

基于计算机视觉技术的人体姿态跟踪方法

刘 恒¹, 高新燕², 苏新新², 李成龙¹

(1. 山东建筑大学 计算机科学与技术学院; 2. 山东山大华天软件有限公司, 山东 济南 250101)

摘 要: 人体姿态跟踪在人体运动分析、智能监控、虚拟现实等应用场景中得到了广泛应用。为解决在多人姿态跟踪过程中因多人遮挡、复杂环境等原因导致的目标易丢失、姿态估计不准确等问题,提出一种结合DeepSORT和OpenPose的人体姿态跟踪方法。选择合适的数据集分别训练人体运动跟踪模型和人体姿态估计模型。在人体运动跟踪阶段,基于改进的YOLOv5与改进的DeepSORT算法进行准确地识别与跟踪,减小了模型内存访问成本,引入模型剪枝方法进一步提高了跟踪速度。在人体姿态估计阶段,基于跟踪阶段获得的准确目标,利用OpenPose算法实现了目标人体的姿态估计。相较于对比方法,改进算法在COCO2017数据集上的FLOPs为4.46 G,检测速度大幅提升,在MOT16数据集、PoseTrack2017数据集上的运行时间有较大优势,分别为85.55 s、81.79 s。

关键词: 人体姿态跟踪; 目标检测; YOLOv5; DeepSORT; OpenPose

DOI: 10.11907/rjtk.241058

开放科学(资源服务)标识码(OSID):



中图分类号: TP391.41

文献标识码: A

文章编号: 1672-7800(2025)004-0136-11

Human Pose Tracking Method Based on Computer Vision Technology

LIU Heng¹, GAO Xinyan², SU Xinxin², LI Chenglong¹

(1. School of Computer Science and Technology, Shandong Jianzhu University;

2. Shandong Hoteam Software Co., Ltd., Ji'nan 250101, China)

Abstract: Human pose tracking has been widely used in application scenarios such as human motion analysis, intelligent monitoring, and virtual reality. To address the issues of target loss and inaccurate pose estimation caused by multiple occlusion and complex environments in the process of multi person pose tracking, the article proposes a human pose tracking method that combines DeepSORT and OpenPose. Select appropriate datasets to train human motion tracking models and human pose estimation models separately. In the human motion tracking stage, based on the improved YOLOv5 and the improved DeepSORT algorithms, accurate recognition and tracking are carried out, reduced the model memory access cost, and model pruning methods are introduced to further improve tracking speed. In the human pose estimation stage, based on the accurate target obtained in the tracking stage, the OpenPose algorithm is used to achieve the pose estimation of the target human body. Compared with the comparison method, the improved algorithm has FLOPs of 4.46 G on COCO2017 dataset, greatly improving the detection speed, and has a greater advantage in the running time on MOT16 dataset and PoseTrack2017 dataset, which are 85.55 s and 81.79 s respectively.

Key Words: human pose tracking; object detection; YOLOv5; DeepSORT; OpenPose

0 引言

近年来,隐式人机交互与以人为中心的计算已经成为

计算机科学研究热点方向之一^[1]。人体姿态跟踪技术可以提供用户的位置朝向等基本数据,是人类行为分析和识别的基础,已广泛应用于肢体语言理解、视频自动标注、运动评估、自动驾驶等领域^[2-4]。在运动障碍疾病诊断与

收稿日期: 2024-03-26

扫描二维码阅读全文:



基金项目: 国家自然科学基金项目(62102235); 山东省自然科学基金项目(ZR2020QF029); 山东建筑大学博士基金项目(XNBS1811)

作者简介: 刘恒(1999-),男,CCF学生会会员,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为计算机视觉;高新燕(1987-),女,硕士,山东山大华天软件有限公司高级工程师,研究方向为计算机图形学;苏新新(1982-),女,硕士,山东山大华天软件有限公司高级工程师,研究方向为计算机图形学;李成龙(1988-),男,博士,CCF会员,山东建筑大学计算机科学与技术学院副教授、硕士生导师,研究方向为计算机视觉、增强现实。本文通讯作者: 李成龙。

治疗以及科学运动等领域,人体姿态跟踪技术可以多维度地量化监控并分析关节的自主运动和被动联动,能够有效地解决传统人工肉眼估计产生的效率低、误差大及标准不统一的问题;在自动驾驶方面,人体姿态跟踪技术可用于自动驾驶汽车的人体捕捉。当前,自动驾驶逐渐进入人们的生活中,对于自动驾驶汽车而言,为了避免危险的发生,不仅需要精准地检测到前方周围的人体,还需要精准地预测估计并且理解人体姿势的下一步行为。

在人体姿态跟踪方面,Michael^[5]引入基于形状模型的人体姿势跟踪方法,为后续姿势跟踪研究奠定了基础。人体姿态跟踪方法经历了3个阶段:①基于机器学习的方法;②基于手工特征的方法;③基于深度学习的方法。

目前,人体姿态跟踪技术在实际应用中还面临着一些问题,主要包括对光照和背景的依赖度、对姿态变化和多人场景的适应性、精度问题等。因此,如何提高姿态跟踪技术对干扰的抵抗力,提升算法的鲁棒性是亟待解决的问题。

针对上述问题,本文提出一种结合DeepSORT和OpenPose的人体姿态跟踪方法。此方法在人体运动跟踪阶段,基于改进的YOLOv5和改进的DeepSORT实现对运动人体的准确识别与跟踪,减小了模型内存访问成本,降低了计算量,实现了模型轻量化,并引入基于BN层的模型剪枝方法,以牺牲较小精度的方式大幅提高了模型检测速度。在人体姿态估计阶段,利用OpenPose算法实现目标人体的姿态估计。在实验阶段,采用优秀的公开数据集用于人体运动跟踪模型和人体姿态估计模型训练,以保证数据多样性。为评估该算法的性能,选择COCO2017、MOT16、PoseTrack2017数据集进行实验验证,其中大部分为多人视频和图像^[6-7]。实验结果验证了该方法的先进性,有效解决了人体姿态跟踪过程中因多人遮挡、意外姿势等引起的目标丢失等问题,能够处理复杂环境下的人体运动,并提高人体姿态跟踪速度、准确度和稳定性。

1 相关工作

计算机视觉技术包含图像分类、目标检测、目标跟踪、语义分割、物体分割和姿态估计等技术。本文实现人体姿态跟踪主要用到了目标检测技术、目标跟踪技术和姿态估计技术,利用目标检测技术实现对运动中人体目标的检测,再通过目标跟踪技术对运动中的人体进行追踪,最后结合姿态估计技术完成对运动中的人体姿态跟踪。

1.1 目标检测算法

目标检测是计算视觉领域非常热门的研究方向,它主要解决一幅图像中有什么物体,以及物体在什么位置的问题,在人体检测、人脸识别等领域有非常广泛的应用。

近年来,目标检测算法由最初的特征提取加分类器的算法发展到了深度神经网络的目标检测算法。目前,主流

的目标检测算法主要分为两阶段目标检测算法(Two-Stage)和单阶段目标检测算法(One-Stage)。

两阶段目标检测算法先获得目标候选区域,再通过候选区域实现对目标位置和类别的预测。常见的两阶段目标检测算法有R-CNN、Fast R-CNN、Faster R-CNN、Mask R-CNN等,这类算法检测精度较高,但速度较慢^[8-11]。

单阶段目标检测算法不需要进行候选区域的提取,直接基于整幅图像预测目标位置和类别。常见的单阶段目标检测算法有YOLO、SSD等,这类算法速度较快,但在精度方面稍弱于两阶段目标检测算法^[12-13]。

由于人体目标检测对实时性要求较高,多数研究都基于单阶段目标检测算法。其中,单阶段目标检测算法中的YOLO系列算法已经由YOLOv1逐步发展到了YOLOv8,在性能上逐步得到提升。

1.2 目标跟踪算法

目标跟踪算法利用了图像在时间上的连续性、空间上的相关性,可通过图像序列中当前图像及前几帧的图像及位置信息获取当前目标的位置。与目标检测算法相比,其可充分利用帧间信息,有利于更快速、稳定地提取目标。根据追踪任务和应用场景不同,又可以分为单目标追踪(Single-Object Track, SOT)和多目标追踪(Multi-Object Track, MOT)。

在单目标跟踪中,使用给定的初始目标位置,在后续视频帧中对给定物体进行位置预测。目前,单目标跟踪算法主要方法分为两大类:①基于相关滤波的跟踪算法,如CSK^[14]、KCF^[15]等;②基于深度学习的跟踪算法,如SiamFC^[16]、SiamRPN^[17]等。相比之下,基于相关滤波的跟踪算法速度更快,基于深度学习的跟踪算法准确性更高。

多目标跟踪算法大多不考虑初始目标位置,目标可自行消失与产生。常见的多目标跟踪算法有Sort^[18]、DeepSORT^[19]、TR-MOT^[20]等。

1.3 人体姿态估计算法

根据从哪个级别(高级抽象或低级像素)开始计算,人体姿态估计方法可分为自顶向下的方法和自下而上的方法。自顶向下的人体姿态估计方法有两个最重要组成部分,即人体区域检测器和一个单人姿态估计器。大多数研究专注于基于现有人体检测器的人体姿态估计,例如Faster R-CNN、Mask R-CNN。自下而上的方法直接预测所有人的所有关节,然后将它们组装成独立的骨架,在复杂环境中正确组合关节是一项艰巨的研究任务,代表算法有OpenPose^[21]、DeepLabCut^[22]。目前,自下而上方法的处理速度非常快,有些方法可以实时运行,但是性能可能会受到复杂背景和人为遮挡的影响。自顶向下的方法在几乎所有基准数据集中都实现了相对更先进的性能,而处理速度受检测到的人员数量限制。

1.4 人体姿态跟踪算法

人体姿态跟踪方法与多人姿态检测一样,也分为自顶

向下和自下而上。自顶向下的方法是在每帧中检测人的边界框,然后检测关键点,最后确定相邻帧相似性,继而跟踪整个视频,例如 LightTrack^[23]、Pose Flow^[24]、BlazePose^[25]。自下而上的方法是在每帧中生成关键点候选点,然后根据时空图求解整数线性规划将时空图分为子图,最终将每个子图对应人体姿势轨迹,例如 PoseTrack^[27]。目前,自顶向下的方法在精度和跟踪速度上都优于自下而上的方法,因为自下而上的方法由于仅仅利用二阶身体部件依赖而失去了全局姿态视图,这直接导致了对关键点的模糊分配。

2 算法总体结构设计

整个算法由3部分组成:人体运动跟踪、模型剪枝和人体姿态估计。算法整体性能主要取决于目标检测速度和精度。因此,在人体运动跟踪阶段,算法基于改进的YOLOv5和改进的DeepSORT实现对运动人体目标的准确识别与跟踪,并减小模型内存访问成本,实现模型轻量化,并引入基于BN层的模型剪枝方法以牺牲较小精度的方式大幅提高了模型跟踪速度。在人体姿态估计阶段,基于跟踪阶段获得的准确目标,利用OpenPose实现目标人体的姿态估计,最终提高人体姿态跟踪速度、准确度和稳定性。

2.1 人体运动跟踪

人体运动跟踪阶段分为检测和跟踪两个方面,改进的YOLOv5充当检测器,用于检测多个人体运动目标位置并显示跟踪帧。DeepSORT是一种将预测与检测相关联的跟踪器,以实现跟踪多个人体运动目标并分配不同的ID,通过对DeepSORT模型中特征提取网络部分的改进,减小模型内存访问成本,实现轻量化。

2.1.1 YOLOv5 模型结构

YOLOv5是基于YOLOv4的一种改进的目标检测算法,其模型结构如图1所示,包括输入端、主干网络、多尺度特征融合模块和检测头部4部分。

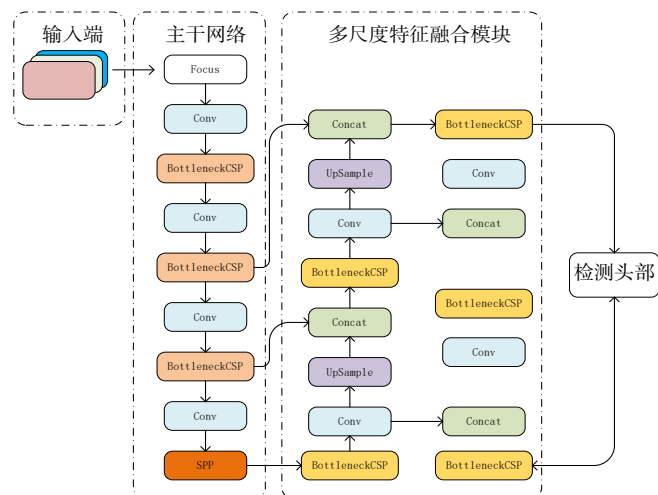


Fig. 1 YOLOv5 model structure

图1 YOLOv5模型结构

(1)输入端。加入了Mosaic数据增强、自适应锚框计算、自适应图片缩放等多种数据预处理方法。其中,Mosaic数据增强利用随机缩放、随机裁剪、随机排布的方式将4张图像进行拼接,丰富了检测物体的背景,并解决了小目标数量少、分布不均造成的学习困难问题;自适应锚框计算可以根据不同的数据集计算出最佳锚框值;自适应图片缩放尽可能减少了统一图像尺寸带来的信息冗余。

(2)主干网络。它是用于提取特征的网络,主要由Focus结构、BottleneckCSP和SPP结构组成。Focus结构将高分辨率的图片切成多个低分辨率的图片,运用隔列采样+拼接的方式,检测速度得以大幅提升;BottleneckCSP由Bottleneck模块和CSP模块两部分构成,Bottleneck为瓶颈—残差结构,可以灵活调整通道数量,提高网络复杂性,CSP可以在减少计算量的同时增强网络的学习能力;SPP采用统一步长但不同尺寸的卷积核组合,实现了特征融合。

(3)多尺度特征融合模块。由特征金字塔FPN结构和路径聚合结构PAN组成。FPN自上而下传递语义信息,PAN自下而上传递定位信息,实现了网络上下信息流的融合,提高了网络检测能力。

(4)检测头部。将输出层用于预测图像特征,生成边界框和类。使用GIOU_Loss作为损失函数,可以区分出IOU相同时预测框与真实框相交情况的不同。并且,使用非极大值抑制操作,去掉多余的预测框。

2.1.2 DeepSORT 算法流程

在DeepSORT出现之前,较为流行的目标跟踪算法是SORT(Simple Online Realtime Tracking)。SORT算法基于目标检测算法得到目标框,利用卡尔曼滤波预测当前帧的跟踪边界框,然后将目标框和跟踪边界框进行IoU匹配。SORT算法通过前后两帧IoU构建相似度矩阵,因而SORT算法计算速度非常快,但是存在目标ID切换频繁的问题。为解决该问题,DeepSORT算法在接收到目标检测结果后,需要进行两级匹配,DeepSORT算法流程如图2所示。DeepSORT算法利用卡尔曼滤波器进行逐帧数据关联,预测目标轨迹,并与检测器进行级联匹配和IoU匹配。

在DeepSORT中,将检测到的结果分为确定态和非确定态,初始时均为非确定态,只有连续3帧的IoU匹配成功才能转为确定态,而转为确定态的预测结果,则会根检测框进行级联匹配。

SORT算法只进行目标框和跟踪边界框之间的IoU匹配,而DeepSORT应用CNN提取目标的特征进行数据关联,引入性能特征的余弦距离,利用匈牙利算法解决线性分配问题。级联匹配流程如图3所示。

使用级联匹配算法在一定程度上解决了长时间遮挡导致的ID切换和跟踪失败问题。该算法从检索到的所有图像和库的图像中提取图像特征,并利用马氏距离消除错误分配。并且,利用余弦距离进行排序,通过门控矩阵限

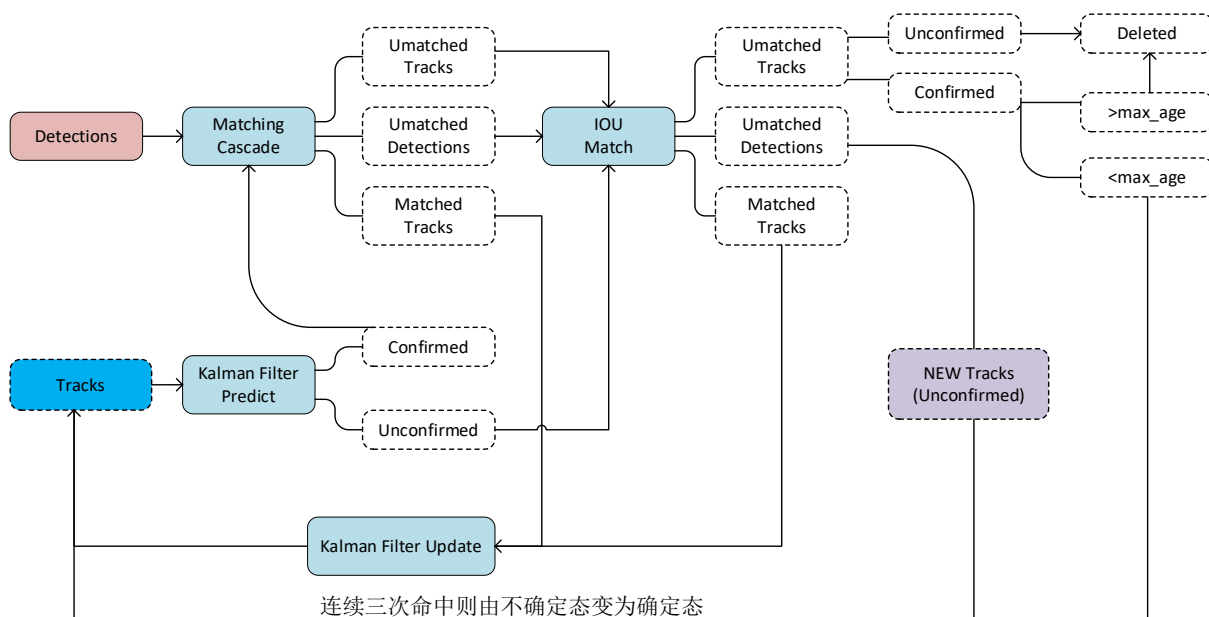


Fig. 2 Algorithm flow of DeepSORT

制代价矩阵,实现帧间同一目标的关联,有效减少了ID切换。

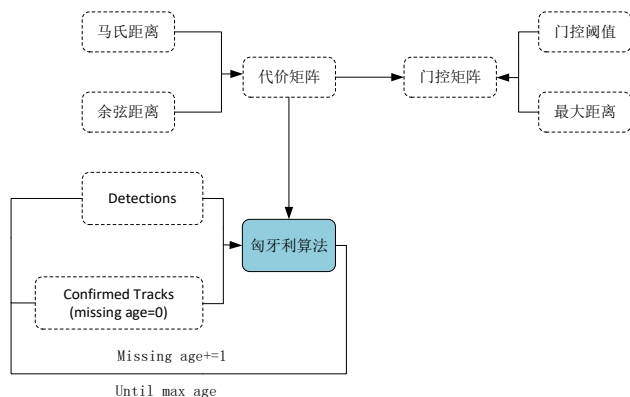


Fig. 3 Cascade matching process

2.2 模型剪枝

在人体运动跟踪过程中,算法整体性能受检测速度影响较大。随着神经网络的发展,模型参数量越来越大,包括YOLOv5在内,算法存在大量参数对最终检测结果影响微乎其微,却占据了大量的计算量和大部分内存。为进一步提高模型检测速度,本文加入了一种基于BN层的模型剪枝方法,将不重要的参数去除,保持模型表达能力不变,极大减少模型计算量。

基于BN层的模型剪枝方法选择BN层的尺度因子作为裁剪通道的衡量指标。在训练时,对BN层的尺度因子添加L1正则化,达到稀疏化的作用,再通过BN层的尺度因子趋于0以识别不重要的通道。

为加快网络收敛,防止过拟合,YOLO算法加入了BN层以对网络每一层的输入作归一化。其计算公式如下:

$$\hat{z} = \frac{z_{in} - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \quad (1)$$

$$z_{out} = \gamma \hat{z} + \beta \quad (2)$$

其中, z_{in} 为上一个卷积层的输出, μ_B 为均值, σ_B^2 为方差, 加入趋于 0 的小数 ϵ 避免分母为 0。BN 层引入了可学习重构参数 γ, β, γ 为 BN 层缩放系数, β 为平移系数, z_{out} 为 BN 层输出值。

选择缩放系数 γ 作为BN层的尺度因子,每个通道激活大小 z_{out} 与系数 γ 成正相关,如果 $\gamma \rightarrow 0$,则激活值也非常小,可以将其舍弃。损失函数为:

$$L = \sum_{(x, \gamma)} l(f(x, W), \gamma) + \lambda \sum_{\gamma \in \Gamma} g(\gamma) \quad (3)$$

其中,前一项为正常的网络损失函数, (x,y) 表示训练的输入和真值, w 表示训练权重,后一项是尺度因子的正则化, γ 是尺度因子, λ 是惩罚稀疏。

剪枝过程如图4所示。

2.3 人体姿态估计

在人体姿态估计阶段,基于跟踪阶段获得的准确目标,利用 OpenPose 实现目标人体的姿态估计。OpenPose 算法是卡内基梅隆大学(CMU)基于卷积神经网络和监督学习而开发的开源库,并在 Caffe 框架中开发,可以实现诸如人体运动、面部表情和手指移动的姿态估计,适用于单人和多人,具有良好的鲁棒性。

OpenPose 算法流程如图 5 所示,将预处理后的图像输入到 VGG-19 中以提取特征,得到一组特征图,然后将特征图输入到一个双分支网络,使用 CNN 网络预测肢体的关键点置信度图(PCM)和关键点亲和度向量场(PAF),置信度数值最大处为预测关节点所在位置,关键点亲和和字段同时保留了肢体的位置和方向信息。根据损失函数优化网

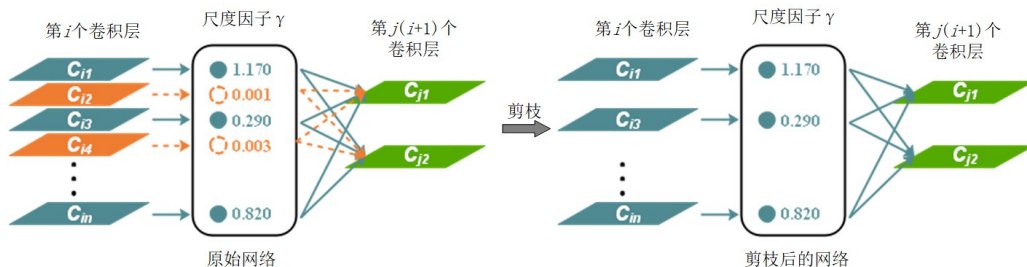


Fig. 4 Model pruning

图4 模型剪枝

络模型参数,同时将肢体部位与人体进行关联,由此可通过比较任意2个关节点的连线与人体部分亲和字段中对应线段的对齐程度选择正确的连接方式,形成人体骨架,最后在原始图像中标记出人体关键点的图像。

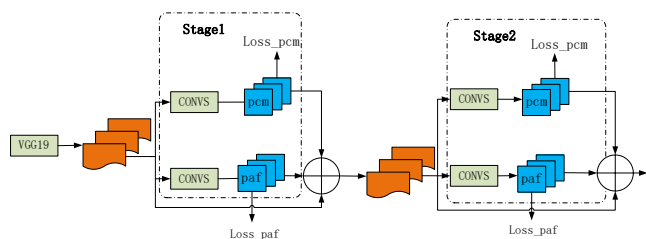


Fig. 5 Algorithm flow of OpenPose

图5 OpenPose算法流程

3 人体姿态跟踪实现

3.1 改进YOLOv5s的人体目标检测实现

YOLOv5s 作为目标检测器,通过特征提取网络从输入图像中提取特征,得到特定大小的特征图。将特征图划分为若干个 $S \times S$ 的网格单元,每个网格单元同时负责目标检测、边界框预测、定位置信度和类别概率预测,最后得出检测结果。其步骤如下:

(1)数据准备。缩放图像以获得特定大小的特征图(下采样),将特征图划分为相等的网格单元。每个网格单元被分配给需根据 Ground Truth 的 IoU 匹配进行预测的样本。

(2)卷积网络。特征图的每个网格单元预测每个类别的条件概率值。卷积网络中每个网格单元输出的预测值个数为 $B \times 5 + C$, B 为每个网格单元生成的盒子个数, C 为类数。每个框预测 5 个回归值(对象检测的属性),包括边界框的 4 个偏移坐标和 1 个置信度分数(由 IoU 表示)。

(3)后处理。多个网格单元可以检测相同的目标,使用非最大值抑制(NMS)比较多个预测框之间的 IoU,去除冗余检测并得到预测框。在人体检测中,非最大值抑制(NMS)用于获得具有最多检测信息的边界框。

YOLOv5 算法目前主要包括 4 个版本:YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x。4 个版本网络宽度、深度各不相同,检测精度和速度有一定区别。其中,YOLOv5s 网络

是 YOLOv5 系列中深度最小、特征图宽度最小的网络。为确保人体姿态跟踪速度提升,选用 YOLOv5s 网络进行改进。

YOLOv5 中, C3 模块表示一个卷积块(Convolutional Block),通常由一系列卷积层、激活函数和池化层组成。C3 模块在图像处理任务中得以广泛应用,用于提取和学习特征。C3TR 是在 C3 的基础上加入了 Transformer 注意力机制,Transformer 注意力机制是一种基于注意力机制的序列建模方法,最初被用于自然语言处理任务,如机器翻译和语言模型。它通过计算查询(query)、键(key)和值(value)之间的相关性权重,实现了全局上下文的交互和信息传递,进而可以提高模型检测精度。

将 YOLOv5s 主干网络中最后一个 C3 模块替换成 C3TR,如图 6 所示。通过此改进引入 Transformer 注意力机制,使得模型具备了更强的上下文建模能力,能够更好地捕捉输入特征之间的关联性。

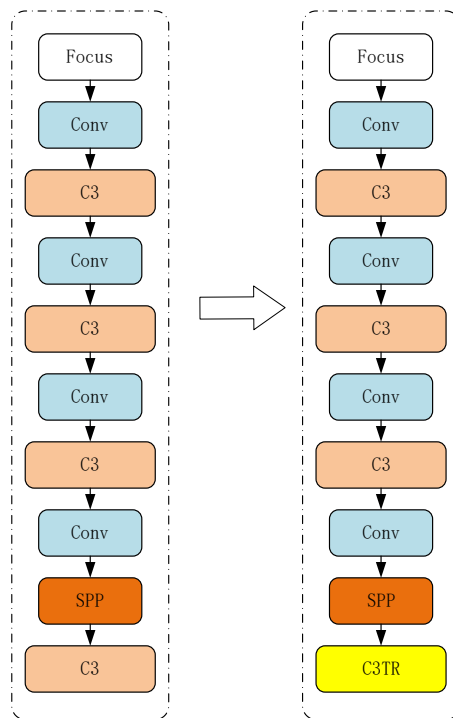


Fig. 6 Replacing the last C3 module in the YOLOv5s backbone network with C3TR

图6 YOLOv5s主干网络中最后一个C3模块替换成C3TR

3.2 模型剪枝

模型剪枝具体步骤主要分为以下3步:

(1)稀疏训练。通过稀疏训练使大多数网络权重趋近于零,方便下一步移除掉对最终网络输出精度影响小的通道。

(2)模型剪枝。它是模型剪枝的关键步骤,去除稀疏训练中识别的不重要通道。在剪枝时,可以设置剪枝率,即规定剪去多少比例的参数,不同的剪枝率决定了模型剪枝后的准确率、速度、模型大小。本文测试了4组剪枝率,分别是50%、60%、70%、80%,输入分辨率为480像素, batch_size 为16。

(3)模型微调。剪枝后模型的mAP值会降低,此时可以通过微调进行精度补偿,能够在一定程度上提升模型精度。

3.3 改进DeepSORT的人体目标跟踪实现

DeepSORT是一种将预测与检测相关联的跟踪器,该算法属于计算密集型算法,占用内存和计算量较大。通过ShuffleNetV2网络对DeepSORT特征提取网络部分进行改进,减小模型内存访问成本,实现模型轻量化。

ShuffleNetV2网络是一种高效的神经网络结构,可以用于图像分类和目标检测等计算机视觉任务。ShuffleNetV2是对ShuffleNetV1的改进版本,利用轻量级的网络结构和特殊设计的ShuffleUnit进一步提高了网络性能和效率。ShuffleUnit是ShuffleNetV2中的基本模块,包括Channel Split和Channel Shuffle两个操作,在每个单元开始处,通过Channel Split将特征通道的输入切分为两个支路,经过分路的卷积计算后通过Concat方式连接,然后通过Channel Shuffle进入下一单元。ShuffleNetV2网络结构如图7所示。

在人体运动跟踪过程中,仅用检测算法对视频进行人体目标检测时容易出现因遮挡、多目标间的互相影响、光照变化等问题而丢失目标的情况,本文通过对改进的DeepSORT算法进行预训练的方式尽可能减少上述情况的发生。

改进后的DeepSORT采用Market-1501数据集进行人体的ReID训练,其中751名人体用于训练,752名人体用于测试,提取相应的边界框特征用于训练。经过40次训练迭代后,验证集和训练集的分类损失都降到了最小值,三元组损失也降到了最小值。DeepSORT中原始特征提取网络生成的权重文件大小为46 382 KB,使用ShuffleNetV2网络加以改进后生成的权重文件大小为2 595 KB,实现了模型轻量化,增加了此方法在不同设备上的普适性。通过训练生成的权重文件用于人体ReID和目标跟踪,可以保证在跟踪过程中尽量少地发生目标丢失和人员ID频繁切换的问题。类别损失和三元组损失如图8所示。

3.4 结合OpenPose的人体姿态跟踪实现

目前,对视频进行人体姿态跟踪的技术存在的主要问

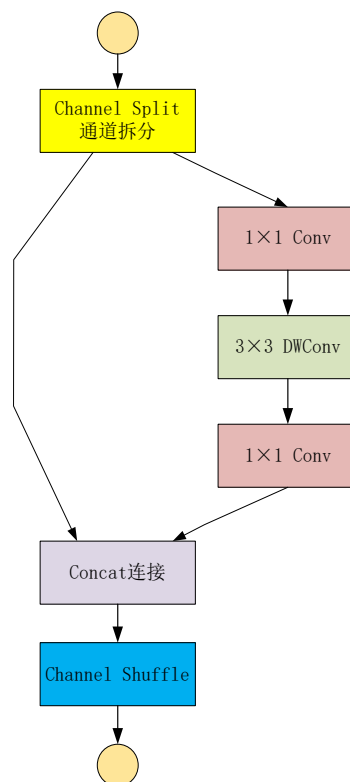


Fig. 7 Model structure of ShuffleNetV2

图7 ShuffleNetV2网络结构

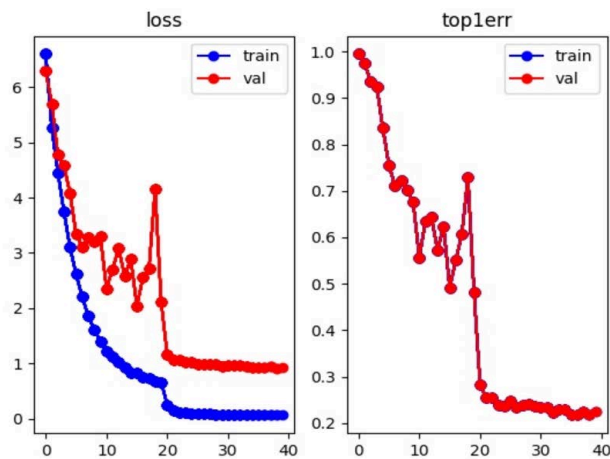


Fig. 8 Class loss and triplet loss

图8 类别损失和三元组损失

题是某一帧出现人体目标被遮挡、目标重叠或不清晰,人体目标难以被检测出来,导致目标跟踪丢失和目标ID频繁变换。

为解决此问题,本文结合OpenPose人体姿态估计模型,加入基于改进的YOLOv5算法和改进的DeepSORT算法实现人体姿态跟踪。DeepSORT算法为每个检测出的人体目标都赋予一个ID值,并持续追踪该目标的运动轨迹,当人体目标出现遮挡、重叠或不清晰的状态,依然能根据上下文信息进行推断,实时性也更强。

向基于DeepSORT和OpenPose的人体姿态跟踪模型

输入一段视频后,按帧进行检测跟踪与姿态估计。基本人体姿态跟踪过程如下:①通过YOLOv5检测模型得到第一帧图像的人体目标边界框,DeepSORT算法根据第一帧的检测结果创建对应轨迹,将卡尔曼滤波的运动变量初始化,通过卡尔曼滤波预测它们的下一个位置,得到预测框,然后使用匈牙利算法将得到的检测框与预测框数据关联,进行级联匹配或IoU匹配,计算其代价矩阵,得到某一目标的跟踪结果,实现对每个视频帧中所有被检测到的人的匹配和跟踪,并为每个人分配一个唯一的ID;②将DeepSORT追踪结果提供给OpenPose模型,对每个视频帧中被DeepSORT确认并追踪过的人进行单独OpenPose姿态估计,获取每个人体关键点坐标,包括头部、肩膀、手肘、手腕、膝盖和脚踝等;③将特征表示和每个人的ID从DeepSORT模型传递给OpenPose模型,在每个人的目标中进行关键点跟踪和预测;④在每个视频帧中,将OpenPose生成的人体关键点坐标与DeepSORT生成的跟踪结果相结合,再次匹配和跟踪所有人,并根据其在当前视频帧中的位置更新其ID;⑤反复重复以上步骤直到视频结束;⑥生成准确的多人姿态跟踪结果并输出。

具体实现流程如图9所示。

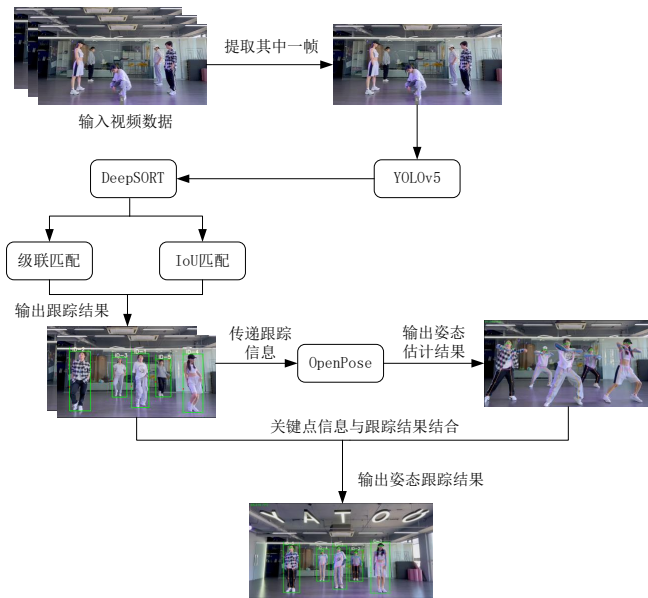


Fig. 9 Flow of human pose tracking

图9 人体姿态跟踪流程

4 实验与结果分析

4.1 实验环境

实验中的计算机配置如表1所示。

4.2 数据集选择

COCO数据集是Google公司开源的大型数据集,是一个可用于图像检测、语义分割和图像标题生成的大规模数据集,其中有250 000个带关键点标注的人体。在关键点检测方面,可以在任意姿态下对人物的关键点进行定位,

Table 1 Experimental configuration

表1 实验配置

参数	型号版本
系统	Windows 11
CPU	Intel Core i7-11800H
GPU	RTX 3060
CUDA 版本	CUDA11.6
框架	Pytorch2.0.0
Python 版本	3.9

该任务包含检测人体及定位到人体的关键点。COCO数据集目前有3个版本,即2014、2015和2017,其中2015版只有测试集,其他两个有训练集、验证集和测试集,本文选用COCO 2017数据集。

MOT(Multiple Object Tracking)数据集是用于多目标跟踪任务的一类广泛使用的数据集。MOT数据集通常提供了丰富的标注信息,以支持算法的评估和性能比较。这些标注信息包括每个目标的边界框位置、目标ID、运动轨迹等。MOTChallenge数据集是较为常用的一个多目标跟踪数据集系列,包括MOT15、MOT16、MOT17和MOT20等版本,这些数据集提供了大量的视频序列和相关的标注信息,可以用于评估和比较不同算法的性能。本文选用MOT16数据集,该数据集共有14个视频序列,其中7个为带有标注信息的训练集,另外7个为测试集。

PoseTrack是一个大规模的视频人体姿态估计和跟踪基准数据集,它包含了一系列具有挑战性的高清晰度人群在密集人群中表演各种各样的活动。PoseTrack2017包含514个视频,其中250个用于训练,50个用于验证,214个用于测试。这些数据集用15个身体关节进行注释,每个关节都被定义为一个点,并与一个唯一的id相关联。

4.3 评价指标

人体检测结果是选择图像中人体目标的边界框。为衡量人体检测情况,引入定位评价公式IoU(Intersection-over-Union)作为中间变量,计算精度评价指标。

$$IoU = \frac{area(box_{pred} \cap box_{truth})}{area(box_{pred} \cup box_{truth})} \quad (4)$$

其中,area表示边界框的区域面积, box_{pred} 是预测的目标框, box_{truth} 是标注的目标框。

IoU量化了两个矩形框的重叠,其表示为两个边界框的交集与它们并集的比率。在人体检测中,两个边界框分别是标注框和前一帧的预测框。

人体检测场景丰富,实际情况复杂,如目标人体运动姿态变化引起的变形、多个人体的重叠问题以及人体在镜头中不完全呈现的情况。为了应对这些情况,需要引入评价参数以改进相关算法。因此,选择精度Precision、全类平均精度mAP和浮点运算次数FLOPs作为评估指标。

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$AP = \int_0^1 p(r) dr \quad (6)$$

其中,TP是指由检测模型正确识别的人体数量,FP是指被错误识别为人体的非人体目标数量。

全类平均精度(mean Average Precision, mAP)是将所有类别检测的 AP 进行综合加权平均而得到,代表模型对所有类别的检测能力,是衡量目标检测算法性能的重要指标。

FLOPs是每秒所执行的浮点运算次数。假设输入特征图大小为 $W_{in} * H_{in} * M$,输出特征图大小为 $W_{out} * H_{out} * N$,卷积核大小为 $D_k * D_k$,则标准卷积 FLOPs 值为 $D_k * D_k * M * W_{out} * H_{out} * N$ 。

由于模型速度与计算量 FLOPs 成正比,模型检测速度不好衡量,可以用 FLOPs 代替,作为模型检测速度衡量标准。本文 FLOPs 值代表模型检测一张图像所产生的计算量。

为衡量人体目标跟踪情况,引入 MOTA(MOT Accuracy)作为评价指标。

$$MOTA = 1 - \frac{\sum_t FN_t + FP_t + IDSW_t}{\sum_t GT_t} \quad (7)$$

其中, t 是时间帧的索引,GT 表示真实框的人体数量,FP(False Positive)表示错误跟踪,即真实情况中没有但跟踪算法误检出有目标存在,FN(False Negative)表示漏检,即真实情况中有但跟踪算法未检测到,IDSW 指在目标跟踪中一个人体目标发生遮挡再重新出现在视野中,若重新赋予了一个序号,则定义为发生一次 IDSW。

MOTA 能非常直观地衡量跟踪器在检测物体和保持轨迹时的性能,与目标检测精度无关。其取值一般小于 100,当跟踪器产生的错误超过了场景中的物体时,MOTA 会变为负数。

4.4 实验结果

在人体姿态跟踪过程中,算法精度和性能主要受制于目标跟踪和姿态估计的算法优化。本文主要针对目标跟踪阶段的目标检测 and 数据处理进行改进,因而在实验阶段先验证改进的有效性,再对整个算法开展对比实验。

4.4.1 改进的 YOLOv5 结果

本文将 YOLOv5s 主干网络中最后一个 C3 模块替换成 C3TR 后,在 COCO 数据集上的检测结果如表 2 所示, mAP@.5 有所提高,说明准确率得到提升。

Table 2 Experimental result of improved YOLOv5

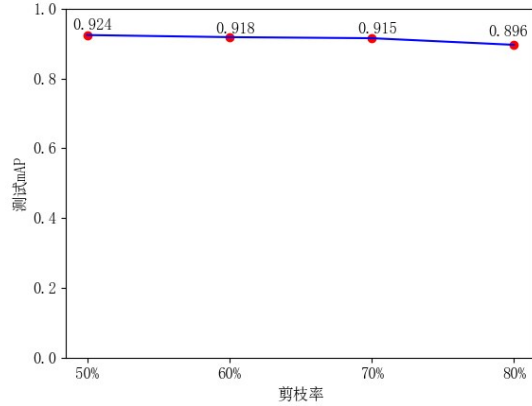
表 2 改进 YOLOv5 实验结果

模型	数据集	mAP@.5
YOLOv5s	COCO2017(val)	0.558
YOLOv5s+TR	COCO2017(val)	0.608

4.4.2 模型剪枝结果

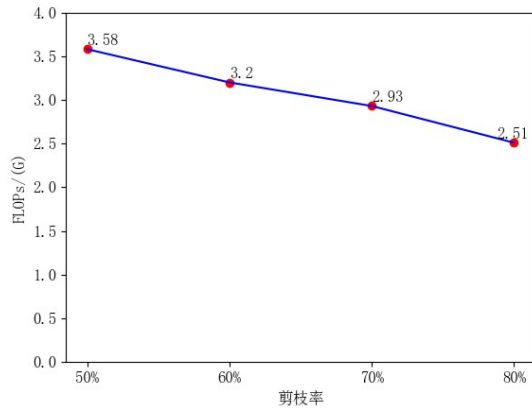
本文在模型剪枝时测试了 4 组剪枝率,分别为 50%、60%、70%、80%,输入分辨率为 480 像素, batch_size 为 16, mAP 值、FLOPs、模型大小变化如图 10 所示。

可以得知,随着剪枝率的提升,模型 mAP 值不断下降,



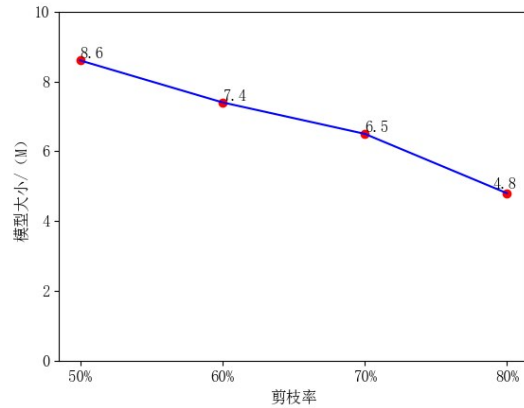
(a) The changing trend of mAP value with pruning rate

(a) mAP 随剪枝率变化趋势



(b) The changing trend of FLOPs with pruning rate

(b) FLOPs 随剪枝率变化趋势



(c) The changing trend of model size with pruning rate

(c) 模型大小随剪枝率变化趋势

Fig. 10 The changing trend of mAP value, FLOPs, and model size with pruning rate

图 10 mAP 值、FLOPs、模型大小随剪枝率变化趋势

但下降速度较缓,FLOPs 和模型大小得到了大幅降低。由于剪枝率为 70% 时,mAP 降低幅度不明显,而剪枝率为

80% 比剪枝率为 70% 时, mAP 降低了 0.019, 相比之前幅度变化明显。为不过多降低检测精度, 本文设置剪枝率为 70%。

经稀疏训练、模型剪枝、模型微调后的模型检测结果如表 3 所示, 并与未剪枝前的模型作了对比。

Table 3 Model pruning results based on YOLOv5s

表 3 基于 YOLOv5s 的模型剪枝结果

步骤	图像尺寸(/ 像素)	Epoch	Batch_ size	最大 mAP	测试 mAP	FLOPs/G	模型 大小/M
未剪枝前	480	200	16	0.951	0.948	4.84	14.2
稀疏训练	480	200	16	0.94	0.933	4.84	14.2
模型剪枝	480	200	16	0.918	0.916	2.9	6.1
模型微调	480	200	16	0.931	0.946	2.9	3.2

可以看出, 模型微调后测试 mAP 值比剪枝后增加了 0.03, 是一种有效的精度补偿方法。将微调后的模型与未剪枝前的模型进行对比, 模型测试 mAP 值只下降了 0.02, 但 FLOPs 值降低了约 2 G, 模型大小降低了 11 M, 在不影响检测精度的前提下极大减少了模型计算量和模型大小, 提升了检测速度, 减少了模型存储空间。由此得出, 在基于 YOLOv5 检测模型的基础上, 采用基于 BN 层的模型剪枝方法, 能够进一步提升模型检测速度, 同时对模型检测精度影响较小。

4.4.3 人体目标检测结果

Faster-RCNN 是一种经典的两阶段目标检测算法, Gold-YOLO 是一种全新的实时目标检器。本文改进的 YOLOv5 算法在人体目标检测上的实验结果如表 4 所示。

Table 4 Experimental results of improved YOLOv5 algorithm in human object detection

表 4 改进的 YOLOv5 算法在人体目标检测上的实验结果

算法	epoch	batch_size	数据集	mAP	FLOPs/G
本文方法	200	16	COCO2017	0.72	4.46
Faster-RCNN	200	16	COCO2017	0.76	21.5
Gold-YOLO	200	16	COCO2017	0.80	12.66

可以得知, 本文检测模型精度虽稍弱于 Faster R-CNN 网络和 Gold-YOLO, 但检测速度有较大优势。YOLOv5 算法相比其它网络, 在检测速度上本就优势明显, 经过改进后, 使得在 YOLOv5 算法的基础上进一步降低了检测网络的计算量, FLOPs 值达到了 4.46 G, 速度得到大幅提升。

4.4.4 人体目标跟踪结果

在人体目标跟踪方面, 本文对比了经典的 SORT 多目标跟踪算法和较新的 TR-MOT 多目标跟踪算法, 实验结果如表 5 所示。运行时间记录的是一段 60 s 的视频在不同目标跟踪模型下的运行时间。

由实验结果可知, 本文方法通过结合改进的 YOLOv5 检测模型和改进的 DeepSORT 算法, 有效降低了跟踪过程中的 ID 切换次数, 通过 MOTA 指标可以得出在跟踪精度上相差不大, 但在运行时间上有较大优势。

4.4.5 人体姿态跟踪结果

在人体姿态跟踪方面, 本文对比了 LightTrack 算法和

Table 5 Experimental result of object tracking

表 5 目标跟踪实验结果

算法	数据集	IDSW	MOTA	运行时间/s
本文方法	MOT16	83	52.4	85.55
SORT	MOT16	155	50.6	101.24
TR-MOT	MOT16	97	52.1	103.20

BlazePose 算法, 实验结果如表 6 所示。可以得知, 本文方法在人体跟踪阶段的改进明显提高了算法整体运行速度, 在姿态跟踪精度方面, 针对 DeepSORT 预训练生成的权重文件, 跟踪效果也有所提升。

Table 6 Experimental result of human pose tracking

表 6 人体姿态跟踪实验结果

算法	数据集	mAP	MOTA	FPS	运行时间(s)
本文方法	PoseTrack2017	0.74	56.1	55.6	81.79
LightTrack	PoseTrack2017	0.71	53.6	50.1	103.20
BlazePose	PoseTrack2017	0.72	53.9	52.2	99.14

4.5 实验效果

4.5.1 人体目标跟踪效果

人体目标跟踪效果如图 11 所示。可以得知, 使用改进的 YOLOv5 和 DeepSORT 算法跟踪人体目标, 根据外观特征为每个边界框分配 ID 编号, 并且在不同目标遮挡、目标外观相似的情况下, 目标 ID 可以实现较少切换, 从而实现准确跟踪。

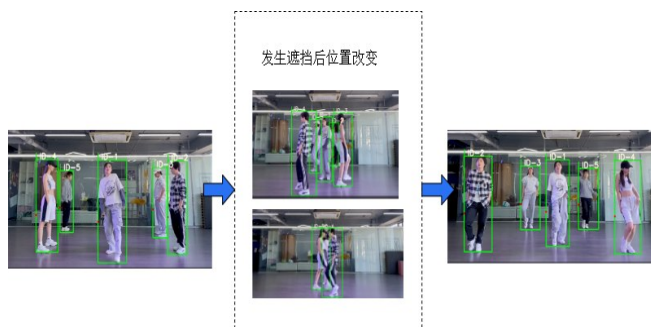


Fig. 11 Tracking effect of human object

图 11 人体目标跟踪效果

4.5.2 人体姿态估计效果

通过 OpenPose 算法可以输出一个人的骨架, 即某个人体目标在图像或者视频中的关节点和它们之间的连接关系, 如图 12 所示。本文对单个人体目标输出 18 个关节点, 输出的关节点可以被分成 4 类: 身体、左肩、右肩和脸。由实验效果可见, 当发生自身遮挡、多个人体目标发生遮挡以及处于特殊姿势时, 依旧可以较好地对人体姿态估计。

4.5.3 人体姿态跟踪效果

图 13 给出了本文算法与 LightTrack 方法的定性比较效果。选取一段视频中的 3 帧进行对比展示, 第一帧为两个方法均检测到多个人体目标的同一时刻, 第二帧为多个人体目标互相遮挡时刻, 第三帧为多个人体目标交换位置后的同一时刻。由实验效果可见, 本文算法在第三帧没有发生人体目标 ID 错误的情况, 而 LightTrack 方法出现 2 个

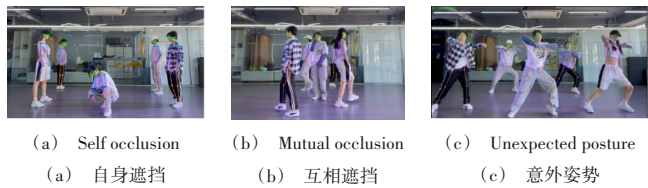


Fig. 12 Experimental effect of human pose estimation in different situations

图 12 不同情况下的人体姿态估计效果

ID 错误的情况。本文方法在检测到多个人体目标后,即为

各目标附一个 ID 值,并不断追踪,同时实现对各人体目标的姿态估计,减少了目标丢失和切换 ID 的情况。

5 结语

本文为解决多人姿态跟踪过程中因多人遮挡、相似、复杂环境等原因导致的目标易丢失、姿态估计不准确等问题,提出了一种结合 DeepSORT 与 OpenPose 的人体姿态跟踪方法。基于改进的 YOLOv5 与改进的 DeepSORT 算法,减小模型内存访问成本,实现模型轻量化,进行准确识别

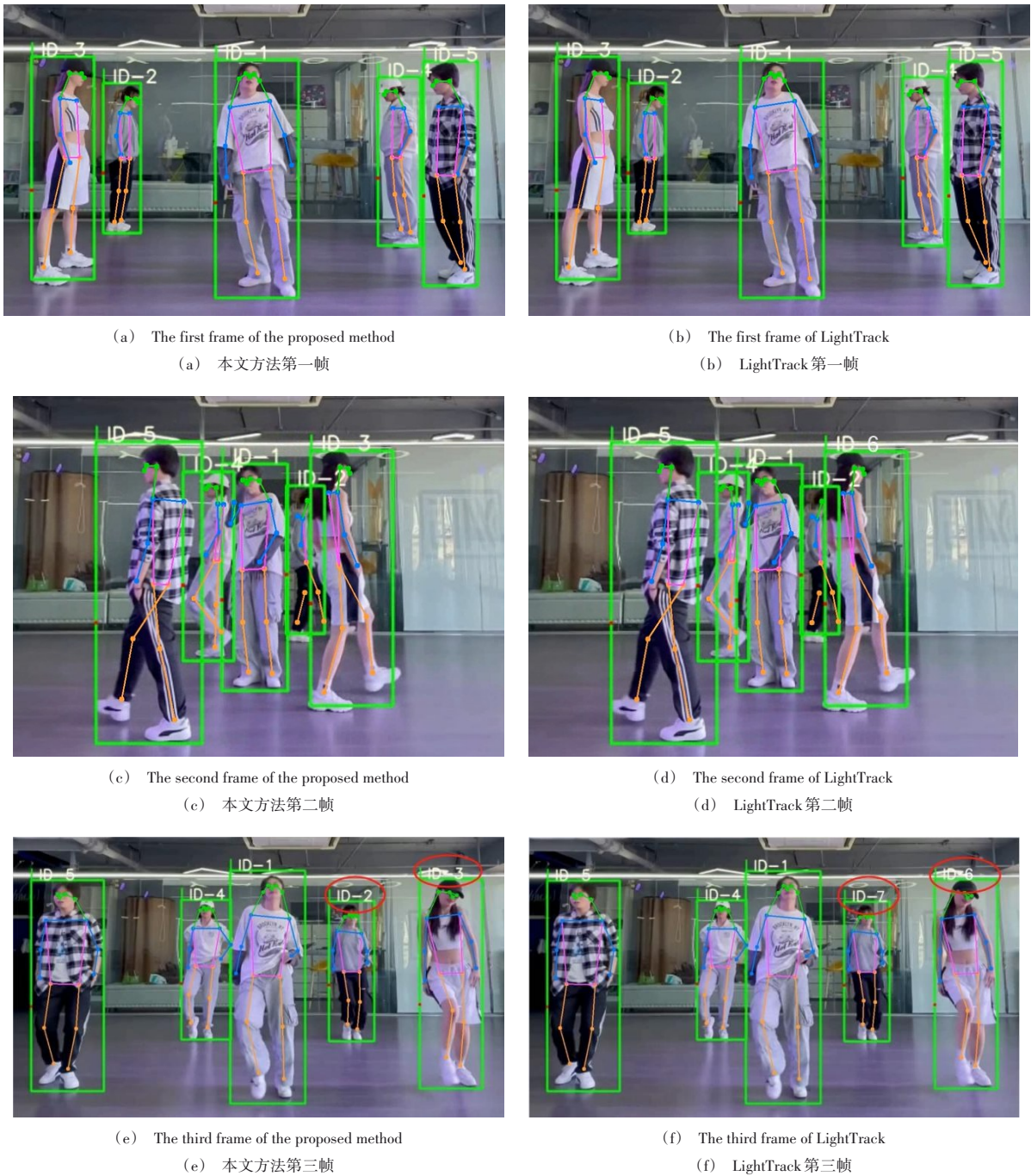


Fig. 13 Experimental effect of human pose tracking

图 13 人体姿态跟踪实验效果

与跟踪,并引入基于BN层的模型剪枝方法提升速度,再利用OpenPose算法实现对目标人体的姿态估计。在COCO 2017、MOT16和PoseTrack2017数据集上的对比实验结果表明,本文算法在速度上有较大提升,精度上也有一定提升,提高了对遮挡、复杂环境、多人场景的适应能力。在后续工作中,可进一步优化姿态估计阶段算法,解决因光照变化等因素对人体目标跟踪和姿态估计准确性的影响,提升人体姿态跟踪整体性能。

参考文献:

- [1] XU G Y, TAO L M, SHI Y C, et al. Human computer interaction for ubiquitous/pervasive computing mode[J]. Chinese Journal of Computers, 2007 (7): 1041-1053.
徐光祐,陶霖密,史元春,等. 普适计算模式下的人机交互[J]. 计算机学报, 2007(7): 1041-1053.
- [2] SHI H L, ZHAO G Y. Human centered action analysis in artificial intelligence[J]. Journal of Northwest University (Natural Science Edition), 2020, 50(3): 328-337.
石恒麟,赵国英. 人工智能中面向人类的行为分析[J]. 西北大学学报(自然科学版), 2020, 50(3): 328-337.
- [3] LIU M. The application framework of virtual reality technology based on computer vision in the field of sports action correction[J]. Adhesion, 2020, 41(5): 185-188, 192.
刘明. 基于计算机视觉的虚拟现实技术在体育动作纠正领域的应用框架[J]. 粘接, 2020, 41(5): 185-188, 192.
- [4] HU Y, CAO Z. The application of deep learning technology in unmanned driving in smart cities[J]. Environmental Technology, 2021, 39(3): 100-105, 110.
胡艳,曹钻. 基于机器视觉和深度学习的自动驾驶方法研究[J]. 环境技术, 2021, 39(3): 100-105, 110.
- [5] MICHAEL J. The SCAPE shape descriptor for articulated 3D shape representation and matching[C]// Proceedings of the 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2000: 716-721.
- [6] YI L, MAIRE M, SERGE J, et al. Microsoft COCO: common objects in context [C]// Proceedings of the 13th European Conference on Computer Vision, 2014: 740-755.
- [7] MILAN A, LEAL L, REID I, et al. MOT16: a benchmark for multi-object tracking[DB/OL]. <https://arxiv.org/abs/1603.00831v2>, 2016.
- [8] ANDRILUKA M, IQBAL U, INSAFUTDINOV E, et al. Posetrack: a benchmark for human pose estimation and tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5167-5176.
- [9] GIRSHICK R. Fast R-CNN [C]// Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.
- [10] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// Proceedings of the International Conference on Neural Information Processing Systems, 2015: 91-99.
- [11] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.
- [12] DAI J, LI Y, HE K, et al. R-FCN: object detection via region-based fully convolutional networks [C]// Proceedings of the Advances in Neural Information Processing Systems, 2016: 379-387.
- [13] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]// Proceedings of the 14th European Conference on Computer Vision, 2016: 21-37.
- [14] HENRIQUES J F, CASEIRO R, MARTINS P, et al. Exploiting the circulant structure of tracking-by-detection with kernels [C]// Proceedings of 12th European Conference on Computer Vision, 2012: 702-715.
- [15] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(3): 583-596.
- [16] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking [C]// Proceedings of the European Conference on Computer Vision, 2016: 850-865.
- [17] LI B, YAN J, WU W, et al. High performance visual tracking with siamese region proposal network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8971-8980.
- [18] BEWLEY A, GE Z, OTT L, et al. Simple online and realtime tracking [C]// Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), 2016: 3464-3468.
- [19] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric [C]// Beijing: 2017 IEEE International Conference on Image Processing (ICIP), 2017.
- [20] CHEN M, LIAO Y, LIU S, et al. TR-MOT: multi-object tracking by reference[DB/OL]. <https://arxiv.org/abs/2203.16621v1>, 2022.
- [21] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2d pose estimation using part affinity fields [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7291-7299.
- [22] MATHIS A, MAMIDANNA P, CURY K M, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning [J]. Nature Neuroscience, 2018, 21(9): 1281-1289.
- [23] YAN B, PENG H, WU K, et al. Lighttrack: finding lightweight neural networks for object tracking via one-shot architecture search [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 15180-15189.
- [24] XIU Y, LI J, WANG H, et al. Pose flow: efficient online pose tracking [DB/OL]. <https://arxiv.org/abs/1802.00977>, 2018.
- [25] BAZAREVSKY V, GRISHCHENKO I, RAVEENDRAN K, et al. BlazePose: on-device real-time body pose tracking [DB/OL]. <https://arxiv.org/abs/2006.10204>, 2020.

(责任编辑:孙娟)