doi:10.3969/j.issn.1001-4616.2025.06.011

基于 X3D 特征和语义融合的篮球运动员检测跟踪方法

韩乾乾[1],顾华宁[2],金　溢[3],赵永强[4]

(1.广东海洋大学体育与休闲学院,广东 湛江 524088)
(2.成都理工大学管理科学学院,四川 成都 610059)
(3.西北民族大学化工学院,甘肃 兰州 730030)
(4.广西师范大学基建处,广西 桂林 54100)

[**摘要**] 针对单机位篮球视频中运动员频繁遮挡、队服相似及相机抖动导致的跟踪难题,本文提出融合 X3D 时空特征与多模态语义的实时检测跟踪方法. 首先,采用单应性稳像将直播流配准至参考平面,抑制背景漂移. 其次,利用轻量化 X3D 网络在 16 帧片段上提取 1 024 维时空描述子(5 GFLOPs 算力约束),捕获各种篮球比赛的关键动作模式,同时满足了边缘部署的延迟要求;最后,设计注意力驱动的特征融合模块,自适应结合几何位移、外观直方图与 X3D 特征. 在 NBA-SYN 和 UCF-Sports-Basket 公开数据集上的实验表明,该方法分别达到 77.43%和 79.30%的 MOTA,以 45.31 FPS 的实时性能,显著优于现有方案,为有效硬件条件下的篮球视频分析提供可靠技术支撑.

[**关键词**] 篮球视频,检测跟踪,X3D 网络,单应性稳像,特征融合
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2025)06-0101-10

Basketball Player Detection and Tracking Method Based X3D Feature and Semantic Fusion

Han Qianqian[1],Gu Huaning[2],Jing Pen[3],Zhao Yongqiang[4]

(1.College of Sports and Recreation,Guangdong Ocean University,Zhanjiang 524088,China)
(2.College of Management Science,Chengdu University of Technology,Chengdu 610059,China)
(3.Northwest Minzu University,College of Chemistry and Chemical Engineering,Lanzhou 730030,China)
(4.Infrastructure Construction Department,Guangxi Normal University,Guilin 541004,China)

Abstract:In order to solve the tracking problems caused by frequent occlusion of players,similar uniforms and camera shake in single-camera basketball videos,this paper proposes a real-time detection and tracking method that integrates X3D spatiotemporal features and multimodal semantics. First,homography stabilization is used to align the live stream to the reference plane to suppress background drift. Secondly,a lightweight X3D network is used to extract 1 024-dimensional spatiotemporal descriptors(5 GFLOPs computing power constraint)on 16-frame segments to capture the key action patterns of various basketball games while meeting the latency requirements of edge deployment;finally,an attention-driven feature fusion module is designed to adaptively combine geometric displacement,appearance histogram and X3D features. Experiments on the NBA-SYN and UCF-Sports-Basket public datasets show that this method achieves 77.43% and 79.30% MOTA respectively,with a real-time performance of 45.31 FPS,which is significantly better than the existing solutions,providing reliable technical support for basketball video analysis under effective hardware conditions.

Key words:basketball video,detection and tracking,X3D network,homography stabilization,feature fusion

篮球职业比赛是一项典型的多人、多策略高对抗运动,比赛节奏快、战术组合复杂. 进攻球员通过挡拆掩护、空切跑位、突然反跑等策略创造进攻空间,防守球员则通过快速换防、夹击、补防等策略予以回应. 这种高频次的攻防互动使得运动员之间经常出现同步冲刺或瞬间拉开的运动模式,导致队形高度动

收稿日期:2025-07-17.
基金项目:国家自然科学基金资助项目(42402278)、甘肃省科技计划资助项目(23YFGA0073)、2023 年度广东省教育科学规划课题(高等教育专项)资助项目(2023GXJK297).
通讯作者:韩乾乾,博士研究生,讲师,研究方向:人工智能,目标识别,运动训练. E-mail:19232022628@163.com

态且密集. 此外,球员之间频繁发生遮挡,进一步增加了视觉跟踪的难度. 除了上述视觉挑战外,更为实际的约束条件则来自于硬件和经济成本. 由于建设和维护成本高昂,以及多数职业篮球场馆的顶部空间有限,很难在篮球赛场中部署多摄像机阵列实现多视角跟踪[1-2]. 绝大多数比赛的视频数据只能来源于官方的赛事直播画面,这些直播画面通常由一到两台摄像机拍摄,位于球场的中线附近. 比赛过程中,摄影师只进行有限的平移和轻微的变焦操作,缺乏场景的完整视图. 因此,如何利用这些有限的单机位直播视频,实现球员的稳定、可靠且准确的跟踪,成为该领域研究的关键问题[3-4].

针对上述复杂而又实际的挑战,本文采用经典的 tracking-by-detection 框架[5-6]进行求解. 具体而言,首先利用目标检测器在每一帧图像中检测出所有可能的球员目标,并生成初始的候选框;随后,为每个候选框提取用于识别与关联的高维描述子,并在连续两帧的候选框之间建立可能的对应关系,通过计算描述子之间的相似性,构造数据关联的代价矩阵;最后,基于匈牙利算法(Hungarian Algorithm)[7]求解该代价矩阵,以最小化总匹配代价的方式确定前后帧之间的全局最佳匹配,从而输出球员的稳定连续轨迹. 在这个过程中,描述子的判别能力和多样性至关重要:描述子必须能够有效地区分不同球员,并且对于姿态变化、局部遮挡、运动模糊等问题具备较好的鲁棒性.

为了实现更可靠的轨迹匹配,本文综合考虑并融合了三类互补的特征描述子,包括几何特征、视觉特征和深度 X3D[8]特征. 其中,几何特征利用检测框质心在图像像素坐标中的位移及方向变化,提供运动轨迹的短期物理约束,有效防止目标瞬间跨越非合理的距离;视觉特征则通过对比检测框及其邻域区域的 RGB 颜色直方图、相关性等视觉属性,区分球员的队服、肤色、号码和鞋袜细节,有效增强球员的视觉辨别力;深度特征则引入了轻量级的 3D 卷积网络 X3D,在视频的 16 帧局部片段上生成 1 024 维时空描述子,以充分捕获投篮、急停、启动和回合转换等关键动作模式. X3D 网络采用了逐轴扩展的策略[8],在 5 GFLOPs 以内的计算资源限制下,将网络的表达能力接近最优,从而在准确性与实时性之间取得了优异的平衡.

本文的主要创新点总结如下:

(1)通过引入轻量 X3D,在 16 帧局部片段上生成 1 024 维时空描述子,捕获投篮、急停、回合转换等动作模式. 把 5 GFLOPs 内的算力利用率逼近上限,既保持了 SlowFast 同级别的表达力,又满足边缘部署的延迟要求.

(2)除了特征融合,我们还引入单应性稳像作为预处理步骤,将直播流配准到统一参考平面,显著抑制因平移或变焦带来的背景漂移.

(3)提出了 X3D 特征+相机稳像+多模态语义融合方案,在单机位篮球视频上实现了鲁棒、实时的球员检测与跟踪,为自动战术分析和比赛统计奠定了可靠基础.

1 相关工作

1.1 多目标跟踪

在视频序列中对多目标进行实时、精确的定位与关联,长期以来是计算机视觉领域的核心课题. 当前主流框架仍以 tracking-by-detection 为主,即先用检测器定位候选目标,再在时间轴上完成数据关联. 该范式随深度检测器的快速迭代而持续受益. Girshick 等[9]开创的两阶段检测器被广泛嵌入跟踪模型中,后续方法将 CNN 检测输出与外观特征或运动先验联合起来进行局部匹配[10-12];另一类方法将多帧检测结果统一建图,通过网络流或整数规划获得跨帧最优关联[13].

除了检测-关联路线,判别相关滤波(Discriminant Correlation Filter,DCF)系列方法[14]以其速度优势在单目标扩展到多目标时仍具参考价值. 早期 DCF 依赖方向梯度直方图(Histogram of Oriented Gradients,HoG)等手工特征,近年开始使用预训练的 CNN 网络进行特征提取并尝试端到端更新相关滤波核. 更有研究探索无监督框架,摆脱昂贵的盒标注[15-16].

若关注人体动作结构,则需要在连续帧中对关键点进行检测与目标标记保持对应,即姿态跟踪. OpenPose 算法[17]和 AlphaPose 算法[18]将自底向上的关键点检测结果通过骨架连接,后续利用时序一致性完成多人物体的持久跟踪,为运动分析提供了细粒度表征.

1.2 体育视频中的目标跟踪

体育转播具有视角单一、遮挡频繁、队服相似度高等特点,使多目标跟踪难度陡增. Voigtlaender 等[19]

引入头部/全身双检测器以增强遮挡鲁棒性. 张炜昕等[20]对现有体育视频检测与跟踪技术作了系统回顾,在 YOLO v5 的基础上进行改进以解决篮球特征单一的问题,并建立分类惩罚机制以降低误检率. 为了在篮球直播中,正确提取比赛特征的知识,Facchinetti 等[21]提出了一种利用篮球运动员的追踪数据自动识别活动时段的算法,在多尺度上提取 CNN 特征,该算法基于运动员运动参数的阈值,阈值的确定是通过比赛视频分析中提取的视觉信息.

2 提出的方法

所提方法的架构如图 1 所示. 首先,对每一帧直播画面进行场地掩膜和稳像处理,去除观众席并消除摄像机平移或变焦带来的抖动;随后,调用 YOLO v7 进行球员检测,获得当前帧的所有球员候选框. 对每个候选框,系统并行提取几何位置信息、颜色纹理外观以及短时动作特征,并通过注意力机制将 3 路信息自适应融合成统一表示. 接着,把上一帧的已有轨迹与当前帧的检测结果构造成代价矩阵,并利用匈牙利算法寻找全局最优匹配,实现跨帧身份的关联;未匹配到的检测会启动新轨迹,长时间未匹配的轨迹则被终止. 如此循环处理整段视频,最终输出连续且唯一的球员轨迹,既保持实时性,又兼顾遮挡和快速换位情况下的匹配准确度.

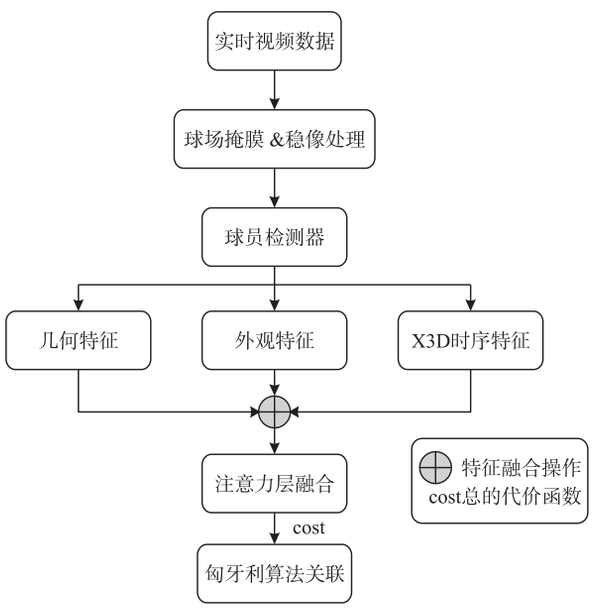


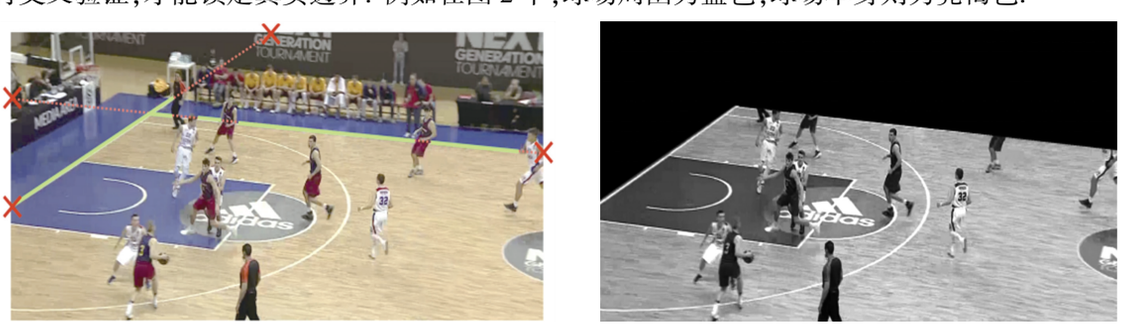
图 1　本文方法的总体框架图

Fig. 1　Overall framework diagram of proposed method

本文采用 X3D+单应稳像+注意力的模块化设计并非简单拼装,主要差异在于:(1)在线稳像与一致性门控,单应估计伴随时序一致性检测,仅在置信度通过门控时才作用于轨迹更新(避免稳像漂移);(2)几何与外观联合注意:注意力同时受球场几何(场地区域/边界)与队伍颜色先验约束,强调“合理位置上的合理外观”. (3)轻量多尺度时序聚合,X3D 的步幅与感受野按球场节奏调整,配合多尺度池化,保证在 1 280×720 分辨率下达到 ≥30 fps 的实时要求.

2.1 预处理

球场检测:虽然球场检测并非本文的研究重点,但想要剔除替补席、裁判等无关候选,首先必须把可见球场轮廓勾勒出来. 本文只假设:①球场边线/底线颜色一致;②观众席与球员区域有明显距离. 在直播画面中,球场通常呈含若干可见边线的梯形. 本文采用无参数直线检测器,先提取整幅图像的线段,再将位于边缘、方向一致且两两相交的线段合并为同一直线. 可见部分最长、且与同向线段总长度成正比者,被视作主方向线段. 由于篮球场包含大量平行标记线(边线、底角三分线、三秒区边线等),必须对多条候选线进行交叉验证,才能锁定真实边界. 例如在图 2 中,球场周围为蓝色,球场本身则为亮褐色.



(a) 分割点的确认　　(b) 最终分割结果

图 2　本文方法中的球场检测

Fig. 2　The detection of proposed method on basketball court

本文仅关心两条主方向:

(1)边线方向——两侧边线交点所在方向;

(2)底线方向——普通机位下只能看到场地一端的底线.

在 HSV 空间做简单颜色过滤即可满足大多数场景:如图 2,观众席呈蓝色,球场本体为亮褐色. 对任一主方向,执行下列步骤:

(1)在图像顶部(边线)或左侧(底线)初始化候选线;

(2)在该线两侧±25 像素 处各绘制一条平行线;

(3)仅统计候选线与两条平行线之间、且满足色彩阈值的像素数:若为边线候选,则比对上下两侧;若为底线候选,则比对左右两侧. 图 2 中,色相 120°~150° 的像素被视为“球场外部蓝色”;

(4)将候选线沿场外方向(边线向下、底线向右)平移 12 像素,并重复步骤 2~3;

(5)遍历所有平移后,取“上下差”或“左右差”最大的候选线作为最终球场边界. 示例中,最优线恰位于褐色与蓝色区域交界处.

此外,为减轻后续球员跟踪的负担,本文可选地加入摄像机稳定步骤来抵消镜头抖动. 由于该步骤会增加额外开销,本文在消融实验中评估其收益. 实现上采用文献[22]的方法:对每帧估计单应矩阵,将整段视频配准到公共参考平面,从而得到稳像序列.

2.2 球员检测

本文方法基于姿态建模技术[17-18],采用基于卷积神经网络的自底向上处理范式,具体包含三个阶段:

(1)通过卷积网络检测人体结构关键点;

(2)运用图论方法连接关键点构建肢体拓扑;

(3)将检测到的肢体结构融合为完整的人体骨骼表示.

针对篮球比赛视频帧的处理,系统对每个球员的姿态估计输出包含两部分:其一为 25×3 的姿态向量,其中每个维度分别对应人体 25 个关键点的屏幕坐标位置及其检测置信度;其二为 26 通道的热力图(heatmap),每个通道编码特定关键点在像素空间中的概率分布. 需要特别说明的是,当出现肢体遮挡或运动模糊导致部分关键点检测失效时,姿态向量中对应位置的坐标值将被标记为未知状态. 系统通过综合分析所有关键点的空间分布,计算每个检测对象的最小外接矩形坐标,最终为球员生成精确的边界框定位.

2.3 视频时序特征的提取

时空卷积(3D 卷积)作为静态图像卷积(2D 卷积)在时序维度的自然延伸,其核心在于通过三维滑动窗口(包含时间、高度、宽度维度)对视频序列进行特征提取. 不同于 2D 卷积仅捕捉空间局部相关性,传统 3D CNN(I3D[23]、C3D[24]、SlowFast[25]等)经常把 2D ResNet 的卷积核直接“充气”成三维核,例如把 3×3 改成 k_t×3×3,其中,k_t 表示时间维度. 从数学原理上看,3D 卷积操作可描述为输入特征图与三维卷积核的多维度互相关运算.

给定输入视频的时空特征图,卷积核 W 定义在三维空间(D,H,W)(分别对应时间深度、空间高度、空间宽度),则输出特征图上坐标(t,j,k)处的像素值 $y[t,j,k]$可通过下式计算:

$$y[t,j,k]=\sum_{d=0}^{D-1}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}x[t+d,j+h,k+w]\cdot W+b, \tag{1}$$

其中,$x[t+d,j+h,k+w]$表示在时空位置$(t+d,j+h,k+w)$处的像素值,W 表示权重参数,b 为偏置项.

目前常用的 I3D 通过膨胀的二维卷积核和双流网络实现了高性能,但计算复杂度较高. C3D 虽然简单高效,但泛化性可能受到限制. 这些特征提取器的计算量随时间维长度 k_t 呈线性爆炸,随通道宽度呈平方级增长,导致在移动或实时场景下难以部署.

X3D 强调先小后大,逐轴扩展:先构造一个极小的 2D 网络(只有 24 个通道、112×112 分辨率、4 帧输入),再沿五个轴——时间长度、空间分辨率、网络深度、通道宽度、瓶颈扩张比——逐步放大,并在每一步检查精度增益与计算开销的比例是否划算.

逐轴扩展策略总结如下:在分辨率轴上,将输入从 112×112 提升到 160×160,可显著提高小目标辨识度;在时间轴上,剪裁或循环补齐,把 4 帧增至 13 → 16 → 32 帧;在宽度轴上,通道数从 24 递增到 48、96

……直至逼近 FLOPs 预算;在深度轴上,网络中段插入更多残差块,弥补加宽后模型容量不足;在瓶颈扩张比上,将 ResNet 骨干网络的展开系数从 1×调到 2×或 4×,提升表达力.

此外,X3D 沿用标准的标签平滑分类交叉熵,其定义为:

$$L_{CE}=-\sum_{c=1}^{C}\left[(1-\varepsilon)l(c=y)+\frac{\varepsilon}{C}\right]\log p_c, \tag{2}$$

其中,一般情况下 $\varepsilon=0.1$,C 是类别的数量,p_c 是 Softmax 概率.

函数 $l(\cdot)$是一个指示函数(indicator function),其定义如下:

$$l(c=y)=\begin{cases}1 & \text{如果当前类别 } c \text{ 就是真实标签 } y\\ 0 & \text{其他情况}\end{cases}. \tag{3}$$

该函数的作用是 1)标记真类别,当 c 正好等于真实标签 y 时,这个位置才取值 1,使得该类别的目标概率变为 $(1-\varepsilon)+\frac{\varepsilon}{C}$;2)抑制非真类别对于所有其他 $c\neq y$,指示函数为 0,则只留下平滑项$\frac{\varepsilon}{C}$,给每个非真类别一个均匀的小概率作为软标签.

X3D 在 Kinetics-400 上采用 30 → 100 → 180 epoch 的三阶段余弦退火学习率调度. 输入随机裁剪、随机水平翻转,对时间轴做随机偏移或循环填补. 预训练权重可无缝迁移到动作检测、视频异常检测、时序字幕对齐等下游任务. 贪心搜索表明:在同样 5 GFLOPs 预算下,X3D 的 Top-1@ K400 提升到 76%,而 SlowFast-R50 只有 72%,且参数量还少一半. 因此本文选用 *X3D* 进行特征提取器.

2.4 特征融合与代价函数

根据前三小节的处理流程,假设得到几何特征为 g_i,外观直方图为 v_i,X3D 时序特征为 f_i. 下面对特征进行如下处理,首先对所有特征进行维度对齐:

$$\tilde{\boldsymbol{g}}_i=\mathrm{ReLU}(\boldsymbol{W}_g\boldsymbol{g}_i+\boldsymbol{b}_g), \tag{4}$$

$$\tilde{\boldsymbol{v}}_i=\mathrm{ReLU}(\boldsymbol{W}_v\boldsymbol{v}_i+\boldsymbol{b}_v), \tag{5}$$

$$\tilde{\boldsymbol{f}}_i=\mathrm{ReLU}(\boldsymbol{W}_f\boldsymbol{f}_i+\boldsymbol{b}_f), \tag{6}$$

其中,$\tilde{\boldsymbol{g}}_i,\tilde{\boldsymbol{v}}_i,\tilde{\boldsymbol{f}}_i\in\mathbf{R}^d$,将三路特征各自投影到同一维度 d.

然后,计算分支级注意力权重:

$$\boldsymbol{s}_k=\boldsymbol{q}^{\mathrm{T}}\tanh(\boldsymbol{U}_k\tilde{\boldsymbol{x}}_k), \tag{7}$$

其中,$\tilde{x}_k$ 代表第 k 路经过线性投影并激活后的特征向量;U_k 是该分支专用的权重矩阵;q 是全局可学习的查询向量. 通过一个双线性打分(先计算 tanh 数值,再内积),模型得到每条分支对当前样本的重要性分值 s_k.

经过 Softmax 归一化后,最终得到我们需要的权重:

$$\boldsymbol{\alpha}_k=\frac{\exp(\boldsymbol{s}_k)}{\sum_{p\in\{g,v,f\}}\exp(\boldsymbol{s}_p)} \tag{8}$$

其中,所有分支的 $\boldsymbol{s}_k$ 经 Softmax 转换为 $\boldsymbol{\alpha}_k$,满足 $\alpha_g+\alpha_v+\alpha_f=1$.

最终的融合表示为:

$$\boldsymbol{z}_i=\boldsymbol{\alpha}_g\tilde{\boldsymbol{g}}_i+\boldsymbol{\alpha}_v\tilde{\boldsymbol{v}}_i+\boldsymbol{\alpha}_f\tilde{\boldsymbol{f}}_i. \tag{9}$$

为了进一步强调维度级别的选择,本文假如元素门控 β_i 如下:

$$\boldsymbol{\beta}_i=\sigma(\boldsymbol{W}_\beta[\tilde{\boldsymbol{g}}_i;\tilde{\boldsymbol{v}}_i;\tilde{\boldsymbol{f}}]), \tag{10}$$

其中,$\sigma(.)$表示 Sigmoid 函数.

经过元素门控操作之后得到:

$$\boldsymbol{z}_i=\boldsymbol{\beta}_i\odot\boldsymbol{z}_i, \tag{11}$$

其中,$\odot$表示逐元素乘积操作.

总的代价函数表示为:

$$\mathrm{cost}=\lambda_1\|\boldsymbol{g}_i-\boldsymbol{g}_j\|_2+\lambda_2\|\boldsymbol{v}_i-\boldsymbol{v}_j\|_1+\lambda_3(1-\cos(\boldsymbol{z}_i,\boldsymbol{z}_j)), \tag{12}$$

其中,$\lambda_1:\lambda_2:\lambda_3=1:2:3$. 上式中第 1 项表示几何距离,确保物理可达性;第 2 项表示外观距离,用于抑制队

服相似带来的误配;第 3 项表示嵌入余弦距离,用于综合时序与上下文信息,区分遮挡与快速换位.

本文算法的实现过程的伪代码如算法 1 所示. 算法中的关联机制与特征流图如下:每帧对活跃轨迹↔新检测进行匈牙利匹配,代价由外观、时序、几何三项加权组成,并设定几何与稳像置信度门控以抑制误配. 三类特征分别为:①外观:对检测框做 RoI 特征并经轻量 ReID 头提取 256 维 L2 归一化嵌入;②时序:对短窗裁剪输入 X3D 形成运动嵌入,用于约束速度与方向连续性;③几何:利用单应矩阵将框映射到场地平面,计算稳像 IoU/中心偏移并叠加合法区域约束,用于过滤不合理匹配.

算法 1　篮球直播球员检测-跟踪流程

输入:视频帧序列$\{F\cdots F_T\}$
输出:带连续 ID 的球员轨迹 TrackSet
初始 TrackSet ←∅
for $t=1\cdots T$ do
　　# 预处理
　　F← 稳像处理(场地掩膜(F_i))
　　# 球员检测
　　DetSet← PlayerDetector(F_i)
　　# 特征提取与融合
　　for $B_i \in DetSet_i$ do
　　　　g_i← 计算几何特征(B_i)
　　　　v_i← 计算外观特征(F_i,B_i)
　　　　f_i← X3D 特征(F_i,B_i)
　　　　z_i← 注意力融合(g_i,v_i,f_i)
　　end for
　　# 构造代价矩阵
　　if TrackSet ≠∅ then
　　　　C ← 计算总代价(TrackSet,$\{z_i\}$)
　　　　Match ← 匈牙利算法(C)
　　　　更新轨迹($TrackSet$,$Match$,$DetSet_i$)
　　else
　　　　将 $DetSet_i$ 中每个 B_i 初始化为新轨迹
　　end if
　　# 轨迹生命周期管理
　　刷新丢失计数并移除长时间未匹配轨迹
end for
return TrackSet

3　实验与分析

3.1　数据集与评价指标

本文选取两类风格各异的篮球直播数据:公开合成集 NBA-SYN 以及经典 UCF-Sports-Basket 子集. 两者均使用固定横向机位拍摄,涵盖常规进攻、防守、快攻和死球场景. 为提升真实转播风格的跨域代表性,本文采用 DeepSportradar 作为补充集. 该数据集由 Sportradar/Keemotion 在法国职业联赛 LNB Pro A 多场馆采集. 为保证标注一致性,对每帧球员、裁判及可见教练标注外接框和全局 ID,同时记录镜头切换与暂停时间戳,方便稳像与分析.

表 1　实验数据集信息

Table 1　The information of experimental datasets

数据集	总帧数	分辨率	训练/验证/测试
NBA-SYN	910k	1 920×1 080	10/3/5
UCF-Sports-Basket	300k	720×480	5/1/2
DeepSportradar	210k	720×480	5/1/2

采用的度量指标是多目标跟踪准确度(MOTA)[6]:

$$MOTA = \frac{\sum_t fp_t + m_t + mm_t}{\sum_t box_t}, \tag{13}$$

式中,fp_t 表示误警,m_t 表示漏警,mm_t 表示失配,box_t 表示所有视频序列中真实包围盒总数量.

此外,本文还采用平均准确率 mAP 和 F1 score 指标用于辅助评价. 这些指标越高表示性能越好.

3.2　实验设置

所有实验在 PyTorch 2.1+CUDA 11.8 环境下完成,硬件为单张 NVIDIA RTX 4080(16 GB). 人体检测器采用 YOLOv7-Pose;优化器为 AdamW,初始学习率 3×10^{-4},余弦退火调度. 该检测器可同时输出框坐标与 17 关键点,为后续几何特征提供额外姿态信息.

在跟踪模块中,本文为每个检测框并行提取三路特征:几何分支记录中心位移与朝向;外观分支由 64 维 RGB 直方图加浅层纹理 CNN 组成;时序分支采用轻量 X3D 处理以目标为中心的 16 帧剪辑,输出 1 024 维动作向量. 三路特征经单头加性注意力融合,得到最终嵌入用于匈牙利匹配. 融合层隐藏维度 256,Dropout 0.5. 跟踪头训练 20 epoch,batch size 32,学习率从 1×10^{-3} 线性递减至 1×10^{-5}. 所有实验设置同一随机种子,重复三次取均值以减小方差.

3.3　实验结果比较

本文方法与其他方法在两个数据集上的性能比较如表 2 所示. 从表 2 可以看出,在三套篮球直播测试集(NBA-SYN、UCF-Sports-Basket 和 DeepSportradar)上,本文方法在三项核心指标 MOTA、mAP、F1 均取得了最优成绩. 以 NBA-SYN 为例,本文 MOTA 达到 77.43%,比文献[20]高出 2.4%;mAP 提升至 85.39%,较所有基线至少高出 1.5%;F1 亦领先第二名约 1.3%. 类似趋势在分辨率更低、遮挡更频繁的 UCF-Sports-Basket 数据集上同样成立. 此外,本文方法取得了与 ByteTrack[26] 相当的性能. 这说明所提方法对不同摄像机参数与拍摄条件具有稳健泛化能力和优秀的检测能力.

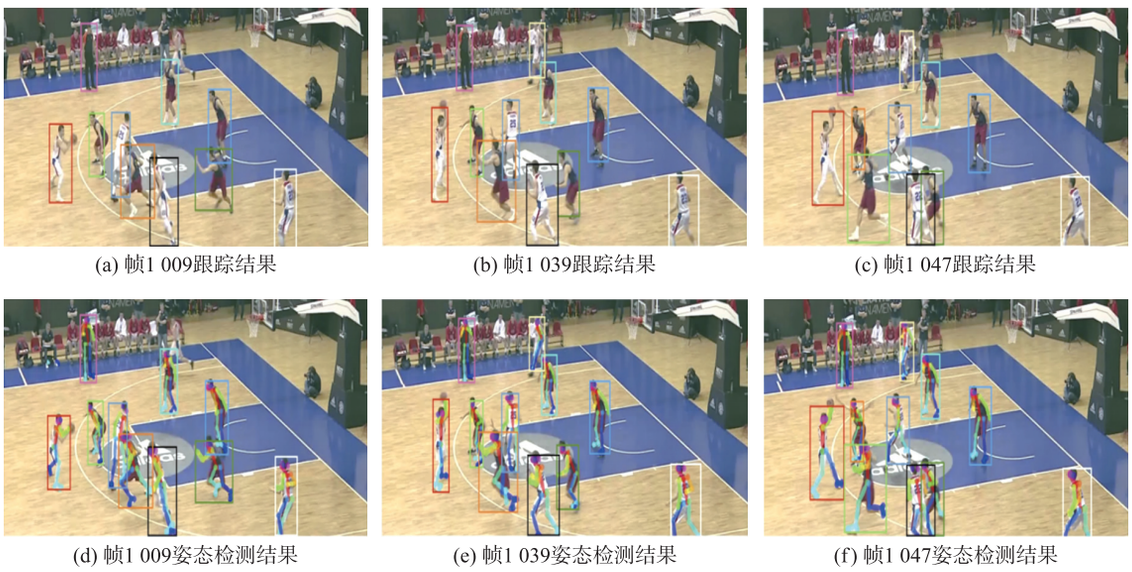
性能提升的主要来源可归因于三个模块:(1)X3D 时空特征为外观高度相似的同队队员提供了动作级区分信息,使 ID 关联更加可靠;(2)单应稳像显著削弱了镜头平移/变焦造成的背景漂移,减少误匹配;(3)注意力驱动的多模态融合在几何、颜色与时序信息之间自动分配权重,充分发挥各自优势. 相比现有工作,本文方法有效提升了检测精度与跨帧一致性,为后续战术分析和数据统计奠定了更高质量的轨迹基础.

表 2　本文方法与其他方法在两个数据集上的性能比较

Table 2　Performance comparison of proposed methods and other methods on two datasets

方法	检测器	NBA-SYN			UCF-Sports-Basket			DeepSportradar		
		MOTA	mAP	F1 score	MOTA	mAP	F1 score	MOTA	mAP	F1 score
文献[12]	CNN	72.91	81.02	81.98	73.18	81.09	80.98	70.19	82.03	80.91
文献[19]	Mask R-CNN	73.22	82.28	82.01	74.13	82.18	81.86	73.19	81.14	82.01
文献[20]	YOLO v5	75.01	83.87	84.98	77.39	84.69	85.02	76.42	85.01	85.17
文献[21]	无	72.49	81.28	82.09	74.38	81.31	81.22	77.55	83.31	82.32
文献[26]	YOLO X	78.10	86.09	86.21	80.11	87.52	87.01	79.91	85.29	85.15
本文	YOLO v7	77.43	85.39	86.31	79.30	87.61	87.11	79.51	86.56	85.18

图 3 呈现了视觉检测与跟踪的实验结果(该结果基于几何特征与深度学习特征的最优组合方案得出). 图中,不同颜色的包围盒对应着检测到的不同球员 ID,通过颜色区分可直观追踪各球员动态. 从实验结果可见:除首帧图像中存在一名球员漏检情况,所提方法在其余 32 次球员关联任务中均实现了准确匹配.



(a) 帧1 009跟踪结果　(b) 帧1 039跟踪结果　(c) 帧1 047跟踪结果

(d) 帧1 009姿态检测结果　(e) 帧1 039姿态检测结果　(f) 帧1 047姿态检测结果

图 3　连续帧中得到的跟踪和姿态检测结果

Fig. 3　Tracking and pose detection results obtained from consecutive frames

各方法的运行效率对比结果如表 3 所示. 从表 3 可见,本文方法在推理效率上全面领先于其他方案. 具体而言,凭借 YOLO v7 检测器与轻量级特征融合框架,我们在 1280×720 分辨率下实现 45.31 FPS,单帧计算量仅为 8.10 GFLOPs. 这表明:

YOLO v7 的更高参数利用率显著降低了前向计算;

我们的几何-外观-时序三分支在检测后端并未引入额外卷积堆叠,保持了极低算量. 因此,本方法同时具备高帧率与低算力双重优势,更适合实时篮球直播分析等部署场景.

表 3　各方法的运行效率对比

Table 3　Running efficiency comparison of different methods

方法	检测器	推理帧率(FPS)	单帧 GFLOPs
文献[12]	CNN	25.17	21.92
文献[19]	Mask R-CNN	30.28	32.08
文献[20]	YOLO v5	41.33	24.97
文献[21]	无	35.90	12.09
文献[26]	YOLO X	43.29	9.12
本文	YOLO v7	45.31	8.10

本文方法在检测跟踪的训练结果如图 4 所示. 可以看出,在两个数据集中,所提方法均在 20 个 epoch 中达到收敛状态.

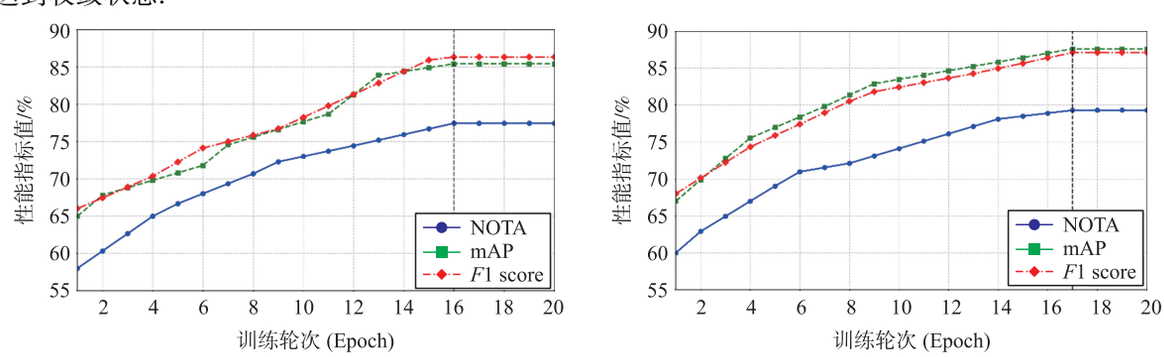


图 4　本文方法在检测跟踪的训练结果

Fig. 4　Training results of proposed method on detecting-tracking

3.4　消融实验

消融实验以 NBA-SYN 和 UCF-Sports-Basket 为实验对象,因为这两个数据集拥有更加完备的数据标注. 为了简化表示,本文用 F_1 表示几何特征,F_2 表示外观特征,F_3 表示时序特征.

表 4 给出了不同特征组合对性能的结果. 结果表明,特征稳定化处理和时序特征的引入对性能提升具有关键作用,其中 F1+F2+F3(稳定化)组合在两个数据集上均达到最佳性能(77.43%和 79.30%),显著优于双特征组合,说明几何、外观和时序特征存在互补性;同时稳定化处理使 F1+F3 组合在 NBA-SYN 数据集上提升 4.73%,验证了时序特征对运动稳定性的敏感性,而 UCF-Sports-Basket 数据集对三特征融合的更大响应(3.19% vs NBA 的 1.45%)则揭示了复杂运动场景对多维特征联合建模的更强需求.

表 4　不同特征组合对性能的影响

Table 4　Impact of different feature combinations on performance　%

特征组合	MOTA	
	NBA-SYN	UCF-Sports-Basket
F_1+F_2	70.31	72.39
F_1+F_3	71.25	74.15
F_1+F_2(稳定化)	73.37	76.11
F_1+F_3(稳定化)	75.98	77.09
$F_1+F_2+F_3$(稳定化)	77.43	79.30

4　结论

本文通过融合 X3D 时空建模、单应性稳像与多模态特征融合,实现了单机位篮球视频的鲁棒实时跟踪. 实验验证表明:稳像预处理显著提升时序特征鲁棒性,三特征融合达到最优性能,轻量 X3D 在 5 GFLOPs 算力下实现高效动作编码,支撑 45.31 FPS 实时处理. 该方法在遮挡与快速换位场景下保持高跟踪一致性,为自动化战术分析奠定基础. 未来将研究自适应稳像权重与动态融合策略,进一步提升复杂交互场景的泛化能力.

[参考文献]

[1] 戴凤麟. 无标定的自适应多视角跟踪系统[D]. 上海:复旦大学,2025.

[2] ZHANG Y,SHANG A,ZHANG W,et al. A measurement fusion algorithm of active and passive sensors based on angle association for multi-target tracking[J]. Information fusion,2024,106(12):1-16.

[3] CAO Z,SIMON T,WEI S E,et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//EEE Conference on Computer Vision and Pattern Recognition. Hawaii,USA:IEEE,2017:1302-1310.

[4] LOZZI D,DI POMPEO I,MARCACCIO M,et al. AI-Powered Analysis of Eye Tracker Data in Basketball Game[J]. Sensors,2025,25(11):1019-1028.

[5] YUE W,XU F,YANG J. Tracking-by-Detection Algorithm for Underwater Target Based on Improved Multi-Kernel Correlation Filter[J]. Remote sensing,2024,16(2):1-16.

[6] 马昌庆. 面向大场景监控视频的行人多目标跟踪算法研究[D]. 沈阳:东北大学,2021.

[7] 朱国晖,漆娜,郭子萱. 车联网中基于进化策略算法与匈牙利算法的资源分配策略[J]. 西安邮电大学学报,2024,29(4):21-29.

[8] FEICHTENHOFER C. X3d: Expanding architectures for efficient video recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington,USA:IEEE,2020:203-213.

[9] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR). Columbus,Ohio:IEEE,2014:580-587.

[10] 龙君芳,马琳娟,李庆珍. 基于 KPCA 和结构化支持向量机的视频目标跟踪[J]. 南京理工大学学报,2023,47(5):671-677.

[11] 张国印,王传博,高伟. 抗遮挡的行人多目标跟踪算法[J]. 智能系统学报,2024,19(5):1248-1256.

[12] 张英俊,白小辉,谢斌红. CNN-Transformer 特征融合多目标跟踪算法[J]. 计算机工程与应用,2024,60(2):180-190.

[13] 张丽娟,周治平. 基于网络流的分层关联多目标跟踪[J]. 计算机辅助设计与图形学学报,2018,30(9):1670-1677.

[14] 苏银强,王宣,王淳,等. 一种用于视觉跟踪的低秩上下文感知的相关滤波器[J]. 计算机科学,2024,51(9):121-128.

[15] WANG W,ZHANG K,LV M,et al. Hierarchical spatiotemporal context-aware correlation filters for visual tracking[J]. IEEE transactions on cybernetics,2020,51(12),6066-6079.

[16] YAN B,PENG H,FU J,et al. Learning spatio-temporal transformer for visual tracking[C]//In Proceedings of the IEEE/CVF International Conference on Computer Vision(CVPR). Nashville,Tennessee:IEEE,2021:10448-10457.

[17] CAO Z,TOMAS S,SHIH-EN W,et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Hawaii,USA:IEEE,2017:7291-7299.

[18] FANG H S,XIE S Q,TAI Y W,et al. Rmpe: Regional multi-person pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision(CVPR). Hawaii,USA:IEEE,2017:2334-2343.

[19] VOIGTLAENDER P, MICHAEL K, ALJOSA O, et al. MOTS: Multi-object tracking and segmentation[C]//Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, California, USA: IEEE, 2019: 7942-7951.
[20] 张炜昕,钟映春,张钢,等. 视频流中移动篮球的检测与跟踪方法[J]. 广东工业大学学报,2025,42(3):62-71.
[21] FACCHINETTI T, METULINI R, ZUCCOLOTTO P. Filtering active moments in basketball games using data from players tracking systems[J]. Annals of operations research, 2023, 325(1): 521-538.
[22] SÁNCHEZ J. Comparison of motion smoothing strategies for video stabilization using parametric models[J]. Image processing on line, 2017, 29(7): 309-346.
[23] CARREIRA J, ANDREW Z. Quo vadis, action recognition? a new model and the kinetics dataset[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Hawaii, USA: IEEE, 2017: 6299-6308.
[24] 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(ICCV). Santiago, Chile: IEEE, 2015: 4489-4497.
[25] 王志明,张佳,彭江南,等. SlowFast 架构下景区异常行为识别算法及预警研究[J]. 南京理工大学学报,2024,48(3):374-383.
[26] ZHANG Y F, PEIZE S, YI J, et al. Bytetrack: Multi-object tracking by associating every detection box[C]//European Conference on Computer Vision. Tel Aviv, Israel: IEEE, 2022: 1-21.

[责任编辑:陆炳新]

(上接第 10 页)

[16] Han Y B, FENG S. PAN H. Stability of F-stationary maps of a class of functionals related to conformal maps[J]. Kodai mathematical journal, 2013, 36(3): 455-478.
[17] HAN Y B, HAN X, XUE Y, et al. Liouville theorems for $\Phi_{S,p,\varepsilon}$-harmonic map[J]. Journal of Sichuan Normal University (natural science), 2023, 46(3): 362-373.
[18] FADDEEV L, NIEML A J. Stable knot-like structures in classical field theory[J]. Nature, 1997, 387: 58-61.
[19] SPEIGHT J, SVENSSON M. On the strong coupling limit of the Faddeev-Hopf model[J]. Communications in mathematical physics, 2007, 272(3): 751-773.
[20] SPEIGHT J, SVENSSON M. Some global minimizers of a symplectic Dirichlet energy[J]. The quarterly journal of mathematics, 2011, 62(3): 737-745.
[21] HAN Y B. Monotonicity formulas of E_F-critical maps with potential[J]. Journal of mathematical research with applications, 2015, 35(6): 681-691.
[22] BRANDING V. The heat flow for the full bosonic string[J]. Annals of global analysis and geometry, 2016, 50(4): 347-365.
[23] CAO X Z, Chen Q. Existence of harmonic maps with two-form and scalar potentials[J]. Vietnam journal of mathematics, 2021, 49(2): 349-361.
[24] JOST J, YAU S T. A nonlinear elliptic system for maps from Hermitian to Riemannian manifolds and rigidity theorems in Hermitian geometry[J]. Acta mathematica, 1993, 170(2): 221-254.
[25] CHEN Q, JOST J, QIU, H. Existence and Liouville theorems for V-harmonic maps from complete manifolds[J]. Annals of global analysis and geometry, 2012, 42(4): 565-584.
[26] WANG G, XU D. Harmonic maps from smooth metric measure spaces[J]. International journal of mathematics, 2012, 23(9): 1250095.
[27] ZHAO G. A monotonicity formula and a Liouville type theorem of V-harmonic map[J]. Bulletin of the Korean mathematical society, 2019, 56(5): 1327-1340.
[28] KOH D. The evolution equation for closed magnetic geodesics[D]. Potsdam: Universitätsverlag Potsdam, 2008.
[29] AFUNI A. Monotonicity for p-harmonic vector bundle-valued k-forms[J]. arXiv preprint: 1506.03499(2015).
[30] FENG S, HAN Y. Liouville type theorems of F-harmonic maps with potential[J]. Results in mathematics, 2014, 66(1-2): 43-64.
[31] HAN Y B, FENG S X. Some results of weakly f-stationary maps with potential.(English)[J]. Journal of mathematics, 2017, 37(2): 301-314.

[责任编辑:陆炳新]