## 基于Lighttrack的足球比赛视频多人姿态估计与追踪

徐 薪 芮仁婷 陈 畅

指导老师：盛 斌

摘  要: 人体骨骼关键点对与描述人体姿态，预测人体行为非常重要。在数字媒体时代，从海量足球比赛视频中通过深度学习抽象、简洁的姿态数据，能够更好地完成足球运动员模型的动作模拟，应用于球队训练指导甚至虚拟现实建模。本小组通过深度学习的人体姿态估计算法完成了对多人比赛视频的人体骨骼关键点估计和人体跟踪的工作。本小组复现了Lighttrack这一新型轻量在线框架，包括人体定位，姿态估计和人体跟踪三个部分。相较于其他模型，此框架生成帧速率高，通用性强，组件可替换。本小组同时针对复现效果做了三个方面的改进：设计Soft NMS算法提高人体检测候选框的召回率、设计人体候选框大小动态调节算法消除背景观众的干扰、设计采样关键帧帧间隔动态调节算法提高人体姿态跟踪准确率，在节点分离和重叠这两个主要问题上有了很大的提升。项目程序源于Lighttrack论文[1]作者上传于GitHub平台的开源代码：[GitHub - Guanghan/lighttrack: LightTrack: A Generic Framework for Online Top-Down Human Pose Tracking](https://github.com/Guanghan/lighttrack)。

**Lighttrack-based multiplayer pose estimation for soccer game videos**

Xin Xu, Renting Rui, Chen Chang

**Abstract**:  Human skeletal keypoints are important for describing human posture and predicting human behavior. Our group accomplishes human skeletal keypoint estimation and human tracking for multiplayer football competition videos by using deep learning human posture estimation algorithms. Our group replicates Lighttrack, a novel lightweight online framework that includes three components: human detection, pose estimation, and human tracking. Compared with other models, this framework generates high frame rates, is highly versatile, and has interchangeable components. Our group has also made three improvements for the reproduction effect, and improved the two main problems of keypoints separation and human overlap. The code in this project is derived from [1]: [GitHub - Guanghan/lighttrack: LightTrack: A Generic Framework for Online Top-Down Human Pose Tracking](https://github.com/Guanghan/lighttrack).

**Key word: Pose estimation, pose tracking, data association, 3D prediction**

# 简介与意义/Introduction

1.1 项目意义和依据/Significance

数字媒体时代，无论是直播平台还是普通视频网站，足球比赛的视频记录都几乎数不胜数。将基于深度学习的人体姿态估计应用于足球比赛视频，获取简练直观、可量化的由人体骨骼关键点表达出的2D乃至3D动作信息，可大大提高视频资料的利用率。于足球运动本身而言，球队日常训练与战术编排中，处理后的动作数据可作为重要的参考资料；并且，对于此时正值起步阶段、正蓬勃发展的虚拟现实技术而言，通过学习抽象、简洁的姿态数据，能够更好地完成足球运动员模型的动作模拟。

然而，原始的足球比赛视频中，由于足球场地大、运动员人数多、跑动快，存在着运动员大小不一、入镜不完整、身影重叠等情况，为确定2D人体骨骼关键点、估计多人姿态、追踪运动员动作带来了一定的阻碍：

在单帧画面内，现有的基于图像的多人姿态估计方法主要分为两大流派，自上而下（Top down）和自下而上（Bottom up）。自上而下方法步骤为先检测出图像中的所有人体，逐一将之缩放至同一大小后，再进行人体骨骼关键点定位，从而获得该图像（或该帧）的人体动作信息；而自下而上方法则率先检测出画面中所有可能的人体骨骼关键点，无需进行人体图像的框定截取与大小调节。虽说自下而上的方法胜在快速与轻量级，但难以适应足球比赛复杂的实际场景。与之相较，自上而下方法更为成熟，在估计精度方面亦优于前者，因而团队选用自上而下的方法--lighttrack模型进行姿态估计。

Lighttrack将视频人体姿态估计任务划分为多人姿态估计和人体姿态跟踪，多人姿态估计遵循自上而下策略先后经过人体检测器与人体姿态估计器模块，获得场景中多人的骨骼关键点；人体姿态追踪则引入前帧画面的关节点信息，计算空间偏移距离、比对动作相似度后输出是否为同一人体的判断，以此为依据对视频内人员进行连续画面的序号标注。并且团队在复现lighttrack的基础上，对人体检测器候选算法以及人体姿态跟踪匹配算法进行改进：设计Soft NMS算法提高人体检测候选框的召回率、设计人体候选框大小动态调节算法消除背景观众的干扰、设计采样关键帧帧间隔动态调节算法提高人体姿态跟踪准确率。

团队还将尝试根据前述神经网络获得的2D人体骨骼关键点，预测3D人体动作，将输出的足球运动员动作数据扩展到三维立体空间。人类可以通过直接观看2D视频获得信息，是因为人类具有通过深度理解空间分布的能力，而将空间推理能力赋予机器对虚拟现实和增强现实等应用有很大的推动作用。这个技术可以辅助机器的动作理解、动作修复、人机交互和动作捕捉。

## 1.2 本方法/系统框架/Article Structure

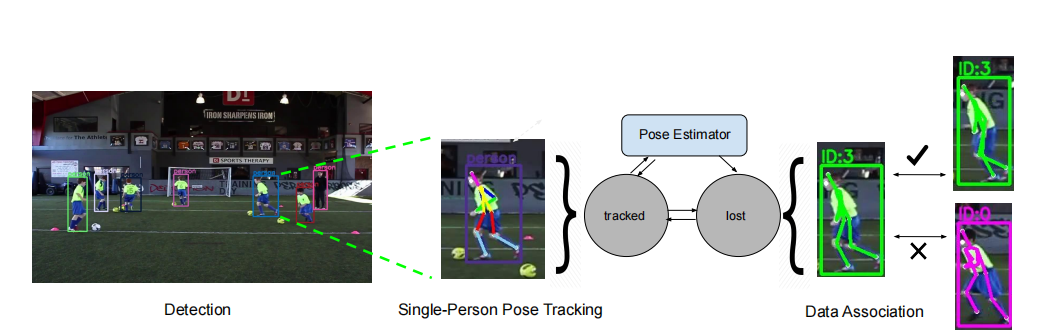


图1 Lighttrack 2D人体姿态估计与追踪

Part1. Lighttrack2D人体姿态估计与追踪：视频数据首先逐帧送入改良后的人体检测器模块，获得选中人体目标框后，在人体姿态估计器内对各框中的人体进行骨骼关键点定位，之后根据与前帧姿态估计的比对结果对人员进行身份确认与ID标注。三模块均含有各自的神经网络训练模型（YOLOv3、MobileNetv1-Deconv、Siamese GCN），并通过相应选择算法连接。

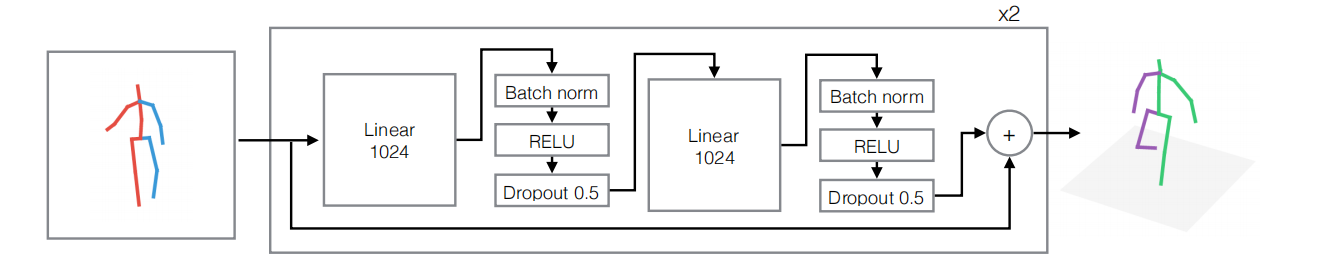


图2 2D人体关键点预测3D人体关键点

Part2.2D人体关键点预测3D人体关键点：网络的基本单元依次由线性层、批标准化、线性整流函数和随机失活组成，两个基本单元通过残差连接，形成一外层单元。整个网络由重复两个外层单元构成，输入2D人体关键点，输出其3D的预测结果。

# 相关工作/Related Works

## 2.1 基于图像的多人姿态估计/Multi-person Pose Estimation in Images

深度神经网络应用于多人图像中人体骨骼关键点的预测由来已久，且成果颇丰[4, 5, 6, 7]，其方法无外乎采用自上而下或自下而上这二者之一。自下而上预测方法，即不预先检测画面中的人体而直接进行关键点标定；自上而下预测方法，即预先确定画面中的人体、缩放至同一大小后进行关键点标定。

具体至自下而上（Bottom up）策略，检测出图像的所有人体关节点，根据人体结构先验对检测出的关节点进行重新组合，使得人体与自身关节点正确匹配，进而实现对每个人的姿态估计。Pishculin[24]采用联合子集划分和标记，首先使用CNN提取身体部位候选区，将每一个候选区域所对应的一个关节点作为一个结点，所有的这些候选节点组成一幅完整的节点图，再用整数规划，将同一个人的关节点归为一类。对检测出来的结点进行标记，确定结点属于人体的哪一个部分。最后通过深度神经网络对属于同一人体的关节点和各关节点的标记类型进行结合。DeeperCut[5]将多人姿态估计问题重新定义为分块与标签的整数线性规划公式，通过求解上述公式实现目标，该方法依赖于强大的人体检测器，但由于人体检测器缺乏全局信息，在较多人的姿态估计中表现欠佳。OpenPose[4]则是引入了人体部位关联度域（Part Affinity Fields）来连接单个人体的各个部位，但在人数众多时效果仍不甚理想。

而自上而下（Top down）策略则是检测出图像中的所有单个人体，对单个人体进行关节点检测，进而估计出每个人的姿态。[25]通过人体检测器检测出图像中的人体，生成一组联合候选框，然后对每个检测到的人体，使用卷积姿态机进行担任姿态的估计。对于关节点被遮挡以及复杂背景下这一特征，[26]采用人体检测框架得出人体候选框，再对每一个人体后选中的人体关节点进行回归，得出人体姿态。通过使用多阶段流水线[6, 7]，解除直接标定人体骨骼关键点的局限性，将多人姿态估计问题划分为人体检测、单人姿态估计以及后续处理操作三阶段，分别利用成熟的人体检测器与姿态估计网络，极大优化了场景较拥挤时的估计效果，更能适应足球比赛的具体应用场景。

## 2.2 基于视频的多人动作追踪/Multi-person Articulated Tracking in Images

将姿态估计从图像延展到视频时，首先需要在单帧层面上进行对人体骨骼关键点的预测，再将不同帧之间的关键点在时空域上进行连接。[8, 9]展示了在应对复杂视频上依旧突出的表现，实现了相邻帧间的平滑处理以及对同一人物的辨认，但难以适用于长视频。[10]团队提出了使用带有时空边界的条件随机场的优化方案，提高了单帧的预测精度，但难以识别同一人体，因而无法进一步解决动作追踪问题。

于是，优化考虑在对视频的目标追踪上展开。在深度学习算法广泛使用之前，相对成熟的方法[11]是利用运算复杂度较高的人工标注的特征，搭建多阶段的流水线系统，在各阶段逐一优化。现如今有了深度学习的帮助，端到端的步骤得以归并整合，[12, 13]使用RNN提取潜在的多样视觉线索，如外形、移动，来追踪目标。在论文[14]中，追踪器因引入了连续两帧间相关性的特征而性能优越，不仅实现了单帧内候选物体的框定，还反映出了该目标随时间的变形情况。目前，端到端模式的人体动作检测与追踪已有诸多成果[15, 16, 17]，受此启发，本项目的人体动作追踪将在深度学习目标追踪的基础上进行。

# 研究内容与方法/Contnts and Methods

* 1. Lighttrack实现方法/Methods of Lighttrack

基于人体检测和视频理解的最新进展，团队寻找到了一个轻量且有效的方法lighttrack，进行迁移学习来研究复杂多人视频中人体关键点的估计和跟踪问题。Lighttrack模型将视频人体姿态估计及人体姿态追踪拆分为三大模块：人体检测器（Detector）、人体姿态估计器（HPE）、人体姿态跟踪（Data association），在各模块内施用现有最优解模型，并设计相应选择算法组合之，以期达到整体最优的效果。原项目分别使用YOLOv3和MobileNetv1-deconv（YoloMD）作为人体目标检测器和姿态估计器，其在PoseTrack'18验证集上平均达到2 FPS，70.4 mAP和MOTA得分55.7％。

第一环节：人体检测器（Detector）

复现该阶段时，团队选用原项目选取的YOLOv3目标检测器，其核心是Darknet-53，具体神经网络架构如下图所示：

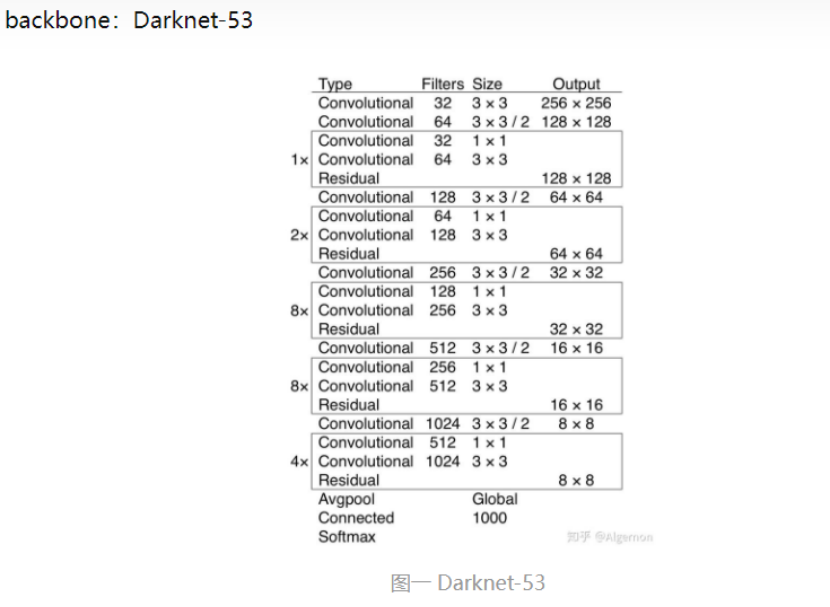


图3 Darknet-53网络结构

训练出的YOLOv3 detector并非施用于视频中的每一帧，而只对指定的关键帧进行人体检测。获得所有候选边界框后第二步进行非极大值抑制（non-max-suppression）处理，因为目标检测在同一位置上可能产生多个候选框，原始候选框间常存在大量重叠，需对其进行选择。

选择的核心在于搜索局部最大值，抑制非极大值，以挑选出最合适的目标边界，消除冗余边界框。具体方法为对候选边界框的置信度排序后，选择置信度最高的从边界框列表中删除、添加到最终输出列表中，随后计算其余候选框与该选中候选框的IoU， 删除IoU大于阈值的边界框。重复上述过程直至边界框列表为空。

第二环节：人体姿态估计器（HPE）

获得的符合条件的人体目标检测框即意味着框内（极有可能）存在一完整人体形态，根据自上而下的姿态估计策略，随后需逐个对目标框内的单一人体进行骨骼关键点的定位。实现流程为从边界框图像中得到放映关键点存在概率的热力图，再根据热力图结果确定关键点的落点。

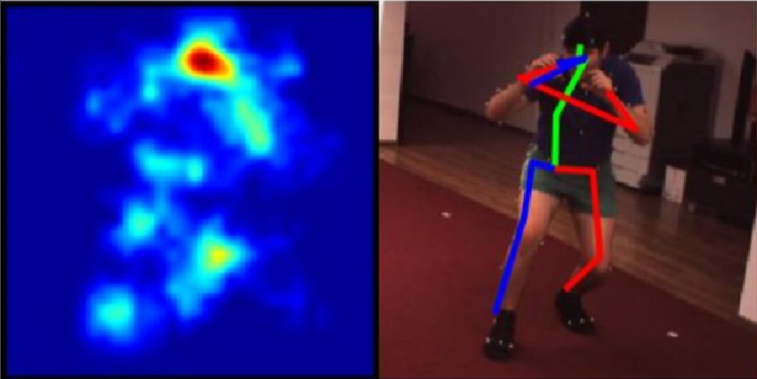


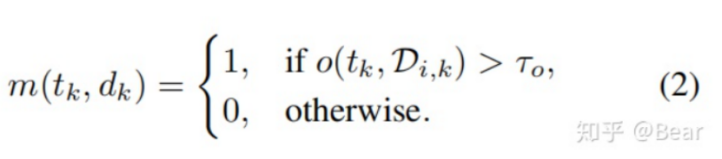
图4 HPE工作原理示意

原项目选择使用MobileNetv1-Deconv人体姿态估计器实现，MobileNetv1-Deconv是为移动和嵌入式设备提出的高效模型，MobileNets基于流线型架构(streamlined)，使用深度可分离卷积(depthwise separable convolutions)来构建轻量级深度神经网络。

第三环节：人体姿态跟踪（Data association）

在对视频中人体跟踪匹配的过程中，原项目提出Re-ID模块，在追踪多个移动人体时保留与更新他们的身份。其中，身份更新的操作将在画面中有人员被检测为丢失（lost）时发生，lost可能由足球视频运动员被遮挡、移动出画面或整体镜头视角切换等等原因导致。

其判断方式为从前一帧的人体关键点估计结果预测后一帧的人体边界框，将边界框每一侧的ROI区域扩大20%，将之视为该人在后一帧出现的局部区域进行关节估计，倘估计结果各关键点的平均置信度低于阈值，则判定为目标丢失。

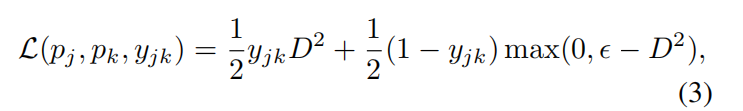


(1)

处理目标丢失，原项目选择将发生lost后的下一帧动态设置为关键帧，在关键帧时人体检测器自动被调用，可重新获得所有人体的边界框，以保证多人目标的追踪准确度。

对于同一人体的身份匹配，原项目考虑空间偏移距离以及姿势相似度进行实现。以相邻帧两边界框的IoU判断空间一致性，IoU高者表示两目标实际位置重合度高；以人体骨骼关键点形式输入的卷积结果反映姿态的匹配程度，在此处原项目引入又一神经网络Siamese GCN完成。

Siamese GCN的输入为人体骨骼关键点与相应骨架连结的图的联合坐标向量，类似CNN，进行若干卷积操作后提取出表征人体姿态的特征向量，蕴含人体关节间的空间关系；输出为两姿态的相似度，用以判断是否匹配。其网络由两个GCN层与一个使用如下边际对比损失函数的卷积层构成：



(2)

其中，为两 norm范数归一化潜在表示的欧氏距离，∈{0,1}为和是否为相同姿态的标签。

* 1. 人体姿态追踪改进/Improvement in Human Pose Tracking

针对原论文的复现效果，输出的足球比赛视频在人不重叠的情况下预测效果良好，但仍具有两个主要问题，一是一个人的关键点分离在不同的人身上，造成节点乱飞的情况；二是在人体重叠的时候，会失去一个人的姿态及关键点跟踪。对于这些问题，本小组进行了如下改进与创新。

改进一：添加边界框不合法判断

针对第一个问题，本小组判断是边界框的预估错误导致，通过输出所有边界框的大小及中心点，发现了异常值。根据异常情况，本小组添加了边界框不合法的判断函数，首先对它的长宽范围进行限制，其次保证边界框的中心点都为非负。

改进二：改进目标检测中的传统非极大值抑制(nms)为soft-nms

针对第二个问题，本小组判断是在非极大值抑制把重叠的另一个人的边界框误判为当前边界框的冗余框直接删除了。通过文献调查，本小组将传统nms改成soft nms，可以提升目标检测的平均准确率。区别在于如果两个框的IoU大于阈值，另一个人的边界框的分数不会直接设置为0，而是将分数设置为一个衰减函数，目的是给这个边界框被保留的机会。



(3)

改进三：动态调整关键帧间隔

针对重叠问题，本小组同时动态调整了关键帧的间隔帧数。如果关键帧的间隔数过大，会导致关键帧的判断错误延续到之后较长时间的非关键帧，重新检测的几率较低。如果关键帧的间隔数过小，关键帧检测错误会影响目标的跟踪，导致非关键帧无法保留关键帧的正确检测数量，所以关键帧的间隔数目非常重要。因此，本小组在非关键帧出现目标丢失情况下将关键帧间隔数降低，使得它能快速重新进行正确的检测，在恢复正常后，再将关键帧间隔数调回原来的值，使得有更长的非关键帧保留正确的检测。

3.3 2D姿态估计预测3D姿态估计/3D Prediction

为了获取3D关键点，本文将其拆解成两个问题：2D人体姿态估计和从2D关键点预测3D关键点。

关于2D关键点预测3D关键点，本问题简化成数学公式便是将一系列2维空间的点输出为3维空间一系列的点，目标是学习到一个函数，降低预测的三维坐标与实际坐标之差。

(4)

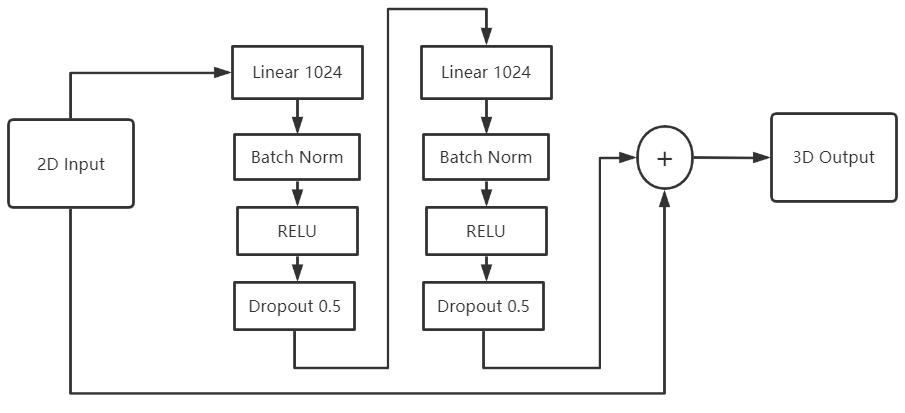


图5 获得3D姿态估计流程图

本文在实现中x是上文经过人体姿态估计转换后的2D坐标，函数运用深度神经网络。图3展示了本文基础的网络架构，包括线性层、批标准化、随机失活、线性整流函数和残差连接。

输入输出：本小组选择2d点作为输入，3d作为输出，因为2d的低维度可以使本小组在训练网络时，很容易地将整个数据集存储在GPU中，并减少整体训练时间，降低设计网络和训练超参数的搜索时间。

线性整流函数层：通过相关工作的文献检索得出，大多数用于3D人体姿态估计的深度学习方法都是基于卷积神经网络，比如应用于整个图像的过滤器，或者二维的关节位置热图。然而，由于本小组将低维点作为输入和输出，本文选择更简单、计算成本更低的非线性层——线性整流函数层，作为在深度神经网络中加入非线性层的标准选择。

残差连接：残差连接可以加快卷积神经网络的训练速度，提升泛化能力和降低训练时间。

批标准化和随机失活：批标准化可以提高网络性能，在训练和测试时间上有所降低。

最大范数限制：本文在每层的权重中都加入了最大范数（max norm）(小于等于1)。再结合批标准化，可以使训练更稳定，且使泛化能力增强。

# 实验结果与分析/Experiment Results and Analysis

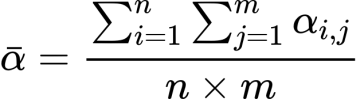
4.1 2D人体姿态估计定量分析/2D HPE Quantitative Analysis

由于专门基于足球比赛视频的人体姿态估计数据集较少，本小组多方搜寻无果后，决定采用迁移学习的方法，利用COCO以及PoseTrack数据集下已训练好的人体目标识别检测器Yolov3的神经网络模型参数来代替在足球比赛视频上的训练过程，同时将搜寻到的经过处理的足球比赛视频作为测试集，评估本小组模型的效果。

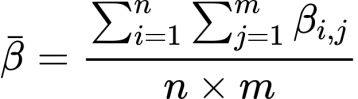
本小组采用的足球比赛测试集经过了以下处理：一是控制视频时长在十五秒内，并且各个视频的时长大致相同，从而保证了后续抽样检测量化分析时的公平性；二是视频中必须包含至少一次的场景快速转换，以检测目标追踪的连续性；三是视频中必须包含至少一次的人体重叠，以检测目标追踪的连续性、是否漏检以及错检；四是视频中至少有三秒以上出现大面积观众背景的情况，以检测模型是否能正确排除背景观众的干扰。

由于所要求的足球场景视频并无带有标记的合适数据集，改进效果评估将采用小数据量的人工检测。本小组采用的人工检测方法是抽样检测，即随机从一个视频中抽取十帧，组成一个样本帧集。若这十帧中不包含场景快速转换、人体重叠或大面积观众背景这三种情况中的一个，则重新抽取十帧；若同一视频抽取三次都没有采样到符合要求的样本帧集，则舍弃该视频，对下一个视频进行采样。直到采样到五个样本帧集，采样结束。

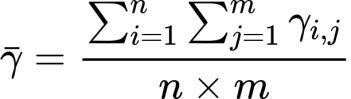
对于模型的表现效果，我们设置了三个检测指标：检测遗漏、估计错误和追踪错误。其中检测遗漏是指每帧中未被检测到的目标人体数。估计错误是指骨骼关节点标错的情况，包括正确检测到人体边界框但错误标识其关节点以及错误检测人体边界框从而错误标识关节点这两种情况。追踪错误是指同一个人体在相邻两帧之间分配了不同的ID，同一个人的一次分配错误计数为1。最后将以上三个指标在整个样本空间内求平均，得到每帧中的错误数。设样本帧集的个数为 IMG_256，样本帧集中帧的个数为 IMG_257，第 IMG_258个样本帧集中第 IMG_259帧的检测遗漏数、估计错误数和追踪错误数分别是 IMG_260、 IMG_261和 IMG_262，则错误数的平均评估指标的计算公式如下所示：



(5)



(6)



(7)

最终的评估结果为：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 检测遗漏（IMG_256） | 估计错误（IMG_257） | 追踪错误（IMG_258） |
| 改进后 | 3.1 | 0.9 | 2.4 |
| 改进前 | 4.8 | 2.2 | 0.2 |

表1 改进前后效果评价

从表格中可以看出，经过Soft NMS方法改进后，检测遗漏数有了明显的降低，说明Soft NMS的方法较原有的标准NMS方法更好，能有效提高边界框的召回率，提高重叠人像的边界框检测率；经过对边界框的大小调整后，估计错误也有了大幅度的降低；经过对关键帧间隔数的动态调整后，模型的追踪错误上升，原因是减少关键帧的间隔数后，ID分配的频率增加。由于ID分配是随机分配的，这会使得追踪错误增加。在今后的改进中，本小组可以针对追踪错误进行优化，比如在关键帧中也要将前一帧的各个目标人体纳入考虑。更一般地，可以针对每个目标人体建立时空姿态轨迹图，使得模型可以根据时空轨迹图进行预判，从而在降低追踪错误的同时，也能提高模型的鲁棒性。

4.2 2D 人体姿态估计定性分析/2D HPE Qualitative Analysis

模型改进前对于足球比赛视频的人体关节估计如下图所示：

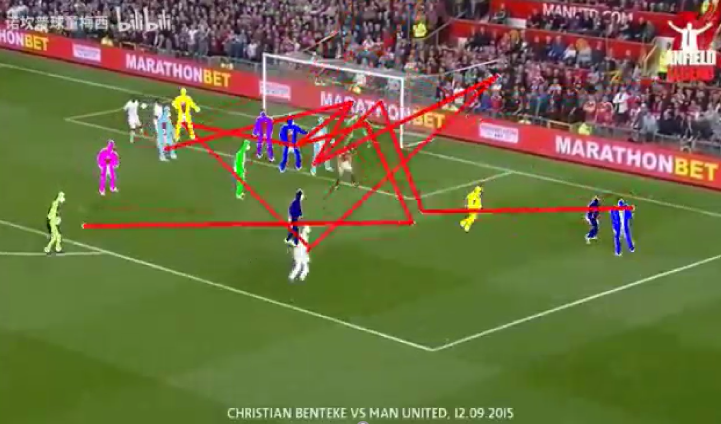


图6 改进前视频某帧姿态估计输出结果

从图中可以看出，原模型错把观众背景当作人体关节点，导致错误关节点几乎占据整个屏幕。并且存在大量漏检的情况。另外，在人体重叠时，漏检情况更为严重。

模型改进后对于同一帧足球视频的人体关节估计情况如下图所示：

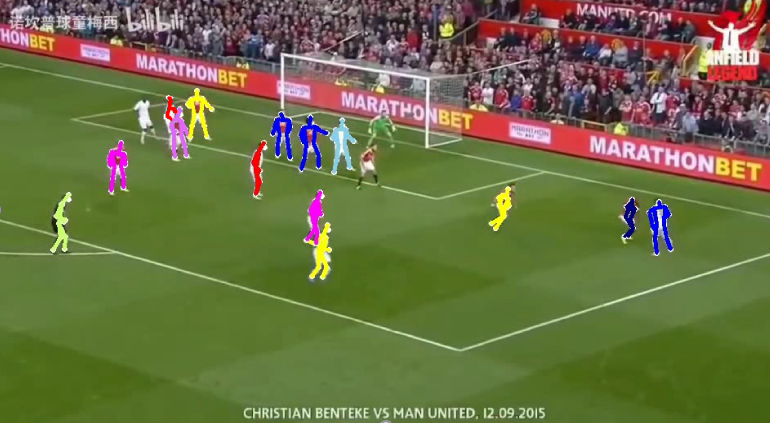


图7 改进后视频该帧姿态估计输出结果

从图中可以看出，经过边界框的大小调整之后，错检的情况明显好转；另外由于Soft NMS方法的运用，重叠人体也能在一定程度上完成姿态估计。

但是，原模型和改进之后的模型都存在漏检的情况，说明本小组的目标人体检测器Yolov3还需要改进。经查阅资料，发现Yolov5已经出现。这给本小组之后的优化指明了方向，可以通过迭代升级目标人体检测器来降低检测遗漏错误数。

4.3 3D 人体姿态估计分析/3D HPE Analysis

当前基于深度学习估计3D人体姿态较为成熟的方法通常需经由2D人体姿态估计实现，即利用2D人体关键点作为输入数据训练神经网络，预测3D人体关键点信息。团队期望将前述实验成果进一步扩展到3D，但由于3D-pose-baseline过分赖Human3.6m数据集标签，泛化性较差，在缺乏相应场景训练数据集的情况下难以适应足球视频的应用场景。加之课程项目时间条件、硬件设施不足，将2D人体姿态估计结果迁移至3D空间的尝试并未成功，有望在日后予以实现。

# 特色与创新/Distinctive or Innovation Points

对比现有的其他算法，本小组选择的Lighttrack模型是在线轻量的，具有实时性，速度更快，保证了足球比赛视频生成的更高帧速率。同时，它是自上而下的在线跟踪框架，通用性强，适合其他姿态估计器和候选匹配机制，可以更快更准确地替换单个组件。

从人体姿态追踪入手，团队创新性地提出了三项算法上的改进，极大提升了整体效果：

一、设计人体候选框大小动态调节算法消除背景观众的干扰

二、设计Soft NMS算法提高人体检测候选框的召回率

三、设计采样关键帧帧间隔动态调节算法提高人体姿态跟踪准确率

1. **致谢/Acknowledgement**

在此，诚挚感谢盛斌教授的优秀教导，为成员打开了计算机图形学的大门；诚挚感谢易冉老师、所向助教以及其他所有助教老师的耐心指导，在必要时为成员指点迷津；最后感谢在项目过程中施以援手的友善同学。

**参考文献**

1. G. Ning, J. Pei and H. Huang. LightTrack: A Generic Framework for Online Top-Down Human Pose Tracking. In CVPR, 2020.
2. R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri and D. Tran. Detect-and-Track: Efficient Pose Estimation in Videos. In CVPR, 2018.
3. Y. Xiu, J. Li, H. Wang, Y. Fang and C. Lu. Pose Flow: Efficient Online Pose Tracking. In BMVC, 2018.
4. J. Martinez, R. Hossain, J. Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In CVPR, 2017.
5. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multiperson 2d pose estimation using part affinity fields. In CVPR, 2017.
6. E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multiperson pose estimation model. In ECCV, 2016.
7. Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. arXiv preprint arXiv:1711.07319, 2017.
8. Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In ICCV, 2017.
9. E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, B. Andres, and B. Schiele. Articulated multi-person tracking in the wild. In CVPR, 2017.
10. U. Iqbal, A. Milan, and J. Gall. Pose-track: Joint multi-person pose estimation and tracking. In CVPR, 2017.
11. J. Song, L. Wang, L. Van Gool, and O. Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In CVPR, 2017.
12. S.-I. Yu, D. Meng, W. Zuo, and A. Hauptmann. The solution path algorithm for identity-aware multi-object tracking. In CVPR, 2016.
13. A. Milan, S. H. Rezatofighi, A. R. Dick, I. D. Reid, and K. Schindler. Online multi-target tracking using recurrent neural networks. In AAAI, 2017.
14. A. Sadeghian, A. Alahi, and S. Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In ICCV, 2017.
15. C. Feichtenhofer, A. Pinz, and A. Zisserman. Detect to track and track to detect. In ICCV, 2017.
16. R. Hou, C. Chen, and M. Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In ICCV, 2017.
17. K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, et al. T-CNN: Tubelets with convolutional neural networks for object detection from videos. arXiv preprint arXiv:1604.02532, 2016.
18. K. Kang, W. Ouyang, H. Li, and X. Wang. Object Detection from Video Tubelets with Convolutional Neural Networks. In CVPR, 2016.
19. K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In ICCV, 2017.
20. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.
21. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
22. Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. PAMI, 2013.
23. J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In CVPR, 2017.
24. C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal residual networks for video action recognition. In NIPS, 2016.
25. Pishchulin L, Insafutdinov E, Tang S Y, et al. DeepCut: joint subset partition and labeling for multi person pose estimation. In CVPR, 2016.
26. Iqbal U, Gall J. Multi-person pose estimation with local joint-to-person associations. In ECCV, 2016.

[27] Chen Y L, Wang Z C, Peng Y X, et al. Cascaded pyramid network for multi-person pose estimation. In CVPR, 2018.