文章编号：2095-2163(2023)11-0001-13 中图分类号：TP391 文献标志码：A

体育视频中动作识别技术研究综述

游义平[1]，季云峰[2]
（1 上海理工大学 健康科学与工程学院，上海 200093；2 上海理工大学 机器智能研究院，上海 200093）

摘 要：随着中国成功举办多项国际体育赛事以及互联网短视频平台的兴起，视频数据呈爆炸式增长，且体育运动越来越受到人们的关注，体育视频中的动作识别成为计算机视觉研究的一大热点问题。本文综述了体育视频中动作识别技术现有应用与研究方法，第一部分回顾了近年来动作识别在体育赛事中的应用现状，将其归纳为辅助判罚、精彩动作集锦、体育新闻自动生成。第二部分总结了体育视频动作识别相关数据集。第三部分回顾了近年来动作识别在体育视频中的实现方法，将其总结为基于传统手工特征的算法和基于深度学习的算法，基于深度学习的算法将其归纳为基于 2D 模型、基于 3D 模型、基于双流/多流模型、基于 Transformer 模型，并总结了各模型的优缺点。最后，讨论了体育视频动作识别的难点与挑战。
关键词：动作识别；深度学习；体育运动

A review of research on action recognition methods in sports video

YOU Yiping[1], JI Yunfeng[2]
(1 School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;
2 Institute of Machine Intelligence, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: With the successful hosting of many international sports events in China and the emergence of Internet short video platforms, video data is exploding and sports are getting more and more attention. Action recognition in sports video has become a hot topic in computer vision research. This paper reviews the existing applications and research methods of action recognition technology in sports video, and the first part of this paper reviews the current situation of action recognition applications in sports events in recent years, and summarizes them as auxiliary penalty, highlight action collection, and automatic sports news generation. The second part summarizes the data sets related to sports video action recognition. The third part reviews the implementation methods of action recognition in sports video in recent years, summarizes them as traditional manual feature-based algorithms and deep learning-based algorithms, and the deep learning-based algorithms are categorized as 2D model-based, 3D model-based, two-stream/multi-stream model-based, and Transformer model-based, and summarizes the advantages and disadvantages of each model. The final part discusses the difficulties and challenges of sports video action recognition.
Key words: action recognition; deep learning; sports

0 引 言

当前，随着 4G、5G 通信技术的发展，视频数据已经成为当下互联网传播信息的重要载体，视频动作识别成为了计算机视觉领域的热门研究方向。相对于图像分类方向，视频中的动作识别根据挑战性去识别视频中的动作信息，需要综合运用多个学科的交叉知识。特别是体育视频中的动作识别，由于体育视频中的动作具有时间上的高依赖性，在处理这类视频时，需要算法设计者更好地聚合动作空间维度和时间维度上的信息。同时，体育视频更多地出现在专业赛场上，因拍摄条件的不同，拍摄视角和拍摄现场的光线与物体的遮挡都将给动作识别带来一定的困难。另有研究指出，视频中包含的信息量远丰富于图像中的信息量，因此，如何消除视频中的冗余信息，捕获并利用视频的中重要信息，成为了基于视频的体育动作识别中的一个难点领域。

基于视频的动作识别研究综述近年来已经有一定进展[1-4]，但这些文献[1-4]对当前基于深度学习的视频中的动作识别算法进行了总结分析，但关注

基金项目：国家自然科学基金(61773083)；上海市浦江人才计划(2019PJD035)；上海市人工智能创新发展专项；上海市引进海外高层次人才工作专项；上海高校特聘教授（东方学者）计划。
作者简介：游义平(1997-)，男，硕士研究生，主要研究方向：动作识别；季云峰(1990-)，男，博士，讲师，主要研究方向：乒乓球机器人。Email：yipingyou919@163.com。
收稿日期：2022-11-16

一些通用人体动作识别数据集，如 UCF101、HMDB51 等。本文将对在体育视频数据集上做出评估的一些动作识别算法进行研究综述，同时，本文还列举了体育视频动作的应用与数据集。希望本文能对广大研究体育动作识别的科研人员有一定的启示作用。

1 应用

体育视频动作识别作为视频分析的主要研究热点之一，分析视频中出现的动作对理解体育运动十分重要，其应用领域也十分广泛，从评估运动员的表现到为用户量身定制的智能设备。大量的研究工作以体育运动数据集[5-13]为基础。学者们在这方面做了许多研究。

1.1 辅助训练

体育视频数据集中包含了大量比赛和训练的片段历史记录，是教练员和运动员分析和提取技战术的良好信息来源。视频动作识别作为一种分析运动员技战术的有效方法之一，可以提供一种直接的方法获取动作，而这些动作的组合与获胜的策略有良好的关联。因此，将动作识别应用在体育视频中，既可以指导运动员的训练，又可以帮助教练员制定训练与比赛计划。文献[14]提出了一种可以识别冰球运动员的姿势和行为的动作识别沙漏网络(ARNH)，这有助于教练评估球员的表现。文献[15]阐述的体育 AI 教练系统，可以根据视频序列提供个性化的运动训练体验。动作识别是人工智能在教练系统中支持复杂视觉信息提取和总结的关键步骤之一。

1.2 辅助判罚

体育比赛中偶因裁判误判引发双方争议，国内外赛事主办方和各运动团队纷纷借助人工智能技术来提升比赛判罚的科学性。文献[16]提出了一个虚拟参考网络来评估跳水动作的执行情况。这种方法是基于视觉线索以及序列中的身体动作。同样对于跳水运动，文献[17]提出了一个可学习时间-空间特征的模型，用来评估相关运动，从而提高动作评估的准确性。文献[18]提出了一个体育裁判员培训系统，该系统采用了一个深度信念网络来获取高质量的手势动作，以此来判断裁判员是否发出了正确的裁判信号。

1.3 精彩动作集锦

体育视频中的精彩动作分割和总结受到体育爱好者的追捧，同时拥有着巨大的市场前景。完成精彩动作集锦的基础就是依靠动作识别技术处理好各种高光动作。文献[19]提出了一种自动高光检测方法来识别花样滑冰视频中的时空姿态。该方法能够定位和拼接花样滑冰动作。花样滑冰中的跳跃动作作为最吸引人的基本内容之一，常出现在精彩动作集锦之中。

文献[20]的主要工作是识别三维跳跃动作和恢复视觉效果不佳的动作。文献[21]将视频亮点看作是一个组合优化问题，并将识别动作的多样性作为约束条件之一。这项工作在一定程度上提高了多样性动作识别的准确性，精彩动作集锦的质量有了极大的改善。

1.4 体育新闻自动生成

体育比赛直播中的新闻信息以比赛中的实况数据为信息源，通过网络平台传播向广大体育粉丝及时转播比赛实况。现有的体育新闻系统通常采用比赛中的统计数字，如足球比赛中的射门数、角球数和任意球数，然后用文字来描述这些信息[22-23]，但大多数情况下这些文字还是依靠体育新闻记者人工撰写，既耗时、还费力。而应用视频动作识别和文字描述图像[24-28]技术，可以直接从视频中生成文字描述，进而自动生成专业的体育新闻。但想要提升自动生成的新闻的质量，仍需对运动员的动作进行更好的识别，而更优的识别结果，可以给自动生成的新闻带来更好流畅性和准确性。

2 体育动作识别相关数据集

在体育视频动作识别研究领域，基于视频预处理和网络结构的改进方法越来越多，但是不同的网络框架也需要一个共同的数据集来衡量性能的优劣。目前体育视频动作识别领域还缺少共同的数据集，本文将会总结体育视频动作识别存在的数据集，供后续研究人员参考。

2.1 乒乓球运动相关数据集

TTStroke-21[29]由 129 个自我录制视频段组成，每段视频采用 120 帧相机录制，视频总时长为 94 h。该数据集的标注工作由法国波尔多大学体育学院的相关专家与学生完成。该数据集共划分了发球反手旋、反手拦网、正手推挡、正手回环等 21 类专业乒乓击球动作，并可应用于乒乓球击球动作识别的综合研究中。需要说明的是，由于此数据集尚未完成对被录制者的隐私保护，从事相关研究的工作者只能从法国波尔多大学处获得部分完成隐私标注的数据集。

文献[30]中的数据集总共收集了22 111个视频片段，这些视频片段由 14 名职业乒乓球运动员做出的 11 种基本击球动作组成。

SPIN[31]提供了一个分辨率为 1 024×1 280、帧率为 150 帧/s 的视频数据集，视频总时长为 53 h，视频中每帧乒乓球的位置用边框标注，每个运动员的骨骼关节点也使用热图标记。该数据集可用在基于球的运动轨迹和球员姿态的跟踪、姿态估计和旋转预测等多项任务中。

OpenTTGames[17]视频采样帧率为 120 帧/s，该数据集包含了 38 752 个训练样本、9 502 个验证样本和 7 328 个测试样本，视频总时长为 5 小时，每个动作样本被标注为乒乓球击球动作、如正面击打。OpenTTGames 中的每个动作样本还对该动作发生前 4 帧、结束后 12 帧处运动员以及记分牌做了标注，故此数据集可用于语义分割、乒乓球的跟踪和击球动作的分类。

P^2A[32]数据集从世乒赛和奥运会乒乓球比赛的转播视频中收集了 2 721 个视频片段，视频总时长为 272 h。该数据集包含 14 类乒乓球击球动作类型。数据集的标注由职业乒乓球运动员和裁判员共同完成。同时对每一个动作样本的起始和结束时间做了精准的标注，该数据集用在动作定位和动作识别任务上。

P^2A 作为目前已知数据量最大、且标注最规范的数据集，将吸引更多研究者在乒乓球动作识别领域开发新的动作识别算法。

2.2　网球运动相关数据集

网球运动也是一项倍受欢迎的运动，吸引了众多学者进行研究。网球动作时间间隔短，而且密集，大多数动作的间隔不到 5 帧，对模型识别动作的快速性提出了很高的要求[33]。

文献[34]中为评估网球比赛中球员的动作制作了一个数据集，数据集来源于澳大利亚网球公开赛女子比赛。该数据集对球员的位置和动作起始与结束时间做了标注。主要将网球击球动作分类了 3 类：击球、非击球和发球。这是一个相对较小的数据集，且运动模糊性较高，是一个具有挑战性的数据集。

THETIS[13]由 8 374 段自录视频组成，包含了 55 位运动员做出的 12 类网球动作：4 类反手击球、4 类正手击球、3 类发球和扣杀球。视频总时长为 7h15 min，除了 RGB 视频外，THETIS 还提供了 1 980个深度视频、1 217 个 2D 骨架视频和 1 217 个 3D 骨架视频，因此可以用于开发多种类型的动作识别模型。

TENNISET[33]包含了超过 4 000 个动作样本，每个样本都采用了帧级别的标注。该数据集包含了 6 类网球动作：近右击球（Hit Near Right）、近左击球（Hit Near Left）、远右击球（Hit Far Right）、远左击球（Hit Far Left）、近发球（Serve Near）、远发球（Serve Far）和其他类。同时，该数据集还对击球动作标注了文本信息，如快速发球是亮点，这可拓展至视频新闻生成任务中。

2.3　足球运动相关数据集

ISSIA[10]为研究足球运动员的检测与跟踪而提出的数据集，数据集由覆盖整个足球场的 6 台分辨率为 1 920×1 080 、帧率为 25 帧/s 的摄像机录制，该数据集共标注了 18 000 帧，是一个小型足球运动数据集。由于足球运动中共有 22 名球员和 3 名裁判员，因此，制作此数据集面临着需标记多个目标的情况，给数据集标签的制作带来了不小的挑战。

Soccer[35]由原始转播视频中挑选精彩时刻的片段组成，该数据集是从 2 019 张图像中手动注释了 22 586个玩家位置。数据集由转播视频组成，因此包含了许多挑战，如不同的玩家外观、姿势、缩放级别、运动模糊、严重的遮挡和杂乱的背景。球员的身高、球员的图像位置和每张图像的球员数量分布广泛，显示了数据集的多样性。例如，玩家的身高从大约 20 像素到 250 像素，并从 150 像素的高度开始有一个长尾分布。

文献[36]中提出的数据集由 14 台摄像机拍摄而成，包含 599 个动作样本，共 132 603 帧。该数据集中，每个球员的位置都使用边界框标注了，该文献将足球运动动作分为了 5 类：传球、运球、射门、解围、无球权犯规。

ITS[37]由 222 个足球转播比赛视频组成，共计 170 个小时。该数据集包含 3 种标注类型：使用边界框标注球员的位置、粗粒度的动作发生与结束时间、细粒度的动作类型。共 11 类粗粒度动作发生与结束时间、15 类细粒度的动作类型。因此，该数据集可用于足球视频分析中的多种任务类型，如动作类型分类、动作定位与球员目标检测。

SoccerNet[38]数据集由来自欧洲 6 个主要联赛的 500 场完整足球比赛组成，涵盖 2014 年至 2017 年三个赛季，总时长 764 h。该数据集主要对以下 3 种主要事件（进球、黄牌/红牌和换人）的发生与结束时间进行了标注，同时该数据集中平均每 6.9 min

出现一个事件。该数据集主要解决长视频中稀疏事件的本地化问题,但关注的动作类型较少,使得任务过于简单。SoccerNet-V2[39] 在 SoccerNet 的基础上进行了拓展,将动作定位从 3 类拓展到 17 类;加入了对相机镜头的时间分割和相机镜头边界检测;重新定义了精彩动作回放任务;这项工作发布了一个足球动作识别基准任务,进一步推动了该领域的研究。

Footballer[40] 是为研究足球运动员的身份重识别与检测而提出的数据集,该数据集包含了 32 支欧洲冠军联赛球员在主场比赛中的 320 名球员、6 800 张图像,该数据集除了标注身份标签以外,还标注了 62 种属性标签信息。

2.4 篮球运动相关数据集

Basket-APIDIS[8] 由 7 台放置在球场周围的摄像机拍摄,但采取了非同步拍摄的方式,球拍摄场地照明条件不佳,导致此数据集是一个非常具有挑战性的数据集。

Basket-1[41] 和 Basket-2[41] 是分别包括一个 4 000帧和一个 3 000 帧的篮球序列。这些视频序列分别由 6 台和 7 台放置在球场周围的摄像机以 25 帧/s的速度同步拍摄。本文研究中对 Basket-1 的每一个第 10 帧和 Basket-2 的 500 个连续帧进行了手工注释,数据集中不仅将篮球动作划分为以下 4 类:扣篮、传球、持球和失球,同时还对篮球的位置进行了标注。

NCAA Basketball Dataset 由 257 个视频长度为 1.5 h 以内的未经修剪的 NCAA 比赛视频组成,经过标注后,该数据集共有 14 548 个动作边界的视频片段。此数据集将篮球动作划分为 3 分球投中、3 分球失败、2 分球投中、2 分球失败、上篮成功、上篮失败、罚篮成功、罚篮失败、灌篮成功、灌篮失败、抢球。此外,NCAA 还提供了共计 9 000 帧球员位置的标注。此项数据集也可拓展至球员位置检测。

2.5 多种类运动相关数据集

UCF Sports[7] 由 150 个分辨率为 720×480 的视频组成,该数据集共包含以下 10 个类别的运动视频:潜水运动(共 14 个视频)、高尔夫运动(共 18 个视频)、足球运动(共 6 个视频)、举重运动(共 6 个视频)、骑马运动(共 12 个视频)、跑步运动(共 13 个视频)、滑板运动(共 12 个视频)、跳马运动(共 13 个视频)、鞍马运动(共 20 个视频)、步行(共 22 个视频)。视频时长为 2.2~14.4 s 不等。与前文相比,该视频数据集较小,且对动作的分类程度较为粗糙。

Olympic Sports[42] 数据集共包含以下 16 类,每类由 50 个视频组成:跳高、跳远、三级跳远、撑杆跳、铁饼投掷、锤子投掷、标枪投掷、铅球、篮球架、保龄球、网球发球、跳台(跳水)、跳板(跳水)、抓举(举重)、挺举(举重)和跳马(体操)。因该数据集是从 YouTube 上获得的奥运比赛转播,故包含严重的相机移动、压缩伪影等情况。该数据集对于动作识别的算法设计提出了巨大的挑战。

Sports-1M 数据集由 100 万个 YouTube 视频组成,共包含 487 类,每个类别都包含 1 000~3 000 个视频。该数据集对类别标签进行了分层设计,父节点采用团体运动、球类运动等粗标签,叶子节点采用如台球的八球、九球等细粒度标签。Sports-1M 为体育运动动作识别任务,提供了一个大型数据集,吸引着更多的学者在这项数据集上进行算法模型的设计。

3 方法部分

目前,基于视频的体育动作识别算法经历了从基于传统的手工特征的算法到基于深度学习方法的转变。其中,基于传统的手工特征算法会涉及到研究人员对各特征的理解程度,直接设计含有物理意义的特征提取器,此设计思想对特征针对性强,但容易忽视数据中的隐含信息,同时对研究人员也提出了较高的领域知识要求;基于深度学习的方法能够很好地解决基于传统方法的不足,但基于深度学习的方法的数学可解释性相对于基于传统的手工特征的稍差。目前来说,基于深度学习的方法在相关的数据集上取得了比基于传统的手工特征更高的准确率。

本部分将回顾基于传统的动作识别算法和基于深度学习的动作识别算法。

3.1 基于传统手工特征的动作识别算法

基于传统方法的动作识别算法中的运动特征是人工提取的,在此基础上建立起表示人体动作的算法模型。

全局特征信息(GIST)[43] 和方向梯度直方图(Histogram of Oriented Gradients, HOGS)[44] 是手工运动特征提取中常采用的方式。采用 HOGS 方式提取视频中每一帧的运动特征,而后在时间上对帧特征进行平均来分类。

文献[45]在 UCF Sports 上对以上 2 种特征提取方式进行了评估,结果表示使用 GIST 特征比使用

HOGS 特征能取得更好的表现（GIST 60.0% vs. HOGS 58.6%）。一种可能的原因是，GIST 特征更容易将运动发生的背景与运动本身相关联，如足球运动通常发生在草坪上。

文献[46]使用 HOG3D 取代 HOG2D 提取视频动作特征，采用多层感知器（Multi Layer Preception，MLP）对动作类型进行分类。文献[34]采用 HOG3D 特征和核化费舍尔判别分析（Kernelized Fisher Discriminant Analysis，KFDA）对网球运动视频进行分析，并在文献[34]提出的自建数据集上取得了 84.5%的准确率。

虽然使用 HOG、HOF 和 SIFT 等提取的时空特征在 UCF Sports 和 Olympic Sports 等运动视频数据集上可以取得相对较好的成绩，但使用这些手工制作特征的方式总体上来说时间花销巨大。此外，由于传统的动作识别模型，特征提取模块和分类器是分开学习的，由此导致了这些模型都不能以端到端的模式训练。综上所述，学者们开始将目光转向基于深度学习的模式，并提出了许多新的方法将动作的准确率提升到了一个新水平。

3.2　基于深度学习的动作识别算法

当前主流的动作识别模型都是以深度学习为基础的，与传统方法相比，基于深度学习的模型能够以端到端的方式进行训练，这给应用深度学习模型带来了良好的实施可行性。

本次研究将对以下 4 种类型的深度学习模型进行归纳总结：基于 2D 模型、基于 3D 模型、基于双流/多流模型。

3.2.1　基于 2D 模型

2D 模型使用 2 维卷积神经网络（Convolutional Neural Networks，CNN）对视频的每一帧做特征提取，再将提取到的特征进行融合，并对融合结果进行预测。文献[47]将 CNN 网络引入了视频动作识别领域，进一步提出了 4 种特征融合方式：

（1）单帧融合：使用一个权重共享的 CNN 网络对视频中的每一帧进行特征提取，并将最后的特征串联起来进行分类。

（2）早期融合：使用一个大小为 11×11×3× T 的 3D 卷积核结合整个时间窗口内的帧信息进行融合。

（3）晚期融合：使用一个权重共享的 CNN 网络对相隔 15 帧的 2 个独立帧之间进行特征提取，并使用一个全连接层来融合单帧的特征表示。

（4）缓慢融合：在第一层实现一个 3D 卷积核，并在网络的更深层缓慢融合帧之间信息。

实验表明，缓慢融合优于其他融合方法，例如，缓慢融合在 Sports 1M [47] 上取得 60.9% 的准确率，而单帧融合、早期融合和晚期融合的准确率分别为 59.3%、57.7% 和 59.3%。但使用 HOG 等手工制作的特征只能达到 55.3% 的准确率，由此远低于使用 CNN 的准确率，这表明基于深度学习的模型可用于体育视频动作识别，并取得较好的效果，这些结果有助于推动后续团队在动作识别领域探索研究更多的深度学习模型。

另一种做法是直接使用长短时记忆（Long Short Term Memory，LSTM）网络[48]来获取动作时间上的联系。文献[49]提出了结合二维 CNN 和 LSTM 的模型，该模型首先使用一个权重共享的二维 CNN 来获取视频帧的空间上的特征信息，然后使用多层 LSTM 网络获取动作时间上的特征信息。在此基础上，文献[50]提出了一种使用两层 LSTM 网络的长期递归卷积网络（Long Tern Recurrent Convolutional Networks，LRCN）。文献[51] 采用基于 LSTM 的自动编码器以无监督方式来学习更好的视频表示。文献[52]提出了一个与文献[49] 中的模型相似的超前神经网络（Lead Exceed Neural Network，LENN），但 LENN 使用网络图像来微调前导网络，以过滤掉不相关的视频帧。

以上学者的研究表明，时间上的动作特征信息在动作识别模型中起着无可替代的作用。

文献[53]提出了由空间 CNN 网络和时间 CNN 网络组成的时间段网络（Temporal Segment Network，TSN），TSN 首先将一个输入视频切分成若干片段，并从这些片段中随机采样由 RGB 帧、光流和 RGB 差值组成的短片段。然后，这些片段被送入空间和时间网络进行预测。接下来，该网络通过聚合各片段的预测分数来获得最终的预测结果。TSN 以 2 种方式获得时间信息：

（1）直接将光流引入框架。

（2）类似于前文提到的晚期融合，TSN 聚合了片段预测的结果。

最后，仅使用 RGB 帧的二维 TSN 获得了令人印象深刻的效果，在 FineGym[54] 上的结果为 61.4%。在通用动作识别数据集 UCF101[55] 上的结果为 87.3%。TSN 的另一个变种 KTSN 不再使用随机采样，而是使用关键视频帧，应用关键视频帧在 FSD-10 上取得了比 TSN 更好的效果[56]（63.3% vs. 59.3%）。

文献[57]提出时间关系网络（Temporal

Relational Network,TRN)以捕获帧之间的时间关系,并摒弃之前学者使用的简单聚合方法,如串联和线性组合,改而使用 MLP 计算这些关系,同时可以插入到任何现有框架中。TRN 在 FineGym[54] 的性能相比 TSN 显著提升,达到了 68.7%的准确率。

然而,在 TRN 中使用 MLPS 计算多帧时间关系时非常耗时,并且不能很好地捕捉有用的低级特征。为了解决这个问题,文献[58]提出了一种简单而有效的模块、即时间移位模块(Temporal Shift Module, TSM)来捕获时间信息,TSM 使用 2D CNNs 提取视频帧上的空间特征,并将 TSM 插入到 2D 卷积块中。TSM 在 FineGym[54] 上取得了 70.6%的准确率,优于 2D TSN、2D TRN 和 I3D[59] 等方法,而且计算复杂度较低。

3.2.2 基于 3D 模型

在二维 CNN 中,卷积应用于 2D 特征图,仅从空间维度计算特征。当利用视频数据分析问题的时候,研究期望捕获多个连续帧编码的运动信息。为此,提出在 CNN 的卷积进行 3D 卷积,以计算空间和时间维度特征, 3D 卷积是通过堆叠多个连续的帧组成一个立方体,并在立方体中运用 3D 卷积核。通过这种结构,卷积层中的特征图都会与上一层中的多个相邻帧相连,从而捕获运动信息。

二维 CNN 中将视频中的图像解码为多个视频帧,并用 CNN 来识别单帧的动作。但这种方法没有考虑多个连续帧中编码的运动信息。为了有效地结合视频中的运动信息,文献[60] 提出可以在 CNN 卷积层中使用 3D 卷积,以捕获动作沿空间和时间维度的特征。该文献中的网络结构由 1 个硬连线层、2 个三维卷积层、2 个子采样层、1 个二维卷积层和 1 个全连接层组成。尽管文献[60]所提出的网络相对较小,也只在小型数据集上进行了评估,但这项工作中的 3D CNN 结构可以从相邻的视频帧生成多个信息通道,并在每个通道中分别执行卷积和下采样,通过将来自视频通道的信息组合获得最终特征表示,取得了比二维 CNNs 更好的性能。文献[56]动作识别中采用 3D CNN 的开创性工作,引领更多学者将 3DCNN 结构应用于动作识别领域。

文献[61]为大型视频动作识别数据集设计了一个深度的三维体系结构(Convolutioal 3D ,C3D),C3D 模型中的三维卷积层为 8 层,每层中的 3D 卷积核大小为 3×3×3。C3D 在 Sports 1M 数据集上取得了 61.1%的准确率。文献[62]使用 C3D 模型,但做了一些改进使得网络层数更浅,在 UCF50 数据集上取得了 97.6%的精度。文献[59]提出了一个新的模型 Two stream Inflated 3D ConvNet(I3D),该模型在动作识别任务上取得了一个新的突破。与 C3D 相比,I3D 网络层次要深得多,其中堆叠了 9 个 3D 初始模块[63] 和 4 个独立的 3D 卷积层。I3D 将 Inception-V1[64] 中大小为 $N \times N$ 的 2D 卷积核扩展为 $N \times N \times N$ 的 3D 卷积核,并且 3D 卷积核的参数也是由预先训练好的 2D 卷积核通过引导得到的。I3D 网络结合了 RGB-3D 网络和 Flow-3D 网络,并且 I3D 网络在比 UCF101 数据集多 400 类的 Kinetics-400 数据集上进行预训练,将预训练的数据进行微调后在 UCF101 数据集上取得了 97.9%的准确率,在 Kinetics-400 数据集上取得了 74.2%的准确率。前述研究工作证明了在视频动作识别任务中,在更大规模的数据集上进行预训练,迁移到较小规模数据集上,做一些参数上的微调,能够取得非常不错的成绩。

直接将大小为 $N \times N$ 的二维卷积核扩展为大小为 $N \times N \times N$ 的三维卷积核可以使网络中可学习的参数量显著增加,并提高模型的容量,但这也会导致计算复杂度的增加,存在过拟合的风险。为了缓解这个问题,文献[65]提出一个伪 3D(Pseudo 3D, P3D)网络,其中 3D 卷积被叠加的 2D 卷积和 1D 卷积所代替。同样,文献[66]研究了不同的体系结构(2D、3D 和(2+1)D),发现将卷积核大小为 $1 \times N \times N$ 的 2D 卷积与卷积和大小为 $T \times 1 \times 1$ 的 1D 卷积核叠加起来,所取得的性能优于其他体系结构。而 S3D[67] 则又将 I3D 中的部分 3D 启动模块替换为 2D 启动模块,以平衡性能和计算复杂度。之后,文献[68]提出了一组称为三维信道分离网络(Channel Separated Networks, CSN),该网络为进一步减少浮点数计算(Floating Point Operations, *FLOPs*), CSN 模型探讨了群卷积、深度卷积和这些方法的不同组合。结果表明,CSN 不但性能比 3D CNNs 好得多,且 *FLOPs* 只有 3D CNNs 的三分之一。

然而,将卷积核从 2D 扩展到 3D 必然会使计算成本增加一个数量级,限制了其实际应用。文献[69]提出了一种简单而有效的方法 STM(SpatioTemporal and Motion Encoding)网络,可将时空和运动特征集成到一个统一的二维 CNN 框架中,无需任何三维卷积计算。

STM[69] 采用 2 个模块-通道时空模块(Channel-wise Spatial Temporal Module, CSTM)和通道运动模块(Channel-wise Motion Module, CMM),其中 CSTM

采用(2+1)D 卷积融合空间和时间特征，而 CMM 只采用二维卷积，但将连续三帧的特征拼接起来。与 P3D[65] 和 R3D[66] 相比，STM 表现更好。

C3D 及其改进模型将 2D 卷积扩展到时空域，默认时域和空域是平等的、对称的，同时处理空域和时域的信息，而 SlowFast[70] 将空域和时域进行拆分处理，也更为符合时域和空域特征的关系。

SlowFast[70] 由 2 个分支组成。一个是低帧率的慢分支，另一个是高帧率的快分支。低帧率的慢分支在底层只使用 2D 卷积，在顶层使用(1+2)D 卷积可以更多地关注空间语义信息，采样率低的慢分支提取随时间变化较慢的空间特征，而快分支在每一层都使用(1+2)D 卷积更多地关注对象运动信息。FAST 分支提取随时间变化较快的运动特征，为了降低该通道的复杂度，卷积核的空间通道数设计得较小，从而使网络变得轻量级的同时还可以学习用于视频动作识别的有用时间信息。

相比于 C3D 及其改进模型，SlowFast 中同样用到了 3D 卷积，但与 C3D 的又不太相同。Slow 通路在底层使用 2D 卷积，顶层使用(1+2)D 卷积(实验发现比全用 3D 卷积效果更好)；Fast 通路每一层用的都是(1+2)D 卷积，但是各层维持时域维度大小不变，尽可能地保留时域信息，而 C3D 中越深的层时域维度越小。此外，SlowFast 将慢速和快速特性横向拼接融合在一起。通过对慢分支、快分支和横向连接的精心设计，SlowFast 在多种流行的动作识别数据集上实现了最先进的性能。

用于视频动作识别的神经网络很大程度上是通过将 2D 图像架构[64, 71-73] 中的网络输入、特征或卷积核扩展到时空维度来驱动的[47, 59, 74-75]；虽然沿时间轴扩展(同时保持其他设计属性)通常会提高准确度，但如果在计算复杂度和准确度之间做一个权衡，这些操作可能不是最优的。

X3D[76] 从空间、时间、深度和宽度四个方面对二维 CNNS 进行了扩展，探索了多种体系结构，发现高时空网络优于其他模型。在 Kinetics-400 上，X3D 比 SlowFast 表现稍差：前者 79.1%、后者79.8%，但 X3D 的参数较少，且训练和推理时间较短。为了进一步减少网络参数和 *FLOPs* 的数量，文献[77]提出能够处理流式视频的移动视频网络(Mobile Video Networks，Movinets)。Movinets 中应用了 2 个核心技术。第一个是神经结构搜索(Neural Architecture Search，NAS)[78]，用于高效地生成 3DCNN 结构；第二个是流缓冲技术，将内存与视频剪辑持续时间解耦，允许 3DCNNs 以较小的恒定内存占用嵌入任意长度的视频流用于训练和推理。使用这 2 种技术，Movinets 只需要 X3D 的 20% 的 *Flops*，就获得了相同的性能。

SlowFast[70] 表明引入不同的时间分辨率有利于动作识别，然而是将一个单独的网络应用于每个分辨率。以上提到的动作识别网络的设计中往往忽略了表征不同动作的一个重要方面：动作本身的视觉节奏。视觉节奏实际上描述了一个动作进行的速度，往往决定了识别的时间尺度上的有效持续时间。在某些情况下，区分不同动作类别的关键是各动作的视觉节奏，比如走路、慢跑和跑步视觉外观上有着高度相似之处，但视觉节奏存在明显不同。时间金字塔网络(Temporal Pyramid Network，TPN)[79] 采用一个主干网，对不同层次的三维特征采用时间金字塔，即低帧率用于捕捉高级特征语义，高帧率用于捕捉低级运动特征信息。TPN 在 Kinetics-400 上实现了 SlowFast 相同的性能，但只采用了一个网络分支。

为了对长视频序列进行建模，文献[80] 将时态全连通操作引入到 SlowFast 中，提出了 TFCNet，文中时间全连接块(TFC Block)是一种高效的组件，可沿时间维度将所有帧的特征通过一个 FC 层组合在一起以获得视频级的感受野，增强时空推理能力。通过将 TFC 块插入到 SlowFast，在真实世界静态无偏数据集 Diving48 上，比 SlowFast 提高了近 11%，性能提高到 88.3%，同时超越了所有以前的方法。

相比于采用 2D 结构的模型，通常采用 3D 结构模型的精度更高，相比于 2D 模型的需要计算的参数量也有了明显的增长。对 GPU 等硬件提出了更高的要求。

3.2.3　基于双流/多流模型

文献[81]首次提出了双流卷积神经网络(Two Stream Convolutional Network)，该模型具有一个空间流卷积神经网络(Spatial Stream ConvNet)分支和一个时间流卷积网络(Temporal Stream ConvNet)分支。以 RGB 图像和相应的光流作为 2 个分支卷积神经网络的输入，分别提取空间特征和时间特征。特征的融合在网络的最后使用支持向量机(Support Vector Machine，SVM)进行分类。研究中提出的双流网络在 UCF101 数据集上取得了 88%的准确率，识别效果优于使用单独的空间流或时间流卷积神经网络。但文献[81]提出的双流网络结构中计算光流所需的计算量大，计算时间较长，这不利于实时视

频行为识别。受此启示,文献[82]通过将光流替换为直接从压缩视频获得的运动矢量应用于实时动作分类中并取得了不错的成绩,但运动矢量缺乏精细的结构,导致了识别性能的下降。

文献[83]受文献[81]在堆叠光流和图像帧上训练的双流卷积神经网络能成功应用于基于视频的动作识别的启发,也以类似的方式考虑了时间维度上的数据。提出了多流网络(Multi Stream Network, MSN)[83]。MSN 是由 2 个双流网络组成的多流卷积神经网络,每个网络由不同的 VGG 网络组成,输入到网络中的是由原始视频拆分而得到的一系列连续 6 帧 RGB 图像,并计算求得其光流(Optical Flow OF)和以人的边界为感兴趣区域(Region of Interest, ROI)。这种多流网络会反馈给全连接层,全连接层向自身馈送给双向长短时记忆网络(Long-Short Term Menory,LSTM)。LSTM 网络的输入来自 MSN 网络的连续输出。这项工作使用像素轨迹而不是堆叠的光流作为运动流的输入,从而显著改善了识别结果。

视频由一系列静态图像组成,此前的工作均是采用静态图像及其计算出的光流输入网络中,但对于视频的最佳表现方式还不是很清楚。文献[84]提出了一种使用顺序池化(Rank Pooling)对 RGB 图像或光流视频等时态数据进行编码得到的动态图像。使用动态图像作为 ResNeXt-50 和 ResNeXt-101 网络输入。研究可知,在 UCF101 数据集上分别达到了 95.4%和 96%的成绩。

人的视觉系统是直观的,不以光流信息作为输入信号,而是以眼睛所看到直观信息来判断运动的种类。文献[85]提出了 ActionFlowNet 模型。这是一种高效的数据表示学习方法,用于学习只有少量标记数据的视频表示。ActionFlowNet 模型直接从原始像素训练单个流网络,用以共同估计光流,减小了计算光流的巨大耗时。与其他不使用预训练的方法相比,该方法在 UCF101 数据集上也取得了 83.9%的准确率。类似的工作还有,文献[86]提出 Motion-Augmented RGB Stream(MARS)。MARS 使用 3D ResNet 训练 RGB 流,以此模仿 OF 特征。作为单个流,MARS 的性能优于单独的 RGB 流或光流。

文献[87]对双流卷积网络的输入、网络结构和训练策略进行了思考,提出了时间段网络(Temporal Segment Networks,TSN),优化了文献[81]提出的双流网络,在 UCF101 数据集上取得了 94.2%的成绩。

3.2.4 基于 Transformer 模型

得益于 Transformer[88]在自然语言处理(Natural Language Processing,NLP)领域取得的巨大成功,文献[89]并未选用 CNN,直接按照 BERT 的模型结构使用了纯 Transformer 的结构提出了 VIT 模型,并在图片分类任务上取得了巨大的成功,实现了计算机视觉(Computer Vision,CV)与 NLP 的融合统一,使得在 NLP 领域成功的模型能迁移到 CV 领域,促进了 CV 领域的发展。由于 Transformer 强大的序列建模能力,CV 领域主流的骨干网络逐渐从 CNN 转为了 Transformer,文献[90]提出了 VTN(Video Transformer Network)模型,该模型摒弃了 3D CNN 的视频动作识别标准方法,引入了一种通过关注整个视频序列信息来对动作进行分类的方法。此模型以给定 8 帧图片为输入,后接一个时间注意力的编码层,获取时空特征。在运行时间方面,与其他方法相比,VTN 方法在推理时间上快了 16.1 倍,运行速度提高了 5.1 倍,同时在 Kinetics-400 数据集上取得了 94.2%的准确率。文献[91]提出了 VidTr 模型,与常用的 3D CNN 相比,VidTr 能够通过堆叠注意力层聚合时空信息,并以更高的效率提供更好的性能。VidTr 在 5 个常用数据集以较低的计算,实现了先进的性能,这项工作证明 VidTr 更为擅长推理长时间序列的行为。

在多项动作识别数据集上,基于 Transformer 的模型取得了最先进的性能,但也存在着许多有待解决的问题。

(1)特征提取问题。Transformer 具有强大的序列建模能力,在 NLP 领域中,特征序列是一维线性排列的,而在视频领域中,图像像素之间的联系是三维的。与 CNN 网络中利用卷积核来获取特征的方式不同,基于 Transformer 的模型目前只能捕捉一维序列中的特征,如何有效地提取视觉特征还需要进一步的研究与拓展。

(2)输入特征冗余问题。基于 Transformer 的模型将输入视频编码为多个 Token 作为模型的输入,VIT 模型中一张 224×224 分辨的图片将产生 196 个视觉 Token,过长的 Token 量将大大增加模型的计算代价,将使模型的的高效训练与推理变得困难。

4 挑战和难点

虽然基于视频的动作识别算法在通用数据集上取得了很不错的成绩,但基于视频的体育动作识别还存在许多的挑战与难点。

4.1　数据集的制作与标注

作为进一步研究视频动作识别方法在体育动作识别的关键问题之一,体育视频数据的收集与标注的质量直接影响着动作识别算法的性能[59, 92-93]。然而,体育视频数据集在制作过程中与其他通用的视频动作识别数据集,如 UCF101、HMDB51、Knietic400 等存在着很大的区别。

(1)版权问题。大多数的体育竞赛视频来自于未经剪辑的直播片段,由于视频版权等原因,这些片段的收集可能会受到版权限制。

(2)自建数据。非专业运动员自制的体育视频可能存在动作质量较低、拍摄角度不佳等问题,在此基础上进行训练的模型的可泛化能力差。

(3)标注的专业性。体育动作识别通常关注特定的运动类别,如花样滑冰、乒乓球、排球等,这些动作相比日常行为如:喝水、跑跳等,需要参与标注的人员有相关的专业知识,且标注者的专业性能很大程度上会影响相关动作识别算法在此类任务上的推广。

4.2　算法应用

(1)密集性动作。流行的动作识别模型[58, 94-95]所研究的对象是每个动作发生的时间间隔为 20 s,或者更长的动作间隔时间。然而,一方面乒乓球比赛中的击球动作通常发生在 0.4 s 或者更短的时间间隔内。传统的低速摄影机难以从具有背景变化的视频中捕捉到更丰富的动作细节[96-97]。另一方面,在乒乓球运动中,运动员双方轮流击球,相比于足球、篮球等动作,击球动作呈现密集分布,这对动作识别算法的识别动作边界提出了更高的要求。当前,虽然有一些学者在这些方面做出了努力,但与常规动作识别任务相比,研究学者所提出的算法性能仍远远低于预期[98-99],这对现有模型来说仍是一项具有挑战性的任务[49, 51]。

(2)动作视角变化。视频动作数据集相比于图像数据集,运动的物体在时间上存在着强关联,目标物体的运动特征的提取质量将直接影响动作识别模型性能[100-102]。此前的一些模型是对由固定摄像机视角拍摄的视频采用光流法[103-104]对运动特征进行提取。然而,随着体育视频集锦的出现,越来越多的体育视频中的相机视角出现了变化,如对视频片段中的精彩动作进行放大。这对成熟的动作识别基准模型[53, 56, 58, 66, 81]提出了巨大的挑战,如文献[57, 105-106]所提出的算法,几乎不能处理动作视角剧烈变化的样本。虽然文献[107-109]考虑了动作视角的变化,但在设计运动描述子时,面对被遮挡和被剪切的动作时,仍然导致了特征空间不一致,使得模型没有达到理想的性能。文献[110-112]通过设计运动描述符的结构和添加注意力机制来解决遮挡问题,但这些工作中的运动描述符仅限于单个目标被遮挡的情况,对于多个被遮挡的对象,效果仍然欠佳。

(3)数据集长尾分布。长尾学习[113-114]是计算机视觉识别最具挑战性的问题之一。视频来源于体育赛事直播中的足球、篮球、乒乓球等比赛。由于类分布的长尾性和不均衡性,使得模型的性能大大降低[115-118]。而考虑到体育类动作的特殊性,对模型中的数据增强方法提出了更高的要求。

5　结束语

本文对最近几年的体育视频中的动作识别算法进行了较全面的综述。由于体育动作与时间上的强关联,在算法设计中引入时序信息,可以有效提升算法的准确性。当前的动作识别算法在各通用数据集上均取得了不错的成绩,但将算法应用在体育视频中的动作识别仍需学者进行更多的研究,特别是在缺乏丰富数据集的情况下,体育视频分析仍然是一项具有挑战性的任务。

参考文献

[1] 彭月, 甘臣权, 张祖凡. 人类动作识别的特征提取方法综述[J]. 计算机应用与软件, 2022, 39(8): 1-14,68.

[2] 朱相华, 智敏. 基于改进深度学习方法的人体动作识别综述[J]. 计算机应用研究, 2022, 39(2): 342-348.

[3] 吴宏俊, 李铭兴, 刘宏哲. 基于深度学习的视频动作识别的发展[C]// 中国计算机用户协会网络应用分会 2021 年第二十五届网络新技术与应用年会. 北京:中国计算机用户协会, 2021: 206-210.

[4] 卢修生, 姚鸿勋. 视频中动作识别任务综述[J]. 智能计算机与应用, 2020, 10(3): 406-411.

[5] BERMEJO N E, DENIZ S O, BUENO G G, et al. Violence detection in video using computer vision techniques [C]// Computer Analysis of Images and Patterns. Berlin/ Heidelberg: Springer, 2011: 332-339.

[6] PERS J. Cvbase 06 dataset: A dataset for development and testing of computer vision based methods in sport environments [DB]. SN, Ljubljana, 2005.

[7] RODRIGUEZ M D, AHMED J, SHAH M. Action MACH a spatio-temporal maximum average correlation height filter for action recognition [C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA: IEEE, 2008: 1-8.

[8] DE VLEESCHOUWER C, CHEN F, DELANNAY D, et al. Distributed video acquisition and annotation for sport-event summarization [J]. NEM Summit, 2008, 8:1010-1016.

[9] PARISOT P, DE VLEESCHOUWER C. Consensus - based trajectory estimation for ball detection in calibrated cameras systems [J]. Journal of Real - Time Image Processing, 2019, 16(5): 1335-1350.

[10] D'ORAZIO T, LEO M, MOSCA N, et al. A semi-automatic system for ground truth generation of soccer video sequences [EB/OL]. [2009]. https://www.xueshufan.com.publication/2144958063.

[11] LI Wanqing, ZHANG Zhengyou, LIU Zicheng. Action recognition based on a bag of 3D points [C]//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. San Francisco, CA: IEEE. 2010: 9-14.

[12] NIEBLES J C, CHEN C W, LI Feifei. Modeling temporal structure of decomposable motion segments for activity classification[M]// DANIILIDIS K, MARAGOS P, PARAGIOS N. Computer Vision-ECCV 2010. ECCV 2010. Lecture Notes in Computer Science.Berlin/Heidelberg: Springer, 2010,6312: 392-405.

[13] GOURGARI S, GOUDELIS G, KARPOUZIS K, et al. THETIS: Three dimensional tennis shots a human action dataset [C]// Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Portland, USA: IEEE, 2013: 676-681.

[14] FANI M, NEHER H, CLAUSI D A, et al. Hockey action recognition via integrated stacked hourglass network [C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, HI, USA: IEEE, 2017:85-93.

[15] WANG Jianbo, QIU Kai, PENG Houwen, et al. AI coach: Deep human pose estimation and analysis for personalized athletic training assistance [C]// Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019.

[16] NEKOUI M, TITO C F O, LI Cheng. FALCONS: Fast learner-grader for Contorted poses in sports [C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New Orleans: IEEE, 2020: 1-9.

[17] VOEIKOV R, FALALEEV N, BAIKULOV R. TTNet: Real-time temporal and spatial video analysis of table tennis [C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, WA, USA: IEEE. 2020: 3866-3874.

[18] PAN T Y, TSAI W L, CHANG C Y, et al. A hierarchical hand gesture recognition framework for sports referee training - based EMG and accelerometer sensors [J]. IEEE Transactions on Cybernetics, 2022, 52(5): 3172-3183.

[19] NAKANO T, SAKATA A, KISHIMOTO A. Estimating blink probability for highlight detection in figure skating videos [J]. arXiv preprint arXiv:2007.01089, 2020.

[20] TIAN Limao, CHENG Xina, HONDA M, et al. Multi - technology correction based 3D human pose estimation for jump analysis in figure skating [J]. Proceedings, 2020,49(1):95.

[21] SHROFF N, TURAGA P, CHELLAPPA R. Video précis: Highlighting diverse aspects of videos [J]. IEEE Transactions on Multimedia, 2010, 12(8): 853-68.

[22] KANERVA J, RöNNQVIST S, KEKKI R, et al. Template-free data-to-text generation of Finnish sports news [J]. arXiv preprint arXiv:1910.01863, 2019.

[23] GONG Junpeng, REN Wen, ZHANG Pengzhou. An automatic generation method of sports news based on knowledge rules [C]// Proceedings of 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). Shanghai: IEEE, 2017:1-4.

[24] WANG Qingzhong, WANG Jiuniu, CHAN A B, et al. Neighbours matter: Image captioning with similar images [EB/OL]. [2020]. https://www.xueshufan.com/publication/3127608482.

[25] WANG Qingzhong, WAN Jia, CHAN A B. On diversity in image captioning: Metrics and methods [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 1035-1049.

[26] WANG Jiuniu, XU Wenjia, WANG Qingzhong, et al. On distinctive image captioning via comparing and reweighting [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023,45:2088-2103.

[27] CHEN Shaoxiang, JIANG Yugang. Motion guided region message passing for video captioning [EB/OL]. [2021]. https://download.csdn.net/download/dwf1354046363/46169190? utm_source=bbsseo.

[28] ZHANG Ziqi, QI Zhongang, YUAN Chunfeng, et al. Open-book video captioning with retrieve-copy-generate network [J]. arXiv preprint arXiv:2103.05284,2021.

[29] MARTIN P E, BENOIS-PINEAU J, PETERI R, et al. Sport action recognition with siamese spatio - temporal CNNs: Application to table tennis [C]//Proceedings of 2018 International Conference on Content - Based Multimedia Indexing (CBMI). IEEE, 2018.

[30] KULKARNI K M, SHENOY S. Table tennis stroke recognition using two-dimensional human pose estimation [C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE. 2021:1-9.

[31] SCHWARCZ S, XU P, D'AMBROSIO D, et al. SPIN: A high speed, high resolution vision dataset for tracking and action recognition in ping pong [J]. arXiv preprint arXiv:191206640, 2019.

[32] BIAN J, WANG Q, XIONG H, et al. A dataset and benchmark for dense action detection from table tennis match broadcasting videos [J]. arXiv preprint arXiv:220712730, 2022.

[33] FAULKNER H, DICK A. TenniSet: A dataset for dense fine-grained event recognition, localisation and description [EB/OL]. [2017]. https://www.xueshufan.com/publication/2778404137.

[34] DE CAMPOS T, BARNARD M, MIKOLAJCZYK K, et al. An evaluation of bags-of-words and spatio-temporal shapes for action recognition [C]//2011 IEEE Workshop on Applications of Computer Vision (WACV). Kona,HI, USA:IEEE, 2011:1-9.

[35] LU Keyu, CHEN Jianhui, LITTLE J, et al. Light cascaded convolutional neural networks for accurate player detection [J]. arXiv preprint arXiv:1709.10230, 2017.

[36] TSUNODA T, KOMORI Y, MATSUGU M, et al. Football action recognition using hierarchical LSTM [C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, HI, USA: IEEE. 2017: 155-163.

[37] YU Junqing, LEI Aiping, SONG Zikai, et al. Comprehensive dataset of broadcast soccer videos [C]//Proceedings of 2018 IEEE

Conference on Multimedia Information Processing and Retrieval (MIPR). Miami, FL: IEEE, 2018:418-423.
[38] GIANCOLA S, AMINE M, DGHAILY T, et al. SoccerNet: A scalable dataset for action spotting in soccer videos [J]. arXiv preprint arXiv:1804.04527, 2018.
[39] DELIEGE A, CIOPPA A, GIANCOLA S, et al. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos [J]. arXiv preprint arXiv:2011.13367, 2021.
[40] 高蕾. 基于深度学习方法实现运动场景上行人重识别的研究 [D]. 武汉:华中科技大学, 2021.
[41] MAKSAI A, WANG Xinchao, FUA P. What players do with the ball: A physically constrained interaction modeling [C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016:972-981.
[42] NIEBLES J C, CHEN C-W, FEI-FEI L. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification [C]//Computer Vision - ECCV 2010. Berlin/ Heidelberg:Springer, 2010: 392-405.
[43] OLIVA A, TORRALBA A. Modeling the shape of the scene: A holistic representation of the spatial envelope [J]. International Journal of Computer Vision, 2001, 42(3): 145-175.
[44] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego:IEEE,2005:886-893.
[45] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: A large video database for human motion recognition [C]//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011: 2556-2563.
[46] IJJINA E P. Action recognition in sports videos using stacked auto encoder and HOG3D features [C]//Proceedings of the Third International Conference on Computational Intelligence and Informatics. Singapore:Springer, 2020: 849-56.
[47] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with Convolutional Neural Networks [C]// Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Ohio, USA: IEEE, 2014: 1725-1732.
[48] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
[49] JOE Y H N, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: Deep networks for video classification [C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE. 2015: 4694-4702.
[50] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description [J]. arXiv preprint arXiv:1411.4389 ,2014.
[51] SRIVASTAVA N, MANSIMOV E, SALAKHUDINOV R. Unsupervised learning of video representations using LSTMs[C]// Proceedings of the International Conference on Machine Learning. New York,USA:ACM,2015:843-852.
[52] GAN Chuang, YAO Ting, YANG Kuiyuan, et al. You lead, we exceed: Labor-free video concept learning by jointly exploiting Web videos and images [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA:IEEE, 2016: 923-932.
[53] WANG Limin, XIONG Yuanjun, QIAO Zhe, et al. Temporal segment networks for action recognition in videos [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2740-2755.
[54] SHAO Dian, ZHAO Yue, DAI Bo, et al. FineGym: A hierarchical video dataset for fine-grained action understanding [J]. arXiv preprint arXiv:2004.06704, 2020.
[55] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild [J]. arXiv preprint arXiv:1212.0402, 2012.
[56] LIU Shenglan, LIU Xiang, HUANG Gao, et al. FSD-10: A fine-grained classification dataset for figure skating [J]. Neurocomputing, 2020, 413:360-367.
[57] ZHOU Bolei, ANDONIAN A, OLIVA A, et al. Temporal relational reasoning in videos [C]// Computer Vision-ECCV 2018. Munich, Germany:Springer International Publishing. 2018: 831-846.
[58] LIN Ji, GAN Chuang, HAN Song. TSM: Temporal shift module for efficient video understanding [J]. arXiv preprint arXiv:1811.08383, 2019.
[59] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[J]. arXiv preprint arXiv:1705.07750, 2017.
[60] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-31.
[61] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015: 4489-4497.
[62] LIMA T, FERNANDES B, BARROS P. Human action recognition with 3D convolutional neural network [EB/OL]. [2017-11]. https://readpaper.com/paper/2785529808.
[63] POUYANFAR S, CHEN S C, SHYU M L. An efficient deep residual-inception network for multimedia classification [C]// Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME). Hong Kong, China:IEEE, 2017: 373-378.
[64] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Xi′an: IEEE, 2015: 1-9.
[65] QIU Zhaofan, YAO Ting, MEI Tao. Learning spatio-temporal representation with pseudo-3D residual networks [C]// Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Los Angeles, California, USA:IEEE, 2017: 5533-5541.
[66] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition [C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE,2018: 6450-6459.
[67] XIE Saining, SUN Chen, HUANG J, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification [C]//Computer Vision - ECCV 2018. Munich, Germany: Springer International Publishing, 2018: 318-335.
[68] TRAN D, WANG Heng, FEISZLI M, et al. Video classification with channel-separated convolutional networks [C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision

(ICCV). Seoul, South Korea:IEEE, 2019: 5552-5561.
[69] JIANG Boyuan, WANG Mengmeng, GAN Weihao, et al. STM: Spatio temporal and motion encoding for action recognition [C]// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Republic of Korea: IEEE, 2019: 2000-2009.
[70] FEICHTENHOFER C, FAN H, MALIK J, et al. SlowFast networks for video recognition [J]. arXiv preprint arXiv: 1812.03982v2, 2019.
[71] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Xi'an: IEEE, 2015: 770-778.
[72] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.
[73] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:1409.1556, 2014.
[74] DONAHUE J, ANNE HENDRICKS L, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description [J]. arXiv preprint arXiv:1411.4389, 2014.
[75] CHRISTOPH R, PINZ F A. Spatiotemporal residual networks for video action recognition [C]// Advances in Neural Information Processing Systems. Barcelona, Spain: NIPS Foundation, 2016: 3468-3476.
[76] FEICHTENHOFER C. X3D: Expanding architectures for efficient video recognition [C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA: IEEE, 2020: 203-213.
[77] KONDRATYUK D, YUAN Liangzhe, LI Yandong, et al. MoViNets: Mobile video networks for efficient video recognition [J]. arXiv preprint arXiv:2103.11511, 2021.
[78] BENDER G, LIU Hanxiao, CHEN Bo, et al. Can weight sharing outperform random architecture search? An investigation with TuNAS [C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA: IEEE, 2020: 14311-14320.
[79] YANG Ceyuan, XU Yinghao, SHI Jianping, et al. Temporal pyramid network for action recognition [C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington, USA: IEEE, 2020: 591-600.
[80] ZHANG Shiwen. Tfcnet: Temporal fully connected networks for static unbiased temporal reasoning [J]. arXiv preprint arXiv:2203.05928, 2022.
[81] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos [J]. Advances in Neural Information Processing Systems, 2014, 27: 568-576.
[82] ZHANG Bowen, WANG Limin, WANG Zhe, et al. Real-time action recognition with enhanced motion vector CNNs [C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 2718-2726.
[83] SINGH B, MARKS T K, JONES M, et al. A multi-stream bi-directional recurrent neural network for fine-grained action detection [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 1961-1970.
[84] BILEN H, FERNANDO B, GAVVES E, et al. Action recognition with dynamic image networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 2799-2813.
[85] NG J Y H, CHOI J, NEUMANN J, et al. Actionflownet: Learning motion representation for action recognition [C]// Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, NV, USA: IEEE, 2018: 1616-1624.
[86] CRASTO N, WEINZAEPFEL P, ALAHARI K, et al. MARS: Motion-augmented RGB stream for action recognition [C]// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 7874-7883.
[87] WANG Limin, XIONG Yuanjun, WANG Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition [C]//Computer Vision-ECCV 2016. Amsterdam, Netherlands: Springer International Publishing, 2016: 20-36.
[88] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.
[89] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020.
[90] NEIMARK D, BAR O, ZOHAR M, et al. Video transformer network [C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Los Alamitos: IEEE, 2021:3163-3172.
[91] LI Xinyu, ZHANG Yanyi, LIU Chunhui, et al. Vidtr: Video transformer without convolutions [J]. arXiv preprint arXiv:2104.11746v1, 2021.104.
[92] ZHANG Yu, CHENG Li, WU Jianxin, et al. Action recognition in still images with minimum annotation efforts [J]. IEEE Transactions on Image Processing, 2016, 25(11): 5479-5490.
[93] ZHANG Jing, LI Wanqing, OGUNBONA P O, et al. RGB-D-based action recognition datasets: A survey [J]. Pattern Recognition, 2016, 60:86-105.
[94] BERTASIUS G, WANG Heng, TORRESANI L. Is space-time attention all you need for video understanding? [J]. arXiv preprint arXiv:2102.05095, 2021.
[95] LONG Xiang, GAN Chuang, DE MELO G, et al. Attention clusters: Purely attention based local feature integration for video classification [C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE. 2018:7834-7843.
[96] HAO Zongbo, ZHANG Qianni, EZQUIERDO E, et al. Human action recognition by fast dense trajectories [C]// Proceedings of the 21st ACM International Conference on Multimedia. Barcelona, Spain: ACM, 2013:377-380.
[97] ANURADHA K, SAIRAM N. Spatio-temporal based approaches for human action recognition in static and dynamic background: A survey [J]. Indian Journal of Science and Technology, 2016, 9(5). DOI:10.17485/ijst/2016/v9i5/72065.
[98] GHADIYARAM D, TRAN D, MAHAJAN D. Large-scale weakly-supervised pre-training for video action recognition [C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA,

USA：IEEE. 2019：12038-12047.
[99] LORRE G，RABARISOA J，ORCESI A，et al. Temporal contrastive pretraining for video action recognition [C]// Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Snowmass，CO，USA：IEEE，2020：651-659.
[100] WANG L，QIAO Y，TANG X. Action recognition and detection by combining motion and appearance features [J]. THUMOS14 Action Recognition Challenge，2014，1(2)：2.
[101] LEE M，LEE S，SON S，et al. Motion feature network：Fixed motion filter for action recognition [C]// Computer Vision-ECCV 2018. Munich，Germany：Springer International Publishing，2018：392-408.
[102] FATHI A，MORI G. Action recognition by learning mid-level motion features [C]//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage，AK，USA：IEEE，2008：1-8.
[103] SEVILLA-LARA L，LIAO Y，GÜNEY F，et al. On the integration of optical flow and action recognition [C]// Lecture Notes in Computer Science. Springer International Publishing. Stuttgart，Germany：dblp，2019：281-297.
[104] PIERGIOVANNI A J，RYOO M S. Representation flow for action recognition [J]. arXiv preprint arXiv：1810.01455v1，2018.
[105] FEICHTENHOFER C，PINZ A，ZISSERMAN A. Convolutional two-stream network fusion for video action recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，NV，USA：IEEE，2016：1933-1941.
[106] HUANG D A，RAMANATHAN V，MAHAJAN D，et al. What makes a video a video：Analyzing temporal information in video understanding models and datasets [C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018.
[107] WANG H，KLäSER A，SCHMID C，et al. Dense trajectories and motion boundary descriptors for action recognition [J]. International Journal of Computer Vision，2013，103(1)：60-79.
[108] WANG Heng，SCHMID C. Action recognition with improved trajectories [C]//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney，Australia：IEEE，2013：3551-3558.
[109] JAIN M，JEGOU H，BOUTHEMY P. Better exploiting motion for better action recognition [C]//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland：IEEE，2013：2555-2562.
[110] WEINLAND D，ÖZUYSAL M，FUA P. Making action recognition robust to occlusions and viewpoint changes [C]// Computer Vision-ECCV 2010. Berlin/Heidelberg：Springer. 2010：635-648.
[111] ANGELINI F，FU Z，LONG Y，et al. 2D Pose-based real-time human action recognition with occlusion-handling [J]. IEEE Transactions on Multimedia，2020，22(6)：1433-1446.
[112] HAO T，WU D，WANG Q，et al. Multi-view representation learning for multi-view action recognition [J]. Journal of Visual Communication and Image Representation，2017，48：453-460.
[113] OUYANG Wanli，WANG Xiaogang，ZHANG Cong，et al. Factors in finetuning deep model for object detection with long-tail distribution [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas，NV，USA：IEEE，2016：864-873.
[114] ZHANG Xiao，FANG Zhiyuan，WEN Yandong，et al. Range loss for deep face recognition with long-tailed training data [C]// Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice，Italy：IEEE，2017：5409-5418.
[115] SOZYKIN K，PROTASOV S，KHAN A，et al. Multi-label class-imbalanced action recognition in hockey videos via 3D convolutional neural networks [C]//Proceedings of 2018 19th IEEE/ACIS International Conference on Software Engineering，Artificial Intelligence，Networking and Parallel/Distributed Computing (SNPD). Busan，Korea (South)：IEEE，2018：146-151.
[116] ZHANG Xing，WU Zuxuan，WENG Zejia，et al. VideoLT：Large-scale long-tailed video recognition [C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal，QC，Canada：IEEE，2021：7940-7949.
[117] DING Wan，HUANG Dongyan，CHEN Zhuo，et al. Facial action recognition using very deep networks for highly imbalanced class distribution [C]//Proceedings of 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Kuala Lumpur，Malaysia：IEEE，2017：368-1372.
[118] WU D，WANG Z，CHEN Y，et al. Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset [J]. Neurocomputing，2016，190：35-49.