardari F, Paiement A, Hannuna S, Mirmehdi M. VI-Net——View-invariant quality of human movement assessment. *Sensors*, 2020, **20**(18): Article No. 5258

37 Parmar P, Reddy J, Morris B. Piano skills assessment. In: Proceedings of the 23rd International Workshop on Multimedia Signal Processing (MMSP). Tampere, Finland: IEEE, 2021. 1−5

38 Wang S L, Yang D K, Zhai P, Chen C X, Zhang L H. TSA-Net: Tube self-attention network for action quality assessment. In: Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event: ACM, 2021. 4902−4910

39 Chen X, Pang A Q, Yang W, Ma Y X, Xu L, Yu J Y. SportsCap: Monocular 3D human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 2021, **129**(10): 2846−2864

40 Parmar P, Gharat A, Rhodin H. Domain knowledge-informed self-supervised representations for workout form assessment. In: Proceedings of the 17th European Conference on Computer Vision (ECCV 2022). Tel Aviv, Israel: Springer, 2022. 105−123

41 Xu J L, Rao Y M, Yu X M, Chen G Y, Zhou J, Lu J W. FineDiving: A fine-grained dataset for procedure-aware action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 2939−2948

42 Zhang S Y, Dai W X, Wang S J, Shen X W, Lu J W, Zhou J, et al. LOGO: A long-form video dataset for group action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023. 2405−2414

43 Liu Y C, Cheng X N, Ikenaga T. A figure skating jumping dataset for replay-guided action quality assessment. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM, 2023. 2437−2445

44 Ji Y L, Ye L F, Huang H L, Mao L J, Zhou Y, Gao L L. Localization-assisted uncertainty score disentanglement network for action quality assessment. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM, 2023. 8590−8597

45 Zahan S, Hassan G M, Mian A. Learning sparse temporal video mapping for action quality assessment in floor gymnastics. *IEEE Transactions on Instrumentation and Measurement*, 2024, **73**: Article No. 5020311

46 Ahmidi N, Tao L L, Sefati S, Gao Y X, Lea C, Haro B, et al. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 2017, **64**(9): 2025−2041

47 Liao Y L, Vakanski A, Xian M. A deep learning framework for assessing physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2020, **28**(2): 468−477

48 Li Y J, Chai X J, Chen X L. ScoringNet: Learning key fragment for action quality assessment with ranking loss in skilled sports. In: Proceedings of the 14th Asian Conference on Computer Vision (ACCV 2018). Perth, Australia: Springer, 2018. 149−164

49 Pan J H, Gao J B, Zheng W S. Action assessment by joint relation graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 6330−6339

50 Lei Q, Zhang H B, Du J X, Hsiao T, Chen C C. Learning effective skeletal representations on RGB video for fine-grained human action quality assessment. *Electronics*, 2020, **9**(4): Article No. 568

51 Gordon A S. Automated video assessment of human performance. In: Proceedings of the AI-ED-World Conference on Artificial Intelligence in Education. Washington, USA: AACE Press, 1995. 541−546

52 Venkataraman V, Vlachos I, Turaga P. Dynamical regularity for action analysis. In: Proceedings of the British Machine Vision Conference. Swansea, UK: BMVA Press, 2015. 67−78

53 Zia A, Sharma Y, Bettadapura V, Sarin E L, Ploetz T, Clements M A, et al. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *International Journal of Computer Assisted Radiology and Surgery*, 2016, **11**(9): 1623−1636

54 Parmar P. On Action Quality Assessment [Ph.D. dissertation], University of Nevada, USA, 2019.

55 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge, US: ACM, 2014. 568−576

56 Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 4489−4497

57 Carreira J, Zisserman A. Quo Vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 4724−4733

58 Qiu Z F, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 5534−5542

59 Xiang X, Tian Y, Reiter A, Hager G D, Tran T D. S3D: Stacking segmental P3D for action quality assessment. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). Athens, Greece: IEEE, 2018. 928−932

60 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico: ICLR, 2016. 928−932

61 Bromley J, Bentz J W, Bottou L, Guyon I, Lecun Y, Moor C, et al. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 1993, **7**(4): 669−688

62 Jain H, Harit G, Sharma A. Action quality assessment using siamese network-based deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, **31**(6): 2260−2273

63 Yu X M, Rao Y M, Zhao W L, Lu J W, Zhou J. Group-aware contrastive regression for action quality assessment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 7899−7908

64 Li M Z, Zhang H B, Lei Q, Fan Z W, Liu J H, Du J X. Pairwise contrastive learning network for action quality assessment. In: Proceedings of the 17th European Conference on Computer Vision (ECCV 2022). Tel Aviv, Israel: Springer, 2022. 457−473

65 Dong L J, Zhang H B, Shi Q H Y, Lei Q, Du J X, Gao S C.

Learning and fusing multiple hidden substages for action quality assessment. *Knowledge-Based Systems*, 2021, **229**: Article No. 107388

66 Lea C, Flynn M D, Vidal R, Reiter A, Hager G D. Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 1003−1012

67 Liu L X, Zhai P J, Zheng D L, Fang Y. Multi-stage action quality assessment method. In: Proceedings of the 4th International Conference on Control, Robotics and Intelligent System. Guangzhou, China: ACM, 2023. 116−122

68 Gedamu K, Ji Y L, Yang Y, Shao J, Shen H T. Fine-grained spatio-temporal parsing network for action quality assessment. *IEEE Transactions on Image Processing*, 2023, **32**: 6386−6400

69 Lei Q, Zhang H B, Du J X. Temporal attention learning for action quality assessment in sports video. *Signal, Image and Video Processing*, 2021, **15**(7): 1575−1583

70 Bai Y, Zhou D S, Zhang S Y, Wang J, Ding E R, Guan Y, et al. Action quality assessment with temporal parsing transformer. In: Proceedings of the 17th European Conference on Computer Vision (ECCV 2022). Tel Aviv, Israel: Springer, 2022. 422−438

71 Xu A, Zeng L A, Zheng W S. Likert scoring with grade decoupling for long-term action assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 3222−3231

72 Du Z X, He D, Wang X, Wang Q. Learning semantics-guided representations for scoring figure skating. *IEEE Transactions on Multimedia*, 2024, **26**: 4987−4997

73 Yan S J, Xiong Y J, Lin D H. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. 7444−7452

74 Gao X, Hu W, Tang J X, Liu J Y, Guo Z M. Optimized skeleton-based action recognition via sparsified graph regression. In: Proceedings of the ACM International Conference on Multimedia. Nice, France: ACM, 2019. 601−610

75 Patrona F, Chatzitofis A, Zarpalas D, Daras P. Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognition*, 2018, **76**: 612−622

76 Microsoft Development Team. Azure Kinect body tracking joints [Online], available: https://learn.microsoft.com/en-us/previous-versions/azure/kinect-dk/body-joints, December 12, 2024

77 Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts. In: Proceedings of the CVPR 2011. Colorado Springs, USA: IEEE, 2011. 1385−1392

78 Felzenszwalb P F, Girshick R B, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(9): 1627−1645

79 Tian Y, Sukthankar R, Shah M. Spatiotemporal deformable part models for action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE, 2013. 2642−2649

80 Cao Z, Simon T, Wei S E, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 7291−7299

81 Fang H S, Xie S Q, Tai Y W, Lu C W. RMPE: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2353−2362

82 He K, Gkioxari G, DollÁR P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2980−2988

83 Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, et al. Real-time human pose recognition in parts from single depth images. In: Proceedings of the CVPR 2011. Colorado Springs, USA: IEEE, 2011. 1297−1304

84 Rhodin H, Meyer F, Spörri J, Müller E, Constantin V, Fua P, et al. Learning monocular 3D human pose estimation from multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 8437−8446

85 Dong J T, Jiang W, Huang Q X, Bao H J, Zhou X W. Fast and robust multi-person 3D pose estimation from multiple views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 7784−7793

86 Celiktutan O, Akgul C B, Wolf C, Sankur B. Graph-based analysis of physical exercise actions. In: Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare. Barcelona, Spain: ACM, 2013. 23−32

87 Liu J, Wang G, Hu P, Duan L Y, Kot A C. Global context-aware attention LSTM networks for 3D action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 3671−3680

88 Lee I, Kim D, Kang S, Lee S. Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 1012−1020

89 Li C, Zhong Q Y, Xie D, Pu S L. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press, 2018. 786−792

90 Li Y S, Xia R J, Liu X, Huang Q H. Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Shanghai, China: IEEE, 2019. 1066−1071

91 Li M S, Chen S H, Chen X, Zhang Y, Wang Y F, Tian Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 3590−3598

92 Shi L, Zhang Y F, Cheng J, Lu H Q. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 12018−12027

93 Yu B X B, Liu Y, Chan K C C. Skeleton-based detection of abnormalities in human actions using graph convolutional networks. In: Proceedings of the 2nd International Conference on Transdisciplinary AI (TransAI). Irvine, USA: IEEE, 2020. 131−137

94 Chowdhury S H, Al Amin M, Rahman A K M M, Amin M A, Ali A A. Assessment of rehabilitation exercises from depth sensor data. In: Proceedings of the 24th International Conference on Computer and Information Technology. Dhaka, Bangladesh: IEEE, 2021. 1−7

95 Deb S, Islam M F, Rahman S, Rahman S. Graph convolutional networks for assessment of physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022, **30**: 410−419

96 Li H Y, Lei Q, Zhang H B, Du J X, Gao S C. Skeleton-based deep pose feature learning for action quality assessment on figure skating videos. *Journal of Visual Communication and Image Representation*, 2022, **89**: Article No. 103625

97 Pan J H, Gao J B, Zheng W S. Adaptive action assessment.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, **44**(12): 8779−8795

98 Nekoui M, Cruz F O T, Cheng L. Eagle-eye: Extreme-pose action grader using detail bird′s-eye view. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2021. 394−402

99 Fawaz H I, Forestier G, Weber J, Idoumghar L, Muller P A. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 2019, **14**(9): 1611−1617

100 Roditakis K, Makris A, Argyros A. Towards improved and interpretable action quality assessment with self-supervised alignment. In: Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference. Corfu, Greece: IEEE, 2021. 507−513

101 Li M Z, Zhang H B, Dong L J, Lei Q, Du J X. Gaussian guided frame sequence encoder network for action quality assessment. *Complex & Intelligent Systems*, 2023, **9**(2): 1963−1974

102 Wang Z, Fey A M. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 2018, **13**(12): 1959−1970

103 Funke I, Mees S T, Weitz J, Speidel S. Video-based surgical skill assessment using 3D convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 2019, **14**(7): 1217−1225

104 Wang Z, Fey A M. SATR-DL: Improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks. In: Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Honolulu, USA: IEEE, 2018. 1793−1796

105 Fawaz H I, Forestier G, Weber J, Idoumghar L, Muller P A. Evaluating surgical skills from kinematic data using convolutional neural networks. In: Proceedings of the 21st International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2018). Granada, Spain: Springer, 2018. 214−221

106 Liu D C, Li Q Y, Jiang T T, Wang Y Z, Miao R L, Shan F, et al. Towards unified surgical skill assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 9517−9526

107 Li Z Q, Gu L, Wang W M, Nakamura R, Sato Y. Surgical skill assessment via video semantic aggregation. In: Proceedings of the 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2022). Singapore: Springer, 2022. 410−420

108 Gao J B, Pan J H, Zhang S J, Zheng W S. Automatic modelling for interactive action assessment. *International Journal of Computer Vision*, 2023, **131**(3): 659−679

109 Anastasiou D, Jin Y M, Stoyanov D, Mazomenos E. Keep your eye on the best: Contrastive regression transformer for skill assessment in robotic surgery. *IEEE Robotics and Automation Letters*, 2023, **8**(3): 1755−1762

110 Fard M J, Ameri S, Ellis R D, Chinnam R B, Pandya A K, Klein M D. Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 2018, **14**(1): Article No. e1850

111 Zia A, Essa I. Automated surgical skill assessment in RMIS training. *International Journal of Computer Assisted Radiology and Surgery*, 2018, **13**(5): 731−739

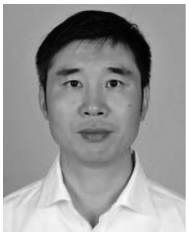
112 Forestier G, Petitjean F, Senin P, Despinoy F, Huaulmé A, Fawaz H I, et al. Surgical motion analysis using discriminative interpretable patterns. *Artificial Intelligence in Medicine*, 2018, **91**: 3−11

113 Wang T Y, Wang Y J, Li M. Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In: Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2020). Lima, Peru: Springer, 2020. 668−678

114 Okamoto L, Parmar P. Hierarchical NeuroSymbolic approach for comprehensive and explainable action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE, 2024. 3204−3213

沈媛媛 北京体育大学体育工程学院讲师. 2020 年获得中国科学院自动化研究所博士学位. 主要研究方向为智能体育与运动表现分析. 本文通信作者.
E-mail: shenyuanyuan@bsu.edu.cn
(**SHEN Yuan-Yuan** Lecturer at School of Sport Engineering, Beijing Sport University. She received her Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2020. Her research interest covers intelligent sports and sports performance analysis. Corresponding author of this paper.)

张燕明 中国科学院自动化研究所副研究员. 2011 年获得中国科学院自动化研究所博士学位. 主要研究方向为结构预测方法, 图神经网络, 概率图模型.
E-mail: ymzhang@nlpr.ia.ac.cn
(**ZHANG Yan-Ming** Associate professor at Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2011. His research interest covers structural prediction methods, graph neural networks and probabilistic graphical models.)



沈燕飞 北京体育大学体育工程学院教授. 2014 年获得中国科学院大学博士学位. 主要研究方向为智能视频分析, 体育大数据, 智能体育装备.
E-mail: syf@bsu.edu.cn
(**SHEN Yan-Fei** Professor at School of Sport Engineering, Beijing Sport University. He received his Ph.D. degree from University of Chinese Academy of Sciences in 2014. His research interest covers intelligent video analysis, sports big data and intelligent sports equipment.)