计算机科学
COMPUTER SCIENCE

视频理解中的动作质量评估方法综述

张洪博，董力嘉，潘玉彪，萧宗志，张惠臻，杜吉祥

引用本文

张洪博，董力嘉，潘玉彪，萧宗志，张惠臻，杜吉祥. 视频理解中的动作质量评估方法综述[J]. 计算机科学, 2022, 49(7): 79-88.

ZHANG Hong-bo, DONG Li-jia, PAN Yu-biao, HSIAO Tsung-chih, ZHANG Hui-zhen, DU Ji-xiang. Survey on Action Quality Assessment Methods in Video Understanding [J]. Computer Science, 2022, 49(7): 79-88.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600028

视频理解中的动作质量评估方法综述

张洪博[1] 董力嘉[1] 潘玉彪[2] 萧宗志[2] 张惠臻[2] 杜吉祥[2,3]
1 华侨大学计算机科学与技术学院 福建 厦门 361000
2 华侨大学福建省大数据智能与安全重点实验室 福建 厦门 361000
3 华侨大学厦门市计算机视觉与模式识别重点实验室 福建 厦门 361000

摘 要 视频中动作质量的评估指对视频中人物对象的动作质量进行评价,如计算动作质量分数、等级或者不同人物表现的优劣,是视频理解和计算机视觉研究中的一个重要方向。从动作质量分数预测、等级分类以及水平排序3个方面对视频中的动作质量评估方法进行总结,然后对这些方法在目前常用数据集上的表现进行分析,最后讨论未来研究中亟待解决的问题。

关键词:视频理解;行为质量评估;质量分数预测;等级分类;水平排序
中图法分类号 TP391.41

Survey on Action Quality Assessment Methods in Video Understanding

ZHANG Hong-bo[1],DONG Li-jia[1],PAN Yu-biao[2],HSIAO Tsung-chih[2],ZHANG Hui-zhen[2] and DU Ji-xiang[2,3]
1 School of Computer Science and Technology,Huqiao University,Xiamen,Fujian 361000,China
2 Fujian Key Laboratory of Big Data Intelligence and Security,Huaqiao University,Xiamen,Fujian 361000,China
3 Xiamen Key Laboratory of Computer Vision and Pattern Recognition,Huaqiao University,Xiamen,Fujian 361000,China

Abstract Action quality assessment refers to evaluate the action quality performed by human in video,such as calculating the quality score,level and evaluating the performance of different people. It is an important direction in video understanding and computer vision research. This paper summarizes the main methods of action quality assessment,including action quality score prediction methods,level classification and ranking methods. The performance of these methods on public datasets is also analyzed. Finally,the challenge problems in future research are discussed.

Keywords Video understanding,Action quality assessment,Quality score prediction,Grade classification,Level sort

1 概述

人体行为识别(Human Action Recognition)一直以来都是计算机视觉领域被广泛研究的课题。在近几十年里,基于计算机视觉的人体行为识别技术和方法发展迅速。在大规模的数据集以及高性能的计算设备的帮助下,越来越多可靠和高效的方法得到快速发展。作为人体行为识别的扩展领域,动作质量评估(Action Quality Assessment,AQA)由于在工业界有着广泛的应用,如病人的康复医学治疗[1-2]、体育运动的自动评分[3-10]、机器人自动化服务的技能判定[11-12]和特定技能的等级评估任务[13-15]等,受到了广泛关注,成为了近年来人体行为理解研究中新的热点问题。

视频中的人体行为质量评估研究也面临着方方面面的挑战,如需要考虑人类运动的多样性和复杂性、摄像机的运动、杂物遮挡、背景杂乱等问题。另一方面,不同于行为识别,行为质量评估是一项更细致化的工作,往往需要对行为的每个阶段分别进行评估,而这就需要解决运动检测与分割以及多层次动作评估度量等关键问题。

为了更好地梳理人体动作质量评估的相关研究工作,本文针对视频中的人体动作质量评估方法进行调研和总结,分析并比较了现有的人体动作质量评估的代表性研究方法及该研究方向存在的问题,最后对该任务未来研究的方向以及待解决的问题进行了讨论。

动作质量评估旨在以不同的方式量化动作的质量,达到衡量动作质量的目的。动作质量的量化往往以不同形式的结果呈现。早期该领域的部分工作通过回归具体分数,对不同任务中的动作质量做出评估。文献[16]使用0到1的分数对康复训练动作进行评价,其中0表示动作执行质量最差,1表示动作执行质量最好。该方法构建了基于几种不同类型神经网络的模型,用于对动作的分数进行回归预测。另有一些工作

到稿日期:2021-06-02 返修日期:2021-10-20
基金项目:国家自然科学基金(61871196);福建省自然科学基金(2019J01082);华侨大学优秀青年科研创新人才项目(ZQN-YX601)
This work was supported by the National Natural Science Foundation of China(61871196),Natural Science Foundation of Fujian Province,China(2019J01082) and Promotion Program for Young and Middle-aged Teachers in Science and Technology Research of Huaqiao University(ZQN-YX601).
通信作者:张洪博(zhanghongbo@hqu.edu.cn)

把质量级别作为分类结果，以此量化执行者动作的质量。Fawaz 等[17]通过对手术执行视频中的人体动作进行质量等级分类来实现对动作质量的评估，将动作质量分成“初学者”“中级”和“专家”3 种级别。除此之外，近年来，国内外学者结合不同的学习方式和网络结构，将等级排序作为评估动作质量的方式，即不进行具体的质量分数预测和等级分类，而是以视频中动作质量排序作为目标，对比视频中动作的质量优劣。

上述研究从不同角度对动作质量评估问题进行建模，设计了不同的网络模型以评价动作质量。基于现有研究，本文从这 3 种评估动作质量的角度出发，对动作质量评估领域的现有研究成果进行分类、总结和分析，如图 1 所示。

图 1 动作质量评估方法的分类

Fig. 1 Classification of action quality assessment methods

(1)以质量分数为评估结果

这类研究以所要评估的动作质量的具体分数作为标签，在构建准确有效的模型的情况下，对输入视频中的动作质量分数进行预测。从机器学习的角度来看，这类工作主要采用回归预测的方法。特别是在深度学习发展后，该类工作主要利用卷积网络提取视频中的时空特征，然后通过对视频片段之间的序列关系进行建模，形成视频级别的特征表示。以这种表示作为网络的输入，从而训练出用于回归的评分估计函数，最终得到视频中动作的具体质量分数。

(2)以等级类别为评估结果

这类研究将动作质量划分为多个水平等级，如“专业”“中等”“业余”。在以等级类别为评估结果的相关工作中，早期工作基于支持向量机(Support Vector Machines，SVM)[18]、动态时间规划(Dynamic Time Warping，DTW)[19]、Adaboost[20]等方法来完成分类任务。随着深度学习的发展和广泛应用，近年来该类工作也通过端到端网络来对动作的技能水平进行分类。

(3)以质量等级排序为评估结果

在动作质量评估领域中，大多数工作均基于上述两种评估方式，而近几年，以质量等级排序为评估结果的工作不断涌现。这类研究对于给定的动作视频序列，能够得到这些视频中动作的水平等级排序结果。该类方法大多结合度量学习(Metric Learning)及 Siamese 孪生神经网络[21]等方式和结构，通过对成对视频的技能水平进行对比和度量，生成技能水平排序。

另一方面，对于不同的动作质量评估任务，所采用的研究数据集和结果的衡量指标也不相同。在以质量分数为评估结果的任务中，预测分数和真实分数值的相关程度被作为评价指标；在以等级分类为评估结果的质量评估方法中，则将分类的准确率作为评价指标。本文将对目前主要的动作评估方法在不同任务中的数据集以及衡量指标上的表现进行总结和分析。

同时，针对视频中的动作理解任务是计算机视觉研究中的经典问题，目前已有许多对动作理解进行综述调研的工作。文献[22]从特征提取、动作识别的方法等方面阐述了该领域的研究现状；文献[23]从视频特征采样、特征描述符选择、特征处理和向量编码等方面综述了视频中的动作识别过程。但是这些工作大多集中于动作识别，对动作质量评估的综述研究较少。Lei 等[24]从不同的视频特征出发，分别从基于手工特征以及基于深度学习特征的角度，对质量分数回归问题进行了分类讨论。不同于上述综述文献，本文分别从人体动作质量评估的不同任务入手，结合近几年的工作，对人体动作质量评估领域中最新的方法进行更加全面的总结和分析。

综上所述，本文第 2 节依据上述 3 个不同任务对动作质量评估方法做出总结和概述，并比较分析不同任务中不同方法的异同；第 3 节将介绍不同任务中常用的数据集及实验结果；最后总结该领域研究成果，并展望未来发展。

2 动作质量评估方法

2.1 动作质量分数预测的方法

在基于质量分数的评估方法中，对于待评估的视频，首先将视频分割成剪辑级别(Clip-level)或帧级别(Frame-level)的数据，然后通过特征提取模块计算动作特征，将特征作为回归函数的输入，以得到质量评估分数。近年来，随着深度学习在动作理解研究中的应用，目前主流的方法采用基于端到端的回归网络，如图 2 所示。同时，随着动作分割方法的不断成熟和发展，以及鉴于个别评估任务难度系数和评价方式的特殊性，一些研究利用动作分割方法代替原来简单的视频剪辑和视频帧的采样方式，将视频划分成几个动作阶段，通过对每个阶段进行评估来实现对整个动作的评估。

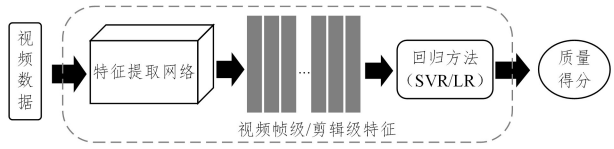

图 2 以质量分数为评估结果的动作质量评估方法

Fig. 2 Action quality assessment method based on quality score

Pirsiavash 等[25]引入深度学习方法来评估体育比赛中运动员的动作质量。该项工作也在一定程度上奠定了以回归分数方式完成质量评估的研究工作基础。该项工作利用三维卷积神经网络[26]来提取视频中的时空特征，将其分别与长短期记忆网络(Long Short-Term Memory，LSTM)[27]和支持向量回归(Support Vector Regression，SVR)[28]方法相结合，提出 3 种可行的动作质量评估框架，最后得到预测分数。在这项工作中，Pirsiavash 等提出“增量标签训练”的概念，用于提供子动作级别的反馈，目的是随着动作的进行，如果动作质量足够好，最终评估分数会逐渐增加，反之分数则减少。在真实赛事中，裁判的评级往往是主观的，如果要设计一个模型来评估选手的运动质量，只能尽可能使模型学习到这种主观的“规则”，而运动员在跳水过程中的动作变化等因素都是模型待学习的因素。换言之，文章从动作质量评估的众多角度之一出发，通过反馈的生成来对跳水动作进行分数评价。此外，这项工作基于人体的姿态信息，对于跳水运动来说，运动员的姿态变化速度快；对于花样滑冰运动来说，运动员的服装会对其动作变换造成一定程度上的遮挡。以上因素会导致姿态信息的提取不够准确，从而影响该文章的效果，且限制了应用场景。

为了探究 AQA 方法的泛化性,2019 年,Parmar 等[5]创建了包含跳水、滑雪等 7 种任务的多任务 AQA 数据集(AQA-7),将 C3D 网络作为骨干网络,对 7 种运动进行分数预测。该项研究在一定程度上说明了不同的运动之间具有共性,使用同样的基础网络均可以实现一定的预测效果。但是从深度学习的角度出发,数据是保证方法准确性的关键因素之一,AQA-7 数据集的规模导致 AQA 泛化性的讨论缺乏数据支撑。

同年,针对跳水任务,Parmar 等[6]使用多任务学习框架实现了动作质量评估任务。该框架包含 3 个与动作质量评估高度相关的任务,即行为识别、视频解说生成以及行为质量评估。该项工作证明了不同行为之间存在共同的特征,而来自多个任务的合并样本有助于改善评估方法的性能,同样使用 C3D 结合 LSTM 网络结构,在关注视频时空特征的同时,一定程度上避免了视频上时序信息的丢失。这项工作得到了较好的实验结果,证明多任务学习方法在影响因素众多的任务中表现良好。

Perše 等[29]提出一种基于贝叶斯网络的篮球活动自动评价方法。这项工作通过对多主体活动识别(Multi-agent Activity Recognition)[30]思想进行扩展,将 3 种不同类型篮球活动的 63 个轨迹段作为训练数据。与文献[30]类似,该评估方法使用了多主体贝叶斯网络、基本事件检测器和时间关系函数。在篮球比赛中,不同的战术对应不同的动作模板,不同的模板中对应的空间和时序分布均不相同。这项工作中的网络结构和时间关系则是自动从上述活动模板中获得的。最后构建出的贝叶斯网络用于学习不同战术的内在关系及外在关系。基本网络结构划分为 4 个层次:第一层代表整体的活动得分,第二层代表个体队员的成绩,第三层代表关键活动事件的执行,第四层代表关键活动之间的关系。前两个层次中的变量值依赖于第三、四层的值,而后两层的值则由基于轨迹的元素检测器和时间关系函数得到。该项工作除了可以更加准确地评估比赛结果,帮助篮球专家发现并消除队员表现不佳的原因之外,还可以扩展到其他场景中。

Xiang 等[8]基于伪 3D(Pseudo-3D,P3D)[31]提出了一种称为 S3D(Stacking Segmental P3D)的分段网络来解决体育活动评分问题。他们利用基于编解码的时间卷积网络(Encoder-decoder Temporal Convolution Network,ED-TCN)[32]将跳水视频分割为 4 段,分别为起跳、下落、入水和结束 4 个部分,并设计了 3 种不同的采样方法,包括随机采样、间隔采样以及在不同阶段的中心进行采样的方法,更具针对性地完成对动作的评估。在中心采样中,该方法结合 4 个不同的阶段完成片段级别(Segment-level)的特征提取,将不同阶段的分数作为代表性分数与任务整体分数进行比较。这项工作在一定程度上证明,将不同阶段的评估结果作为整体评估贡献的一部分能够更加有效地建立评分标准。同时,这项工作也在一定程度上体现了运动的子阶段可以作为评价体育运动执行的关键因素。

类似地,Li 等[4]引入了一种由关键片段分割部分(Key Fragment Segmentation,KFS)和评分预测(Score Prediction,SP)组成的新型评分网络。通过筛选跳水运动中的关键片段,包括起跳和入水两个阶段,来生成“执行分数”和“难度分数”两种分数,最后由上述两种分数生成视频最终分数。该项工作的分数预测方法遵循奥运会的评分标准,更加客观地实现了打分机制。该方法的思路在于关键片段的选取,故其只摘取整个运动视频中的部分片段,但这对于体育运动中的评价分数来说有失偏颇,因为在真实赛事中并不能要求运动员只完成运动的某一个阶段,运动员的整个表现过程对最后的评分至关重要。

Xu 等[33]的工作旨在对花样滑冰运动视频进行评分。该项工作设计了由自注意力 LSTM 模型(Self-Attentive LSTM)和多尺度卷积跳跃 LSTM 模型(Multi-scale Convolutional Skip LSTM)结合的网络结构,分别关注视频局部信息以及全局信息,并分别得到视频整体得分(Total Element Score,TES)以及组成成分子得分(Program Component Score,PCS)。整体得分用于判断所有技术动作的难度和执行力,组成成分子得分旨在评估表演者的表现以及整体表现对音乐的诠释程度。通过对分数的解耦,更好地实现了长视频体育运动的质量评估。由于该工作基于时空特征,因此对于有遮挡或背景颜色过深的视频帧,PCS 的预测容易出现误差。

由于运动的评估是基于与真实赛事视频进行比较来完成的,Jain 等[34]提出了一种较为新颖的动作质量评估方法,它将评估问题转化为比较给定动作视频和参考视频的问题。文章提出基于度量学习的 Siamese 孪生神经网络,通过对比成对输入的运动视频,得到相关度分数;然后将相关度分数与片段级别的子分数相结合,形成最终分数。同时,文章引入了一种无监督的技术来提供子动作级别的反馈,使分数的评估更有应用价值。

除了体育运动的自动评估,在健康应用领域,AQA 方法也涌现了一些以分数为评价结果的研究工作。具体任务包括对中风、卒中患者的康复训练辅助,设计针对特殊疾病人群的辅助穿戴设备等。早期,文献[35]提出一种用于帮助手语词汇学习的方法,它可以替代手语老师来帮助聋哑人自主学习手语词汇。文章对手语素材进行收集,并对手语视频、手语关键姿态、插图图示以及文字进行标识。该文使用传统的特征提取方法对在线捕捉的手语手势视频进行姿态和关节特征的提取,以对手语种类做出识别以及回归,并评估得出手语类型以及受试者做出手语动作的得分。这项工作是较早应用于健康领域的 AQA 研究,无论从实验设计还是整体方法来看,该文章都缺乏一定的创新性和研究价值,只是单纯地满足了手势评估的功能需求。

此后,随着深度学习在计算机视觉领域的广泛应用,健康领域也出现了一些基于学习的 AQA 研究。Vakanski 等[16]提出了 3 种基于端到端的深度学习模型,用于量化康复病人的运动表现。该项工作分别基于 CNN(Convolution Neural Network),RNN(Recurrent Neural Network)以及 HNN(Hierarchical Neural Network)来完成,最后通过学习得到评分函数,并把分数归一到 0～1 的范围内,以评价受试者的行为质量。从效果来看,这项工作相较之前的工作有了一定的创新和提升,但针对网络结构来说,3 种骨干网络都过于基础,对于特定任务缺乏针对性。

近年来，越来越多的学者和专家关注到 AQA 领域，并针对医疗健康、机器人辅助等领域开展了各种研究工作。Liu 等[15]提出了一种新的基于 RNN 的空间注意力模型，用于评估视频中动作的技能水平，动作包括使用筷子、婴儿抓握物品等。该框架由两个循环神经网络组成，一个用于每一帧的空间注意力估计，另一个用于技能分数的估计。该方法不仅关注视频中每一帧的即时视觉信息和针对具体任务的高层次相关知识，还通过注意力信息的获取和辅助实现动作技能的评估。

Parmar 等[36]创建了跨视、听两个模态的钢琴演奏评估数据集，并设计方法以实现钢琴演奏水平的评估任务。在新冠肺炎还未得到彻底解决的大环境下，这项工作也为线上教学和考核工作做出了一定程度的贡献。

除此之外，还有许多研究基于人体姿态和待评估视频的时空注意力特征实现不同领域的 AQA 任务。Nekoui 等[37]结合人体骨架以及姿态信息，使用图卷积等方法对跳水运动进行分数评估。与之相似，文献[38-39]更关注姿态信息。其中，文献[38]设计了基于注意力机制的上下文感知结构，以评估长视频中的运动质量；文献[39]设计了一种双流网络，分别关注人体关节点和帧级别的表观特征，以实现动作质量评估。同样用于体育运动评估的工作还有文献[40]，这项工作以裁判给出分数的分布来分析和计算选手的最终得分。许多待评估任务从视觉上分析是对称的，如体育运动中的双人跳水，一些日常动作执行中左右手的协调性等。文献[41]针对这些特殊任务进行对称性建模，以评估任务执行的质量。文献[42]将时序卷积和空间卷积解耦成两阶段方法，并结合注意力机制，实现体育运动质量分数评估。还有一些端到端的模型，如用于小众运动种类——高尔夫球[43]、瑜伽[44]等运动的分析与评估。

2.2 动作质量等级分类方法

此类任务的研究与视频中的动作识别类似，因此大多数方法均在行为识别研究的基础上进行扩展，将类别信息根据评估的需求进行改进。该类方法先对待评估的视频数据进行时空特征的提取，再由分类器分出不同等级的技能水平，如图 3 所示。

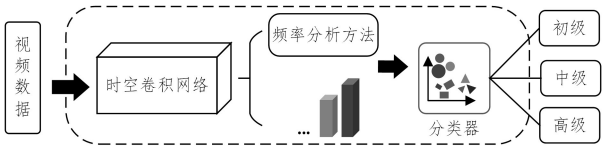

图 3 以等级类别为评估结果的动作质量评估方法

Fig. 3 Action quality assessment methods based on grade classifications

在早期的工作中，Parmar 等[45]对物理治疗康复中不同的执行等级进行分类，分类方法包括支持向量机、单层和双层神经网络以及增强决策树方法，最后通过类别信息确定动作质量是“好”还是“坏”。该项工作数据集的规模较小，不利于模型的训练和测试。

与直接评判技能类别类似，Baptista 等[46]使用动态时间规划方法，以得到受试者与模板动作相似的子序列的间隔。评估结果将对受试者提供反馈建议，以纠正其在操作上的错误。文章使用的数据集并非真实中风患者的行为采集数据，而是健康人士模拟中风患者的运动行为。由此可见，使用真实数据进行模型训练是医疗健康领域在 AQA 任务中亟待解决的问题之一。

Wang 等[7]通过分析多变量时间序列数据，首次将深度学习应用到评估手术技能上，这也是 AQA 任务应用在手术技能评估上的比较具有代表性的研究工作之一。这项工作提出了一个 10 层的卷积神经网络，包括 5 种类型的操作，即卷积、池化、扁平化、全连接和 softmax 回归，以提取用于自动技能评估的动力学特征。然而，有监督深度学习的分类准确率很大程度上依赖于数据集上标注的样本，由于 JIGSAWS 数据集[47]缺乏技能水平的 ground-truth 标签，因此这项评估工作只能对视频中的操作水平做出简单解释，无法从专业角度得出更详细、具体的评估结果。

Tao 等[48]提出了一种稀疏隐马尔可夫模型（Hidden Markov Model，HMM），以不同技能水平的外科医生的手术操作视频作为观测数据，设计方法以对医生在术中实操技能水平进行分类。文章提出的算法用于学习每个手势和描述不同手势之间转换的 HMM 语法，然后使用上述手势数据和语法来分类新的数据。与之相似，Fand 等[12]提出了一种基于运动轨迹的技能评估方法。该项工作建立了一个分类框架，用于自动评估不同专业水平外科医生的表现。与以往的手术技能评估工作不同，该项工作致力于评价机器人辅助的外科手术任务，将手术技能水平分为“专业”和“初级”两类。与文献[45]的问题相同，这项工作的局限性也在于数据的规模太小，评估的效果受限。

Fawaz 等[17]设计了一种 CNN 方法，其不但能够获取数据中的手势信息，还可以获取与手术技能水平相关的全局信息，最后得到 3 种不同类别的评估结果，即“初学者”水平、“中级”水平以及“专家”水平。该方法较为直接，使用简单的 CNN 结构对手术的技能水平进行分类，相比相同任务上的其他工作，该方法对技能的执行水平缺乏一定的针对性和解释性。

Zia 等[11]从手术视频中的频率域提取时空兴趣点[49]（Space-Time Interest Points，STIPs）特征，获得行为的多维时间序列表征；然后将视频运动信息编码为一个时间序列，并使用两种不同的频率分析方法[50-51]，即离散余弦变换（Discrete Cosine Transform，DCT）和离散傅里叶变换（Discrete Fourier Transform，DFT）对时间序列进行分析；最后选择最佳频率，区分 3 种技能等级，从而将医生技能表现分为业余水平、中级水平和专家水平。从数据角度来看，该项工作使用的数据和 JIGSAWS 数据集不同，JIGSAWS 的拍摄视角为机械臂的操作镜头；而该文章中数据集的拍摄视角为执行者的整个上半身，这更贴合现实中医护人员执行手术操作的场景，更具有现实应用价值。

2.3 动作质量等级排序的方法

这类方法多用于特定动作的质量评估任务，使用排序思想解决针对特殊技能表现的评估问题。评估结果为特定动作的质量等级排序，技能表现出色的排名靠前，技能表现逊色的则排名靠后。图 4 给出了该类方法中网络的训练过程。该类方法多使用度量学习的思想，结合孪生神经网络的结构完成权重共享，学习评分函数，用于排序任务。

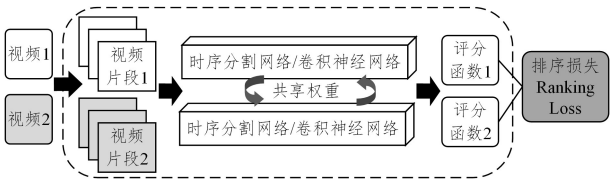

图 4 以质量等级排序为评估结果的动作质量评估方法

Fig. 4 Action quality assessment methods based on quality ranking classifications

Carvjal 等[52]提出一种用于自动检测环球小姐晚礼服比赛中的获胜者的方法。该项工作将所要解决的问题分解为两个子问题，分别是针对同一年度整体参赛选手的整体排序(Listwise Ranking)，以及针对同年度两名选手的成对排序(Pairwise Ranking)。文章提出了一种同时解决这两个问题的方法，并将所提出的方法与文献[30]中用于动作分析的视频描述符结合使用。其中，视频描述符从像素基上提取，并利用梯度和光流作为有效的视频表示。然后使用 Parmar 等[10]提出的堆叠 Fisher 向量(Stacked Fisher Vector，SFV)方法[53]对视频描述符进行编码，完成两种排序任务。由于该项任务较为新颖，因此也面临数据集规模小的问题。其次，由于晚礼服比赛需要考虑到模特的走姿等，因此在评估方法中加入姿态信息也是该项任务未来值得尝试的工作之一。

Doughty 等[13]提出了一种用于视频中技能确定的深度成对排序方法，并将所提方法应用于各种任务，如外科手术中的缝合、烹饪中的卷饼及使用筷子等动作。该项工作利用双流 CNN 提取所有成对视频的时空特征表示，最后在损失函数中计算成对视频的相似度，以度量特定任务的完成质量。

在文献[13]的后续研究中，Doughty 等[14]引入了新的数据集，用于研究长视频中的技能确定。对于输入的成对技能视频，使用全新的网络结构来完成视频中技能水平等级的排序。该项工作提出了一个由 I3D(Inflated 3D ConvNets)[54]网络提取特征的等级感知时间注意力网络，其中的注意力模块使用多个注意力过滤器来挖掘长视频中的重要部分，从而完成技能质量的评估。除此之外，该项工作使用 3 种类型的损失函数训练网络，并取得了较好的结果。

对于文献[13-14]提出的技能评价任务来说，除了要在已知的任务类别上取得较好的效果外，更重要的是在未知的任务种类上有更好的泛化能力。从长期的研究目标上看，这也是 AQA 任务需要考虑的问题之一。

3 数据集和结果分析

本节将对不同评估任务中的数据集以及不同方法在该数据集上的结果进行对比和分析。

3.1 质量分数预测任务

3.1.1 评价指标

在基于质量分数的动作质量评估工作中，大部分研究采用斯皮尔曼等级相关系数(Spearman Rank Correlation)作为评价指标。该评价指标没有明确强调真实的分数值，而是表示真实值与预测值的相关程度。

斯皮尔曼等级相关系数反映两组变量之间联系的密切程度，取值在 $-1\sim+1$ 之间。若两个变量完全单调相关，则斯皮尔曼等级相关系数为 $+1$ 或 -1。

对于测试数据而言，假设测试数据的样本容量为 n，x_i 表示评估方法得到的第 i 个样本的预测分数值，y_i 表示第 i 个样本的分数真实值。对 x_i 和 y_i 按从小到大排序，记 x_i' 和 y_i' 为原始 x_i 和 y_i 在排序列表中的位置，称它们为 x_i 和 y_i 的秩次，则秩次差 d_i 为：

$$d_i = x_i' - y_i' \tag{1}$$

则两个变量的斯皮尔曼等级相关系数 ρ_s 为：

$$\rho_s = \frac{6\sum d_i^2}{n(n^2-1)} \tag{2}$$

3.1.2 常用数据集

以回归分数作为评估结果的方法中多使用与体育运动相关的数据集。该领域常用的数据集及其属性如表 1 所列。

表 1 基于质量分数的体育运动数据集

Table 1 Physical activity datasets based on quality scores

Dataset	Release Time	Number of Action Classes	Number of Samples	Category of Samples
MIT Olympic Scoring[25]	2014	2	68	Diving, Figure skating
UNLV Olympic Scoring[10]	2017	3	716	Diving, Figure skating, Vault
UNLV-AQA-7[5]	2018	7	1 189	Diving(including single/synchronized 10 m platform diving, synchronized 3m platform diving), Vault, Skiing, Trampoline
MTL-AQA[54]	2019	1	1 412	Diving
Fis-V[33]	2019	1	500	Figure skating
GolfDB[43]	2019	1	1 400	Golf
Yoga Vid Collected[44]	2019	6	88	Yoga

MIT Olympic 数据集[25]包含跳水和花样滑冰两种运动样本。其中，跳水数据集包含 2012 年奥运会男子 10 m 跳台预赛的 159 个视频；花样滑冰数据集包含 150 个视频，每秒 24 帧。这些视频的平均时长为 2.5 min，且其视角随着表演不断变化。这些数据的标签为比赛过程中评委的打分，数值为 0～100。

UNLV Olympic 动作评估数据集[5,10]中包含跳水和跳马两种动作。其中，跳水部分是对 MIT Olympic 数据集跳水的扩展，该数据集包括半决赛和决赛的视频，共计 370 个，每个视频包含约 150 帧；而跳马部分数据集包括 176 个视频，这些视频相对较短，平均长度约 75 帧。跳马分数由执行分数和难度分数的总和决定，最终分数值范围为 0～20。

Parmar 等[5,10]提出的用于多任务动作质量评估的数据集——UNLV-AQA-7 包含 1 412 个跳水视频，比赛类型包括 10 m 跳台和 3 m 跳板，运动员种类包含男运动员和女运动员，数据集从不同视角收集。数据集中包含了每个样本的动作质量评估分数、跳水类型以及对比赛过程的解说。在 MTL-AQA 数据集中，除了上述跳水运动之外，还包含了用于跨任务评估的其他几种运动类型，包括体操跳马、滑雪、蹦床。

Parmar 等[55]后续又对 MIT-diving 数据集进行了二次扩充，提出了 MTL-AQA 数据集。该数据集用于多任务学习方法下对跳水运动进行评估，其跳水样本数量达到 1 412 个，是

UNLV 跳水数据集的 4 倍之多。相比现有评估跳水方法在 UNLV 数据集上的效果，该方法效果更好。除此之外，该项工作还对 UNLV 数据集进行了进一步的扩充，得到新的 MTL-AQA 数据集。

Xu 等[33]对 MIT Olympic 数据集中的花样滑冰数据进行扩展，提出了 Fis-V 数据集。该数据集的视频数比 MIT 花样滑冰数据集多出 3 倍。

文献[43-44]对高尔夫和瑜伽运动进行了数据集的收集和标注，构建了 GolfDB 数据集和 YogaVidCollected 数据集。其中，GolfDB 数据集包括 1 400 个数据样本，视角不断变化；YogaVidCollected 数据集包含 6 个种类的瑜伽运动，共 88 个数据样本。

3.1.3 实验结果对比

表 2 总结调研了各方法在不同数据集上的实验结果，这些结果均采用斯皮尔曼等级相关系数作为衡量指标。

表 2 各类方法在基于分数的数据集上的表现

Table 2 Performance of various methods on score-based datasets

Method	Year	MIT-Diving	UNLV-Skating	UNLV-Diving	UNLV-Vault	AQA-7
文献[25]	2014	0.41	0.35	—	—	—
文献[56]	2015	0.45	—	—	—	—
文献[10]	2017	0.74	0.53	0.79	0.68	—
文献[5]	2018	—	—	0.61	0.67	—
文献[3]	2018	—	0.57	0.80	**0.70**	—
文献[4]	2018	**0.78**	—	0.84	**0.70**	—
文献[8]	2018	—	—	0.86	—	—
文献[33]	2019	—	**0.59**	—	—	—
文献[55]	2019	—	—	**0.88**	—	—
文献[34]	2020	—	—	0.69	—	—
文献[40]	2020	—	—	—	—	0.81
文献[38]	2020	—	0.62	—	—	—
文献[42]	2020	—	0.71	0.85	0.76	—
文献[39]	2021	—	—	0.83	0.74	0.81

在 MIT 跳水数据集以及 UNLV 体操跳马数据集上，Li 等的两项工作[3-4]在上述两个数据集上均取得了截至目前最好的结果。在 MIT 跳水数据集上，Li 等[4]提出的基于关键片段的分割网络效果最好，相关系数达到 0.78。这在一定程度上说明了体育运动中的“关键片段”对质量评估至关重要。与这两种方法类似，Xiang 等[8]提出的基于时序阶段的方法也在 UNLV 跳水数据集上得到了 0.86 的相关性。除了跳水任务，在 UNLV 体操数据集上，Li 等[3]提出的端到端的评估动作质量网络也在体操数据集上达到 0.70 的相关值。

在花样滑冰任务中，Xu 等[33]的工作效果最佳，该项工作提出的两种模型可以分别有效地学习局部信息和全局信息，得到视频整体得分以及组成成分子得分，这两个分数实际上是对花样滑冰选手表现的两个不同方面的评价。组成成分得分这一思想也从一定程度上印证了上文中提到的基于关键片段或时序阶段信息的重要性和有效性。

在 UNLV 跳水数据集上，使用多任务学习方法的 Parmar 等[55]的工作效果最佳，相关性达到 0.88。由此可见，除了利用时序信息，还应充分利用数据中的其他信息，如运动片段的类内信息、比赛过程中的解说信息等。

AQA-7 数据集包含 7 类运动比赛任务的数据。在多个运动类别上，文献[39-40]中方法的平均相关系数值均达到 0.81。其中，文献[39]基于姿态和人体骨架信息，文献[40]从裁判打分的分布信息角度出发实现 AQA 任务。在 AQA 任务的泛化性研究上，上述两种方式均为相关工作奠定了一定的基础。

3.2 等级水平分类和排序任务

3.2.1 评价指标

由于等级水平分类与水平排序的工作采用的方法和数据集较为类似，因此本文将上述两类任务相结合进行讨论分析。

在基于等级水平分类的动作质量评估工作中，将技能等级的分类（如将技能水平分为“新手”“中级”“专业水平”3 种类别）作为评估结果，并将“分类准确率”作为最终评价指标。假设数据集中有 n 个样本，分类正确的样本数量为 m，则分类准确率描述为：

$$acc=\frac{m}{n}\times 100\% \tag{3}$$

对于等级水平排序的工作，采用“排名准确率”作为评价指标。“排名准确率”即技能水平的相关性排序预测结果的准确率，该类任务不考虑单个执行者的执行分数，而是着重于对每个视频的相对技能进行排名。对于已有的工作，使用的准确率均为成对视频的预测精度（Pairwise Precision），即排序结果中正确排序对的百分比。对于成对视频 (p_i, p_j)，定义 $E(p_i, p_j)$ 为评估结果函数：

$$E(p_i,p_j)=\begin{cases}1, & p_i>p_j\\ -1, & p_i<p_j\\ 0, & p_i=p_j\end{cases} \tag{4}$$

其中，$p_i>p_j$ 表示成对视频中第 i 个视频的表现优于第 j 个视频；$p_i<p_j$ 表示第 j 个视频的表现优于第 i 个视频；$p_i=p_j$ 表示成对视频表现相同。通过训练，得到动作质量得分函数 $f(\cdot)$。对于成对视频，假设标签真实值 $E(p_i,p_j)=1$，若模型输出的得分函数值 $f(p_i)>f(p_j)$，则判断预测准确。对于存在 n 对样本视频的数据，若判断准确的成对视频数量为 m，则最后的排名准确率描述为：

$$acc_{\text{rank}}=\frac{m}{n}\times 100\% \tag{5}$$

除上述两种指标之外，文献[52]对于成对的环球小姐晚礼服比赛数据采用归一化折损累计增益（Normalized Discounted Cumulative Gain，NDCG）作为评价指标。该项工作将这一指标用于计算动作质量评价模型的预测值与每年的实际排名（真实值）的相似程度。NDCG 方法常用于搜索算法的评估，该项工作将每个视频样本作为待排序的样本，即为每个视频样本分配 1～10 的分数，10 分代表排名最高，1 分代表最低。NDCG 方法定义如下：

$$NDCG_{@b}=\frac{DCG_{@b}}{IDCG_{@b}} \tag{6}$$

其中，$DCG_{@b}$（Discount Cumulative Gain）表示特定排名位置 b 的折现累积收益，定义为：

$$DCG_{@b}=\sum_{j=1}^{b}\frac{2^{r(j)}-1}{\log_2(\max(2,j))} \tag{7}$$

其中，$r(j)$ 表示第 j 个样本的排序结果。

式(6)中的 $IDCG_{@b}$ 表示理想情况下最大的 DCG 值。预测完全准确时 $NDCG_{@b}$ 值为 1。

3.2.2 常用数据集

在该方向的研究中，多项研究工作建立了用于评估受试

者在特定任务中的行动表现的数据集,如外科手术技能训练、老年人日常生活的监测和协助等。

表 3 整理了该应用领域已经引入的一些公开可用的数据集,包括 JIGSAWS 数据集[47]、EPIC-Skills 2018 数据集[13]、BEST 2019 数据集[14]以及早期工作中用于环球小姐礼服大赛评选的 MU(Miss Universe)数据集[52]等。

表 3 用于技能判定评估的数据集

Table 3 Datasets for skill determination assessment

Dataset	Release Time	Number of Action Classes	Number of Samples	Category of Samples
JIGSAWS[47]	2016	3	103	Suturing, KnotTying, Neede passing
MU[52]	2016	1	105	Miss Universe Dress Selection
EPIC-Skills 2018[13]	2018	6	216	Painting, Wraps, Surgery, Using chopsticks
BEST 2019[14]	2019	5	500	Beat eggs, Braid hair, Tie knots, Origami cranes, Draw eyeliner
PIano Skills Assessment (PISA)[36]	2021	1	992	Piano playing

JIGSAWS 数据集[47]包含外科医生在台式模型上执行的 3 项基本手术任务的数据。这项 3 任务包括缝合、打结、穿针,通常被作为外科技能培训课程的一部分。该数据集的标注分为两种,一种为不同任务的手势标注,如手势标签"G1"表示"右手拿针"这一姿势,数据集中每次用于测试的所有帧都分别对应一个手势标签;另一个标注为技能水平等级标签,其也是技能水平评估领域的常用指标。技能水平评分被定义为 3 个分数:1,3,5。对于不同的动作,3 种分数也分别对应了不同的注释。

EPIC-Skills 2018 数据集[13]作为 JIGSAWS 数据集的扩充,在包含 3 种类型的手术动作的基础上,加入了包括绘画(Drawing)、卷饼(Dough-Rolling)和使用筷子(Chopstick-using)3 种类别的动作。手术类别来源于 JIGSAWS 数据集,无须再次标注。其余 3 种类别动作的标注使用亚马逊的自动标注机器人(Amazon Mechanical Turk,AMT)完成。AMT 是一个众包(Crowdsourcing)网站,为服务请求者雇佣远程的"众包工人"来执行计算机目前无法完成的任务。AMT 的工作人员同时观看成对的视频,并选择显示较高水平技能的视频来完成给定的任务。每一对视频都由 4 位不同的工作人员进行注释。成对视频的相对排名情况作为最后的排名标签。

BEST 2019 数据集[14]为 EPIC-Skills 2018 数据集[11]的后续工作,包括 500 个视频。由于 EPIC-Skills 2018 数据集的视角单一,因此 BEST 2019 数据集中的视频从 YouTube 上检索和编辑,其中包含 5 项技能任务:炒鸡蛋、编辫子、系领带、折纸鹤和画眼线。每个视频都标注了包含"B""I""E"的类别标签。"B"表示初学者(Beginner),"I"表示中级水平(Intermediate),"E"表示专业水平(Expert)。除此之外,在数据集中 40%的视频上进行成对视频技能等级排序结果的标注。

MU 数据集[52]包括从 1996—2010 年间有视频和官方成绩年份(共 10 年)的环球小姐晚礼服比赛视频。数据集中包括 105 个视频,18 343 幅描绘每位候选人在晚礼服比赛中走秀的画面。数据集标注为各年份比赛中的参赛真实成绩。

3.2.3 实验结果对比

表 4 列出了以等级类别和质量等级排序为评估结果的两类工作在常用数据集上的实验结果。

表 4 各类方法在技能评估常用数据集上的表现

Table 4 Performance of various methods on common datasets for skill assessment

(单位:%)

Method	Year	Evaluation Metric	JIGSAWS			EPIC-Skills 2018				BEST 2019	MU
			Suturing	KnotTying	Needle Passing	Dough-Rolling	Drawing	Chopstick-Using	Infant-Grasp		
文献[48]	2012	Classification Accuracy	97.4	96.2	94.4	—	—	—	—	—	—
文献[52]	2016	Normalized Discounted Cumulative Gain	—	—	—	—	—	—	—	—	**87.02**
文献[57]	2017	Classification Accuracy	89.7	96.3	61.1	—	—	—	—	—	—
文献[58]	2018	Classification Accuracy	**100.0**	**99.9**	**100.0**	—	—	—	—	—	—
		Correlation Coefficient	0.75	0.63	0.46	—	—	—	—	—	—
文献[12]	2018	Classification Accuracy	89.9	95.8	82.3	—	—	—	—	—	—
文献[7]	2018	Classification Accuracy	93.4	89.8	84.9	—	—	—	—	—	—
文献[59]	2019	Classification Accuracy	**100.0**	—	96.4	—	—	—	—	—	—
文献[13]	2018	Classification Accuracy		73.3	80.4	83.2	71.5	—	—	—	—
文献[15]	2019	Classification Accuracy		73.1		**82.7**	**85.3**	**85.5**	**86.1**	—	—
文献[14]	2018	Classification Accuracy	—	—	—	—	—	80.3	**81.2**	—	—

由于数据集标注有限，2019 年以前的评估方法大多将技能的分类准确率作为评价指标。近年来，文献[13-14]将技能评估任务与度量学习方法相结合，并在现有数据集的基础上进行数据和标注的扩充，以成对视频的排名准确率作为评价指标。

在用于手术技能的 JIGSAWS 数据集上，Zia 等[58]提出的方法从机器人运动学数据(Robot Kinematic Data)中提取 4 种不同类型的整体特征——序列运动纹理特征(Sequential Motion Texture，SMT)、离散傅里叶变换、离散余弦变换和近似熵(Approximate Entropy，ApEn)特征；然后利用这些特征进行技能分类和技能得分预测，得到了 100%的分类准确率。同样地，Funke 等[59]提出的基于深度学习的方法也在该数据集上得到了 100%的分类准确率。该项工作从手术视频中提取若干连续帧的堆栈，并在训练过程中将网络扩展为时域网络来完成对技能等级的分类任务。

但是，JIGSAWS 数据集规模较小，这也是上述工作中较为局限的地方。数据集的规模对模型性能影响很大，数据集过小很容易造成模型的过拟合，也会使得模型的泛化能力不强。

对于后续扩充的 EPIC-Skills 数据集，Li 等[15]提出的方法在扩充的 3 个新任务上的排名准确率均达到了目前最高。除此之外，该项工作还进一步对数据集进行了扩充，加入了婴儿动作数据，用于评估婴儿的抓握能力。这项工作基于注意力机制，从一定角度体现了注意力机制 AQA 在任务中的有效性。

Doughty 等[14]在扩充的 BEST 2019 数据集上对成对任务的排序工作做出了重要贡献。该项工作提出的方法在两个数据集上的平均性能都超过了 80%。对于文献[52]提出的 MU 数据集，文章所提方法在 1998 年比赛数据上的测试效果达到最好，指标系数为 87.02%。

结束语 视频中的人体行为识别已经成为计算机视觉中炙手可热的研究方向，而人体行为质量评估问题在近年来也受到了广泛关注。在以往的研究中，人体动作质量评价或动作评分一直被看作一个回归问题，但在近几年，研究表明基于回归的简单解决方案缺乏可解释性，而且对结果的评价过于简单，于是产生了不同的评估方式(如基于质量等级分类的质量评估方式)。除此之外，研究人员还提出了基于等级排序的评估方法，在基于排序的评估方式中，使用了度量学习的方式对成对视频进行比较，达到了评估受试者技能的目的。

然而，在该领域的研究中，仍然存在一些亟待解决的问题。

(1)现存的评估模型仅面向特定动作任务。虽然有一些多任务的数据集用于动作质量评估，如 AQA-7 数据集、JIGSAW 数据集等，但是目前的研究大多数都集中于对单一的动作建立模型，即针对每种动作单独训练模型进行评估。构建一个开放的模型，使其能够自适应不同动作之间的差异以进行多种动作的评价，并解决复杂动作的评估问题，扩展动作质量评估的应用，是动作质量评估领域一个可分析和解决的问题。

(2)对动作质量的分数预测分析过于单一。目前，对动作质量分数的预测工作主要使用回归模型对给定视频中的动作进行分数预测。然而，不同的动作或者行为具有不同的复杂性，如简单的挥手动作、高尔夫球的挥杆动作或者一套跳水动作。对于这些复杂层次不同的动作，采用同一种方法构建的模型过于简单，无法对动作进行更加细致的分析。对于如“跳水”之类的复杂动作，应当进行多阶段多层次的评价，才能得到更加有效和准确的评估结果，更有助于动作质量方法在真实场景中的应用。

(3)虽然已经有部分工作进行分阶段评估，但是这些方法对视频多个阶段的分割方式较为简单。现存的大多数方法都采用等分视频的方法来减小学习网络的参数规模，只有一些近期的方法使用了较为准确且有针对性的时序分割模型来对原始视频进行阶段层面的分割。除此之外，由于分割过度或错误分割可能导致重要的时间信息丢失，因此分割的准确率也将作为后续评估准确率的重要影响因素。

(4)以排序作为评估结果的动作质量评估方法中，现有研究工作只对成对视频进行相对排序。在排序学习算法中，除了成对排序之外，还有单任务分类排序(Point Wise Ranking)、列表排序(Listwise Ranking)两类排序方式。而近年来的研究方法均基于成对排序思想对视频中的技能进行评估排序，缺少使用其他排序思想进行评估任务的工作。

(5)数据集仍较为欠缺。针对特定领域的专家进行专业注释的人力成本巨大，特别是对复杂动作的多阶段评价，所需要的代价更大，目前仍然缺乏具有多种动作类别和多种应用领域的大规模注释数据集。大多数已有模型都是基于图像和视频分类的大规模数据集进行预训练，因此基于弱监督和半监督的方法也是当前动作质量评估研究的主要方向。

参 考 文 献

[1] ANTUNES M, BAPTISTA R, DEMISSE G, et al. Visual and human-interpretable feedback for assisting physical activity [C]// European Conference on Computer Vision. Cham: Springer, 2016: 115-129.

[2] PAIEMENT A, TAO L, HANNUNA S, et al. Online quality assessment of human movement from skeleton data [C]// British Machine Vision Conference. BMVA press, 2014: 153-166.

[3] LI Y, CHAI X, CHEN X. End-to-end learning for action quality assessment [C]// Pacific Rim Conference on Multimedia. Cham: Springer, 2018: 125-134.

[4] LI Y, CHAI X, CHEN X. ScoringNet: learning key fragment for action quality assessment with ranking loss in skilled sports [C]// Asian Conference on Computer Vision. Cham: Springer, 2018: 149-164.

[5] PARMAR P, MORRIS B T. Action quality assessment across multiple actions [C]// 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019: 1468-1476.

[6] PARMAR P, MORRIS B T. What and how well you performed? A multitask learning approach to action quality assessment [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 304-313.

[7] WANG Z, FEY A M. Deep learning with convolutional neural

network for objective skill evaluation in robot-assisted surgery [J]. International Journal of Computer Assisted Radiology and Surgery,2018,13(12):1959-1970.

[8] XIANG X,TIAN Y,REITER A,et al. S3d:Stacking segmental p3d for action quality assessment[C]// 2018 25th IEEE International Conference on Image Processing(ICIP). IEEE,2018:928-932.

[9] XU C,FU Y,ZHANG B,et al. Learning to score figure skating sport videos[J]. IEEE Transactions on Circuits and Systems for Video Technology,2019,30(12):4578-4590.

[10] PARMAR P,MORRIS B T. Learning to score olympic events [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017:20-28.

[11] ZIA A,SHARMA Y,BETTADAPURA V,et al. Automated assessment of surgical skills using frequency analysis[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham:Springer,2015:430-438.

[12] FARD M J,AMERI S,ELLIS R D,et al. Automated robot-assisted surgical skill evaluation:Predictive analytics approach [J/OL]. The International Journal of Medical Robotics and Computer Assisted Surgery,2018,14(1). https://onlinelibrary.wiley.com/doi/10.1002/rcs.1850.

[13] DOUGHTY H,DAMEN D,MAYOL-CUEVAS W. Who's better? Who's best? Pairwise deep ranking for skill determination [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:6057-6066.

[14] DOUGHTY H,MAYOL-CUEVAS W,DAMEN D. The pros and cons:Rank-aware temporal attention for skill determination in long videos[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:7862-7871.

[15] LI Z,HUANG Y,CAI M,et al. Manipulation-skill assessment from videos with spatial attention network[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019.

[16] LIAO Y,VAKANSKI A,XIAN M. A deep learning framework for assessing physical rehabilitation exercises[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering,2020,28(2):468-477.

[17] FAWAZ H I,FORESTIER G,WEBER J,et al. Evaluating surgical skills from kinematic data using convolutional neural networks[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer,2018:214-221.

[18] DRUCKER H,WU D,VAPNIK V N. Support vector machines for spam categorization[J]. IEEE Transactions on Neural networks,1999,10(5):1048-1054.

[19] BERNDT D J,CLIFFORD J. Using dynamic time warping to find patterns in time series[C]// Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (AAAIWS '94). 1994:359-370.

[20] FREUND Y,SCHAPIRE R E. Experiments with a new boosting algorithm[C]// ICML. 1996:148-156.

[21] BROMLEY J,GUYON I,LECUN Y,et al. Signature verification using a"siamese" time delay neural network[J]. Advances in Neural Information Processing Systems,1993,6:737-744.

[22] HU Q,QIN L,HUANG Q M. A Survey of Human Action Recognization based Vision[J]. Chinese Journal of Computers,2013,36(12):2512-2524.

[23] LUO H,WANG C J,LU F. Survey of video behavior recognition [J]. Journal on Communications,2018,39(6):169.

[24] LEI Q,DU J X,ZHANG H B,et al. A survey of vision-based human action evaluation methods[J]. Sensors,2019,19(19):4129.

[25] PIRSIAVASH H,VONDRICK C,TORRALBA A. Assessing the quality of actions[C]// European Conference on Computer Vision. Cham:Springer,2014:556-571.

[26] TRAN D,BOURDEV L,FERGUS R,et al. Learning spatiotemporal features with 3d convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:4489-4497.

[27] HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[28] DRUCKER H,BURGES C J C,KAUFMAN L,et al. Support vector regression machines[J]. Advances in Neural Information Processing Systems,1997,9:155-161.

[29] PERŠE M,KRISTAN M,PERŠ J,et al. Automatic evaluation of organized basketball activity using bayesian networks[M]. NA,2007.

[30] CARVAJAL J,SANDERSON C,MCCOOL C,et al. Multi-action recognition via stochastic modelling of optical flow and gradients[C]// Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis. 2014:19-24.

[31] QIU Z,YAO T,MEI T. Learning spatio-temporal representation with pseudo-3d residual networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:5533-5541.

[32] LEA C,FLYNN M D,VIDAL R,et al. Temporal convolutional networks for action segmentation and detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:156-165.

[33] XU C,FU Y,ZHANG B,et al. Learning to score figure skating sport videos[J]. IEEE Transactions on Circuits and Systems for Video Technology,2019,30(12):4578-4590.

[34] JAIN H,HARIT G,SHARMA A. Action quality assessment using siamese network-based deep metric learning[J]. IEEE Transactions on Circuits and Systems for Video Technology,2020,31(6):2260-2273.

[35] CHAI X,LIU Z,LI Y,et al. SignInstructor:an effective tool for sign language vocabulary learning[C]// 2017 4th IAPR Asian Conference on Pattern Recognition(ACPR). IEEE,2017:900-905.

[36] PARMAR P,REDDY J,MORRIS B. Piano Skills Assessment [J]. arXiv:2101.04884,2021.

[37] NEKOUI M,CRUZ F O T,CHENG L. FALCONS:Fast Learner-grader for Contorted poses in Sports[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020:900-901.

[38] ZENG L A,HONG F T,ZHENG W S,et al. Hybrid Dynamic-

static Context-aware Attention Network for Action Assessment in Long Videos[C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020:2526-2534.

[39] NEKOUI M, CRUZ F O T, CHENG L. EAGLE-Eye: Extreme-Pose Action Grader Using Detail Bird's-Eye View[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021:394-402.

[40] TANG Y, NI Z, ZHOU J, et al. Uncertainty-aware score distribution learning for action quality assessment[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:9839-9848.

[41] GAO J, ZHENG W S, PAN J H, et al. An asymmetric modeling for action assessment[C]// European Conference on Computer Vision. Cham: Springer, 2020:222-238.

[42] WANG J, DU Z, LI A, et al. Assessing Action Quality via Attentive Spatio-Temporal Convolutional Networks[C]// Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Cham: Springer, 2020:3-16.

[43] MCNALLY W, VATS K, PINTO T, et al. Golfdb: A video database for golf swing sequencing[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[44] YADAV S K, SINGH A, GUPTA A, et al. Real-time Yoga recognition using deep learning[J]. Neural Computing and Applications, 2019, 31(12):9349-9361.

[45] PARMAR P, MORRIS B T. Measuring the quality of exercises [C]// 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2016:2241-2244.

[46] BAPTISTA R, ANTUNES M, AOUADA D, et al. Video-based feedback for assisting physical activity[C]// 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP). 2017.

[47] GAO Y, VEDULA S S, REILEY C E, et al. JHU-ISI gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling[C]// MICCAI Workshop: M2cai. 2014.

[48] TAO L, ELHAMIFAR E, KHUDANPUR S, et al. Sparse hidden markov models for surgical gesture classification and skill evaluation[C]// International Conference On Information Processing in Computer-assisted Interventions. Berlin: Springer, 2012:167-177.

[49] LAPTEV I. On space-time interest points [J]. International Journal of Computer Vision, 2005, 64(213):107-123.

[50] AHMED N, NATARAJAN T, RAO K R. Discrete cosine transform [J]. IEEE Transactions on Computers, 1974, 100(1): 90-93.

[51] WEINSTEIN S, EBERT P. Data transmission by frequency-division multiplexing using the discrete Fourier transform[J]. IEEE transactions on Communication Technology, 1971, 19(5): 628-634.

[52] CARVAJAL J, WILIEM A, SANDERSON C, et al. Towards Miss Universe automatic prediction: The evening gown competition[C]// 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016:1089-1094.

[53] PENG X, ZOU C, QIAO Y, et al. Action recognition with stacked fisher vectors[C]// European Conference on Computer Vision. Cham: Springer, 2014:581-595.

[54] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? Anew model and the kinetics dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:6299-6308.

[55] PARMAR P, MORRIS B T. What and how well you performed? A multitask learning approach to action quality assessment [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:304-313.

[56] VENKATARAMAN V, VLACHOS I, TURAGA P K. Dynamical Regularity for Action Analysis[C]// BMVC. 2015:1-12.

[57] FORESTIER G, PETITJEAN F, SENIN P, et al. Discovering discriminative and interpretable patterns for surgical motion analysis[C]// Conference on Artificial Intelligence in Medicine in Europe. Cham: Springer, 2017:136-145.

[58] ZIA A, ESSA I. Automated surgical skill assessment in RMIS training[J]. International Journal of Computer Assisted Radiology and Surgery, 2018, 13(5):731-739.

[59] FUNKE I, MEES S T, WEITZ J, et al. Video-based surgical skill assessment using 3D convolutional neural networks[J]. International Journal of Computer Assisted Radiology and Surgery, 2019, 14(7):1217-1225.

ZHANG Hong-bo, born in 1986, Ph.D, associate professor, master tutor, is a member of China Computer Federation. His main research interests include computer vision, machine learning and video understanding.

(责任编辑:喻藜)