基于实时视频流的 3D 人体姿势和形状估计

朱　越[1],黄海于[1],罗学义[2]
(1. 西南交通大学 计算机与人工智能学院,四川 成都 611756;
2. 消防救援局昆明训练总队,云南 昆明 650217)

摘　要:为满足元宇宙、游戏及虚拟现实等应用场景中对实时视频流 3D 人体姿势和形状估计准确性和真实性的要求,提出了一种基于时间注意力机制的 3D 人体姿势和形状估计方法。首先,提取图像特征,并将其输入运动连续注意力模块以更好地校准需要注意的时间序列范围;随后,使用实时特征注意力集成模块以有效地组合当前帧与过去帧的特征表示;最后,通过人体参数回归网络得到最终结果,并使用基于图卷积的生成对抗网络判断模型是否来自真实的人体运动数据。相较于之前基于实时视频流的方法,在主流数据集上加速度误差平均减少了 30% 的同时,网络参数与计算量减少了 65%,在实际测试中实现了每秒 55 ~ 60 帧的 3D 人体姿态和形状估计速度,为元宇宙、游戏及虚拟现实等应用场景提供更好的用户体验和更高的应用价值。

关键词:三维人体重建;SMPL 模型;实时特征注意力集成;图卷积神经网络;机器学习
中图分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1673-629X(2024)04-0042-06
doi:10.20165/j.cnki.ISSN1673-629X.2024.0007

3D Human Pose and Shape Estimation Based on Live Video Streams

ZHU Yue[1],HUANG Hai-yu[1],LUO Xue-yi[2]
(1. School of Computer and Artificial Intelligence,Southwest Jiaotong University,Chengdu 611756,China;
2. Kunming Training Corps of Fire Rescue Bureau,Kunming 650217,China)

Abstract:A 3D human body pose and shape estimation method based on a temporal attention mechanism is proposed to meet the requirements of real-time accuracy and realism in 3D human body pose and shape estimation for applications such as the metaverse, gaming,and virtual reality. First,image features are extracted and input into a motion continuity attention module to better calibrate the time sequence range that requires attention. Then,a real-time feature attention integration module is used to effectively combine the feature representations of the current frame and past frames. Finally,the human parameter regression network is used to obtain the final results,and a graph convolutional generative adversarial network is used to determine whether the model comes from real human motion data. Compared with previous methods based on real-time video streams,the proposed method reduces the acceleration error by an average of 30% on mainstream datasets,while reducing the network parameters and computational complexity by 65%. The proposed method achieves a 3D human body pose and shape estimation speed of 55 ~ 60 frames per second in practical tests,providing better user experience and higher application value for applications such as the metaverse,gaming,and virtual reality.

Key words:3D human reconstruction;skinned multi-person linear model;real-time feature attention integration;graph convolutional neural network;machine learning

0　引　言

对于许多以人为中心的 3D 计算机视觉应用等来说,估计人体的姿态和形状是一件至关重要的任务[1]。随着该领域的发展,基于模型的方法被设计用于估计具有几个参数的三维人体姿态和形状。尽管单个图像和视频片段的准确性和鲁棒性有所提高,但是随着 Virtual Reality(VR)游戏的发展以及元宇宙的兴起,对于人体姿态的估计的实时性与真实性需求越发强烈,在实时视频流方面提高人体姿态和形状估计的准确性和时间一致性方面仍然缺乏探索。

基于单个图像的方法[2-4],在空间误差以及关节点误差方面取得了令人印象深刻的成绩。然而,这些

收稿日期:2023-06-26　　**修回日期**:2023-10-26
基金项目:应急管理部消防救援局科技创新项目(2020XFCX29)
作者简介:朱　越(1999-),男,硕士,CCF 会员(P2474G),研究方向为 3D 人体重建、计算机视觉;黄海于(1970-),男,通信作者,副教授,硕导,研究方向为人工智能、计算机图形学。

方法中每个图像的估计误差是高度独立的,当它们应用于实时视频流时,即使是微小的空间波动也会对最终的估计结果产生巨大的影响,导致动作出现剧烈的抖动。时间序列信息的缺乏使得此类方法很难获得时间上稳定的姿态和形状估计。

基于视频的方法[5-6],将帧序列输入到预先训练的基于单帧图片训练的特征提取网络中以获得每张图片的静态特征,随后利用时间编码器提取时间特征,最后网络参数回归器通过输入时间特征为每一帧输出名为 Skinned Multi-Person Linear Model (SMPL)[7]的基于模板的参数。然而,这类方法通常依赖来自未来帧的信息,这使得它们无法在实时视频流上应用。为了获取更丰富的时间序列特征,这些方法通常要求输入长时间的视频,同时时间编码器的推理时间也相对较长,这限制了它们在实时应用中的可行性。

为了能够实现实时视频流上的人体姿态和形状估计,Wang Z 提出了 TePose[8]方法,将先前的姿态估计作为输入结合静态特征进行时间特征的提取。然而,由于在视频开始的一段时间内无法获得相应的人体估计参数,该方法无法准确估计人体的姿态和形状信息。此外,推理速度较慢,无法适用于实时视频流。同时,该方法中的动作鉴别器部分仅对当前姿势进行动作判别,尽管输入序列包含之前的姿势估计结果,但之前的估计结果已不在计算图内,因此运动鉴别器对生成真实动作的影响微乎其微。

针对上述问题,该文提出了一种基于实时视频流的时间注意 3D 人体姿势和形状估计的方法,利用过去帧与当前帧的特征信息,通过注意力机制来有效地捕获运动中的人体姿势,以从实时视频流中估计准确且真实的 3D 人体姿态和形状。主要贡献描述如下:

(1)开发了一个实时时间特征注意力集成模块(RFA),该模块以分层注意力集成的方式有效地组合了过去帧和当前帧的特征表示,以加强时间相关性并细化当前帧的特征表示,同时推理速度具有很高的实时性。

(2)引入基于图卷积网络(GCN)的运动鉴别器,在没有 3D 标注的情况下提供 3D 估计评估。

(3)实现了基于实时视频流的 3D 人体姿势和形状估计,并在广泛使用的实时人体姿势和形状估计上实现最先进的性能。

1　相关工作

1.1　基于单个图像的 3D 人体姿势和形状估计

现有的对于单图像的 3D 人体姿势和形状估计大都采用参数 3D 人体模型,如 SMPL,即通过深度网络模型估计姿态、形状和相机参数,然后通过 SMPL 模型将其解码为人体的 3D 网络。Kanazawa 等人[9]提出一种端到端的人体网格恢复(Human Mesh Recovery, HMR)模型,通过最小化关节的重投影损失来直接从图像像素中估计姿态和形状的参数。Kolotouros 等人[10]提出了一种自改进框架(SPIN),该框架集成了 SMPL 参数回归器和迭代拟合的方案,以更好地估计 3D 人体姿势和形状。在此基础上华为诺亚方舟实验室[2]将估计的包围盒坐标信息、图片像素信息以及相机焦距信息加入到回归框架中,获得了更加准确的结果。尽管上述方法对静态图像有效,但是每张图片估计都是相互独立的,很难在实时视频序列上生成时间上连续且平滑的 3D 人体姿势和形状,即,可能会出现抖动、不稳定的 3D 人体运动[11]。

1.2　基于视频的 3D 人体姿势和形状估计

与基于单个图像的方法类似,现有的基于视频的 3D 人体姿态和形状的估计方法主要基于 SMPL 模型。比较有代表性的工作是 Kocabas 等人[5]提出的基于视频流的循环网络方法(VIBE),在 SPIN 的基础上,引入了具有自我注意力机制的对抗性学习的方法,该框架利用 AMASS 数据集[12]来监督人体姿势和形状回归网络。Choi 等人[6]提出了一种时间一致性网络恢复(TCMR),通过两个单向门控循环单元(GRU)和一个双向 GRU,整合过去帧、当前帧与未来帧的特征信息进行参数估计。Wei 等人[13]提出了一种运动姿态和形状网络(MPS-Net),借助图注意力机制捕获连续的人体运动从而生成更加平滑的姿态。尽管上述方法实现了更为平滑的 3D 人体姿态和形状的估计,但是像是类似 TCMR 的方法,需要使用到未来帧的信息,这在实时视频流中是无法取得的,其余的基于视频的方法往往需要较长的视频序列信息以获得足够的时间信息,因此推理速度十分缓慢。

1.3　基于人体骨架的图神经网络

简单的将骨骼数据表示为由循环神经网络(RNN)处理的向量序列,或由图卷积网络(CNNs)处理的 2D/3D 图,不能完全建模身体关节的复杂时空配置和相关性。这表明拓扑图可以更适合于表示骨架数据。Yan 等人[14]通过引入时空图卷积网络(ST-GCN)将时空图卷积网络(GCN)用于基于骨架的人体动作识别(HAR),该 GCN 可以从骨架数据中自动学习空间和时间信息。Liu 等人[15]对其进行改进,将多尺度时空图卷积网络[16](MS-GCN)扩展到其解耦版本称之为 MS-G3D。

2　基于注意力机制与 GCN 的模型设计

基于实时视频流的时间注意 3D 人体姿势和形状估计(TaPose)的总体框架如图 1 所示。它主要包含三

个部分:第一个部分是运动注意力机制模块;第二个部分是实时特征注意力集成(RFA)模块;最后一个部分是基于 GCN 的运动鉴别器。

2.1 运动注意力模块

该文将从实时视频流中接收到的实时帧定义为当前帧 I_t,当前帧之前的 T 帧 $\{I_{t-T},\cdots,I_{t-1}\}$。首先使用 Kolotouros 等人[10]预先训练的 ResNet-50[17]网络来提取每个帧的静态特征信息形成特征序列 $P=\{p_n\}_{n=t-T}^{t}$。其中 $p_n \in R^{2\,048}$。随后将提取到的 P 发送到运动注意力机制模块以计算时间特征表示序列 $Q=\{q_n\}_{n=t-T}^{t}$,其中 $q_n \in R^{2\,048}$。运动注意力模块由 MPS-Net[13]提取出来,如图 2 所示,其中 $[\rho,\gamma,\varphi,\psi]$ 表示卷积运算,$\otimes$ 表示矩阵乘法,$\oplus$ 表示元素和。

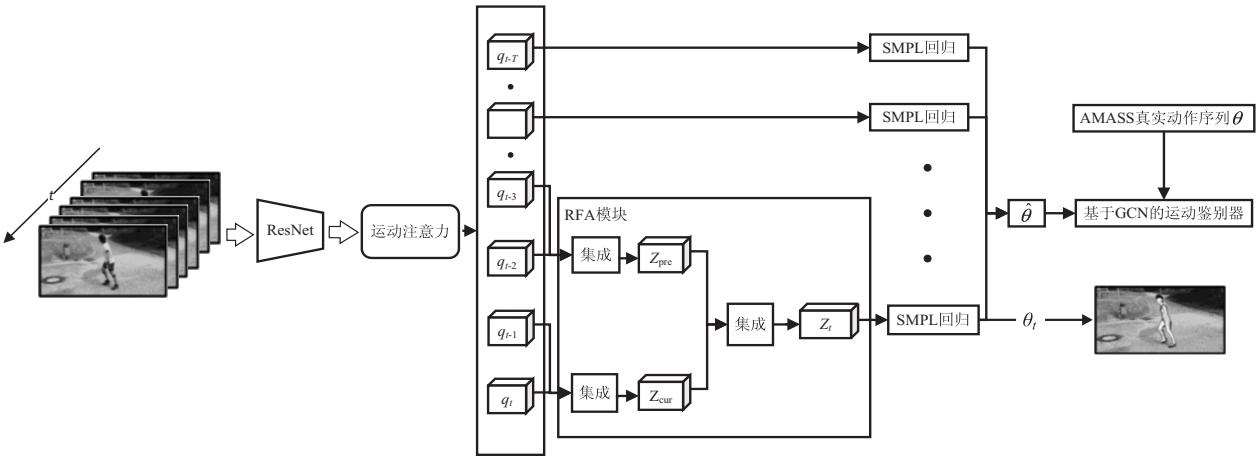

图 1 TaPose 总体框架

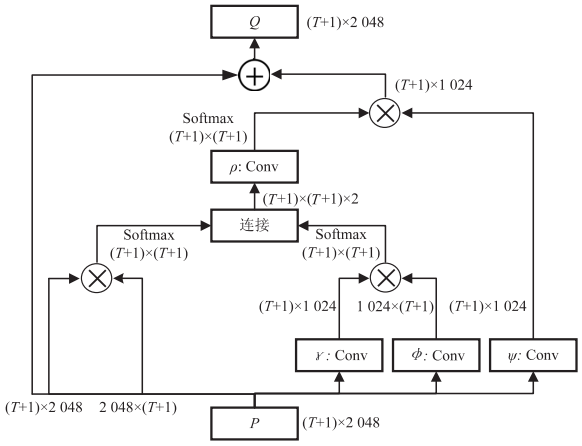

图 2 运动注意力模块结构

2.2 实时时间特征注意力集成模块

给定时间特征表示序列,RFA 模块的目标是通过整合从过去帧中观察到的相邻时间特征来细化当前帧 q_t 的时间特征,以增强它们的相关性并获得更好的姿态和形状估计。具体而言,在 RFA 模块中每个集成分支如图 3 所示。该文将两个相邻帧视为一个特征组,组之间的相邻帧不重叠,并通过共享的全连接(FC)层将它们分别从 2 048 维调整为 256 维,以降低计算复杂性。随后将其级联并重新调整时间特征大小为 $Z^c \in R^{512}$,并传递到三个 FC 层和一个 Softmax 激活,以探索它们之间的依赖关系,从而计算注意力值 $A=\{a_k\}_{k=1}^{2}$。然后,将注意力值加权回每个对应的帧,以放大重要帧在时间特征集成中的贡献,从而获得聚合的时间特征。由底部分支产生的聚合时间特征将被传递到上层,并以相同的方式进行集成,以产生最终的细化 Z_t。它将为 SMPL 参数回归器提供机会使其能够学习估计准确且时间一致的 3D 人体姿态和形状。

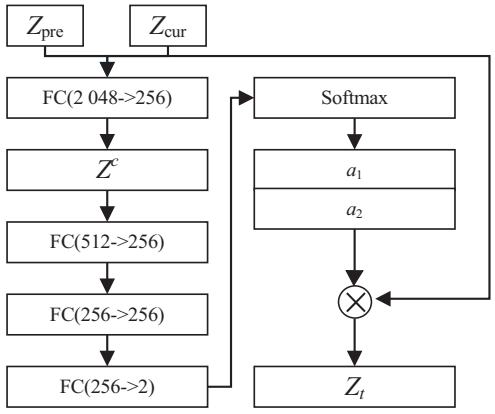

图 3 RFA 模块中每个集成分支的网络结构

2.3 基于 GCN 的运动鉴别器

相较于之前 VIBE 与 TCMR 所采用的基于 GRU 的运动鉴别器,基于 GCN 的运动鉴别器可以更适合于表示骨架数据。同时能够实现更快的推理速度和准确度。该文的 TaPose 经过训练可以生成尽可能真实的 3D 骨架序列,以对抗的方式欺骗运动鉴别器。

基于 GCN 的运动鉴别器网络如图 4 所示,包含三个 GCN 块和全局池化与全连接层,每个 GCN 块将具有剩余连接的 MS-G3D 和 MS-GCN 单元的输出相加。输出通道的数量对于三个 GCN 块分别为 64,128 和 256。在 2.1 节中的时间编码器的输出为时间特征表示序列 $Q=\{q_n\}_{n=t-T}^{t}$。该文将时间特征表示序列 Q

$=\{q_n\}_{n=t-T-1}^{t}$ 部分与 Z_t 合并,随后输入到 SMPL 回归器中得到 SMPL 估计姿势参数序列 $\hat{\theta}=\{\hat{\theta}_n\}_{n=t-T}^{t}$,其中 $\hat{\theta}_n$ 表示 SMPL 模型的姿势参数。该文将估计的姿势序列 $\hat{\theta}$ 与来自运动捕捉数据集的真实姿势序列 θ 统一为骨架序列 $G\in R^{(T+1)\times N\times C}$,其中 N 是关节数,C 是每个关节的坐标维度。骨架序列首先由三个连续的 GCN 块进行处理。然后,使用全局平均池化层和全连接层来输出输入骨架序列是真实的还是生成的最终决策。对于每个 GCN 块,通过 MS-GCN 和 MS-G3D 处理图特征。然后将它们的输出与剩余连接相加。最后,使用激活函数对图形特征进行处理,以获得鉴别器输出,用于判断当前的输入骨骼序列是否为真实的骨骼序列。

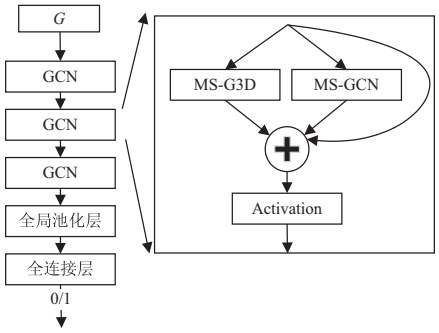

图 4　运动鉴别器架构

2.4　损失函数构造

在具有 3D 注释和没有 3D 注释的数据集上进行训练。受到先前工作的启发[6],该文的损失函数为:

$$L=L_{2D}+\exists_{3D}L_{3D}+\exists_{\Theta}L_{\Theta}+L_{adv} \tag{1}$$

其中,$\exists_{3D}$ 表示是否带有 3D 标签,$\exists_{\Theta}$ 表示是否带有 SMPL 参数标签。将 2D 损失与 3D 损失定义如下:

$$\begin{aligned}L_{3D}&=\|\hat{X}_t-X_t\|_2\\L_{2D}&=\|\hat{x}_t-x_t\|_2\\L_{\Theta}&=\|\hat{\theta}_t-\theta_t\|_2+\|\hat{\beta}_t-\beta_t\|_2\end{aligned} \tag{2}$$

其中,X_t 表示在 t 时刻 3D 关节点的真实值,$\hat{X}_t$ 表示 3D 关节点的估计值。x_t 为 2D 关节点的真实值,$\hat{x}_t=\pi_c(\hat{X}_t)$ 是使用投影透视法(π_c)将估计的 3D 关节点投影到图像平面,进而得到 2D 关节点的估计值。θ_t 表示 SMPL 姿势参数的真实值,β_t 表示 SMPL 形状参数的真实值,$\hat{\theta}_t$ 表示 SMPL 姿势参数的估计值,$\hat{\beta}_t$ 表示 SMPL 形状参数的估计值。对抗损失表示为:

$$L_{adv}=[D_M(\hat{G})-1]^2 \tag{3}$$

其中,D_M 表示基于 GCN 的运动鉴别器,$\hat{G}$ 表示骨架序列的估计值。使用交叉熵损失来训练该运动鉴别器。

$$L_{D_M}=E_G[(D_M(G)-1)^2]+E_G[D_M(\hat{G})^2] \tag{4}$$

其中,G 表示来自 AMASS 数据集的骨架序列的真实值。

3　实验结果与分析

3.1　实验配置

遵循 VIBE 及 TCMR 中的一些设置,视频序列的帧率设置为 25 ~ 30 帧。模型中的主干和回归器使用预训练的 SPIN 模型进行初始化。所有输入的人体帧都使用数据集中提供的边界框进行裁剪,并调整大小为 224×224。将遮挡增强应用于裁剪的帧,以更好地概括模型。该文选择一个小的时间长度 $T+1=6$。鉴别器中所有 MS-GCN 和 MS-G3D 单元的核尺度的数量分别设置为 13 和 6。该网络使用一个 NVIDIA RTX3080Ti GPU 进行训练。估计器和鉴别器的学习率初始化为 5—5 和 1—4,当 3D 姿态精度在每 8 个 epoch 后没有提高时,学习率降低 10 倍。Adam 优化器[18]用于更新具有 32 的小批量大小的网络权重。

3.2　实验数据集

该文同时使用 2D 数据集与 3D 数据集进行训练。对于 3D 数据集,该文采用 3DPW[19],MPI-INF-3DHP[20],Human3.6m[21]进行训练。对于 2D 数据集,该文使用 PoseTrack[22]和 InstaVariety[23]进行训练。AMASS[12]用于训练运动鉴别器。在评估方面,使用了 3DPW,MPI-INF-3HP 和 Human3.6M。其中,Human3.6M 是一个室内数据集,而 3DPW 和 MPI-INF-3HP 包含具有挑战性的室外视频。

具体而言:

3DPW 是一个具有挑战性的实景数据集,包括 60 个视频,以 30 帧每秒的速度由手机拍摄。此外,IMU 传感器被用于获取近乎真实的 SMPL 参数,即姿态和形状。其中训练集、验证集和测试集分别由 24,12 和 24 个视频组成。Human3.6M 是一个大规模的数据集,在受控室内环境下收集,包括 360 万个视频帧。该文在 5 个受试者(即 S1,S5,S6,S7 和 S8)上训练模型,在 2 个受试者(即 S9 和 S11)上进行测试。

MPI-INF-3DHP 是一个复杂的数据集,使用无标记运动捕捉系统在室内和室外场景中捕捉。通过多视图方法计算 3D 人体姿态注释。训练集和测试集分别由 8 个和 6 个受试者组成。每个受试者在室内或室外环境中拍摄了 16 个视频。总视频帧数为 130 万。该文使用官方的训练和测试数据集划分。

InstaVariety 是从 Instagram 收集的 2D 人体姿态数据集。它包括 28K 个视频,平均长度为 6 秒。

PoseTrack 也是一个用于多人姿态估计和跟踪的 2D 人体数据集,包括 1.3K 个视频。按照 VIBE 的方法,该文使用 792 个视频进行训练。

3.3 性能评估指标

对于性能评估,该文使用四种标准度量[5-6],包括平均每关节位置误差(MPJPE)、与 Procrustes 对齐的 MPJPE(PA - MPJPE)和平均每顶点位置误差(MPVPE),以及基于时间的度量,加速度误差(ACC-ERR)来计算估计误差。具体而言,MPJPE 被计算为在将骨盆关节对准真实位置之后,真实的和估计的 3D 关节位置之间的欧几里得距离的平均值。PA-MPJPE 的计算类似于 MPJPE,但在估计的姿势与真实姿势刚性对齐之后计算距离的平均值。MPVPE 为真实的和估计的 3D 网格顶点之间的欧几里得距离的平均值(由 SMPL 模型输出)。ACC-ERR 为每个关节的真实的和估计的 3D 加速度之间的平均差。这些位置误差在估计的三维坐标和真实三维坐标之间以毫米为单位测量。

3.4 实验结果分析

表 1 显示了 TaPose 与主流的基于视频的方法在 3DPW,MPI-INF-3HP 和 Human3.6M 上的性能比较。在 TCMR 之后,所有方法都在包括 3DPW 在内的训练集上进行训练,但不使用从 Mosh 获得的 Human3.6M SMPL 参数进行监督。由于法律问题,Mosh 的 SMPL 参数已从公共访问中删除[24]。

为了能够得到更加真实的人体动作估计,如表 1 所示,相较于 TePose,TaPose 在仅牺牲了 2% 单帧准确性的情况下,在加速误差方面取得了 65% 的提升,避免了模型在估计人物动作时剧烈抖动的现象。

表 1 不同方法性能对比

模型	3DPW				MPI-INF-3DHP			Human3.6M			T+1
	PA-MPJPE	MPJPE	MPVE	ACC-ERR	PA-MPJPE	MPJPE	ACC-ERR	PA-MPJPE	MPJPE	ACC-ERR	
VIBE[5]	57.6	91.9	~	25.4	68.9	103.9	27.3	53.3	78	27.3	16
MPS-NET[13]*	52.1	84.3	103.2	7.4	62.8	96.7	8.5	52	73.6	3.6	16
TCMR[6]*	52.7	86.5	102.9	7.1	63.5	97.3	8.5	52	73.6	3.9	16
TePose[8]	52.3	84.6	100.3	11.4	63.1	96.2	16.7	47.1	68.6	12.1	6
TaPose	53.5	87.2	102.5	10.1	64.9	102.2	10.8	53.8	73.2	5.8	6

注:“*”表示方法无法实时估计,“~”表示不可用结果,最后一列“T+1”表示模型所需要的帧数。

在模型参数量与计算量方面,如表 2 所示,相较于 TePose,文中模型的参数量与计算量相较于 TaPose 下降了 65%,结合 TinyYolo[25] 轻量级人体框选模型,在 720p 分辨率和 3080Ti 的环境下,TaPose 的推理速度能达到 55 ~ 60 fps,满足了对于实时性的需求。

表 2 网络参数量、计算量比较

模型	#Parameters/M	FLOPs/G
VIBE[5]	72.46	4.17
MPS-NET[13]*	108.89	4.99
TCMR[6]*	39.63	4.45
TePose[8]	39.15	0.15
TaPose	12.58	0.06

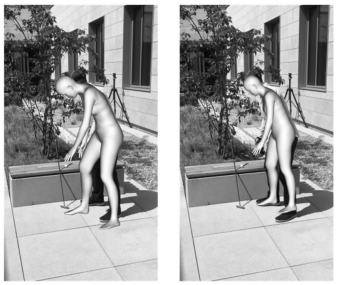

图 5 TePose 与 TaPose 的估计结果比较

与 TePose 和 VIBE 进行视觉效果比较,如图 5 所示。在对人物第一帧动作进行估计时,左图为 TePose 的估计结果,右图为文中模型的估计结果。可以看出,TePose 在估计第一帧人体动作时存在较大的误差,而文中 ss 的估计结果更为准确。

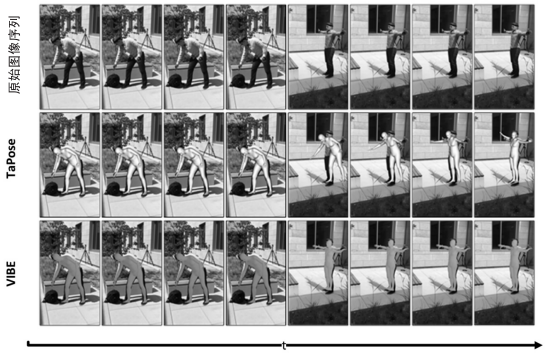

图 6 VIBE 与 TaPose 的估计结果比较

为了验证 TaPose 能够学习到连续的人体运动表示,该文使用了一个较为极端的例子进行展示。从数据集中抽取了两张不同姿势的图片并将其多次复制形成一个序列,随后将其发送到 VIBE 和 TaPose 中进行 3D 人体姿势和形状的估计,如图 6 所示。可以明显地看出,TaPose 在姿势之间产生了平滑的过渡效果,这种过渡符合人体运动的连续性,这也解释了为什么文中方法能够实现更低的 ACC-ERR 误差。相比之下,VIBE 过于依赖当前帧的信息,无法真正学习到人类运动的连续性,因此其 ACC-ERR 很高。

4 结束语

该文提出了 TaPose 网络,实现了基于实时视频流

的 3D 人体姿态和形状估计。TaPose 的总体框架包含三个部分:运动注意力机制模块、RFA 模块和基于 GCN 的运动鉴别器。该算法的损失函数结合了 2D 与 3D 数据集,并使用了对抗损失。实验结果表明,相比于其他基于视频的方法,TaPose 在精度与速度之间取得了较好的平衡。TaPose 还能够学习到连续的人体运动表示,产生了平滑的过渡效果,符合人体运动的连续性,与现有的基于实时视频流的方法相比,TaPose 具有较高的实时性与预估动作的真实性。

参考文献:

[1] ZHENG C,WU W,CHEN C,et al. Deep learning-based human pose estimation:a survey[J]. arXiv:2012.13392,2020.

[2] 裘志超,姚剑敏,严群等. 基于单张图像的人体准确姿势 3D 重建研究[J]. 传感器与微系统,2023,42(5):61-64.

[3] 叶韶华. 基于单张图像的人体姿态模型重建研究及其应用[D]. 武汉:华中科技大学,2022.

[4] KOCABAS M,HUANG C H P,HILLIGES O,et al. PARE:part attention regressor for 3D human body estimation[C]//Proceedings of the IEEE/CVF international conference on computer vision. Montreal:IEEE,2021:11127-11137.

[5] KOCABAS M,ATHANASIOU N,BLACK M J. VIBE:video inference for human body pose and shape estimation[J]. arXiv:1912.05656,2019.

[6] CHOI H,MOON G,CHANG J Y,et al. Beyond static features for temporally consistent 3d human pose and shape from a video[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville:IEEE,2021:1964-1973.

[7] LOPER M,MAHMOOD N,ROMERO J,et al. SMPL:a skinned multi-person linear model[J]. ACM Transactions on Graphics,2015,34(6):1-16.

[8] WANG Z,OSTADABBAS S. Live stream temporally embedded 3D human body pose and shape estimation[J]. arXiv:2207.12537,2022.

[9] KANAZAWA A,BLACK M J,JACOBS D W,et al. End-to-end recovery of human shape and pose[J]. arXiv:1712.06584,2017.

[10] KOLOTOUROS N,PAVLAKOS G,BLACK M J,et al. Learning to reconstruct 3D human pose and shape via model-fitting in the loop[C]//Proceedings of the IEEE/CVF international conference on computer vision. New York:IEEE,2019:2252-2261.

[11] ZENG A,YANG L,JU X,et al. Smoothnet:a plug-and-play network for refining human poses in videos[C]//Computer vision - ECCV 2022:17th European conference. Tel Aviv:Springer,2022:625-642.

[12] MAHMOOD N,GHORBANI N,TROJE N F,et al. AMASS:archive of motion capture as surface shapes[C]//Proceedings of the IEEE/CVF international conference on computer vision. Los Alamitos:IEEE,2019:5442-5451.

[13] WEI W L,LIN J C,LIU T L,et al. Capturing humans in motion:temporal-attentive 3D human pose and shape estimation from monocular video[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans:IEEE,2022:13211-13220.

[14] YAN S,XIONG Y,LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI conference on artificial intelligence. San Francisco:ACM,2018.

[15] LIU Z,ZHANG H,CHEN Z,et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle:IEEE,2020:143-152.

[16] LIC,CUI Z,ZHENG W,et al. Spatio-temporal graph convolution for skeleton based action recognition[C]//Thirty-second AAAI conference on artificial intelligence. San Francisco:ACM,2018:3482-3489.

[17] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:770-778.

[18] KINGMA D P,BA J. Adam:a method for stochastic optimization[J]. arXiv:1412.6980,2014.

[19] VON MARCARD T,HENSCHEL R,BLACK M J,et al. Recovering accurate 3d human pose in the wild using imus and a moving camera[C]//Proceedings of the European conference on computer vision (ECCV). Munich:Springer,2018:601-617.

[20] MEHTA D,RHODIN H,CASAS D,et al. Monocular 3d human pose estimation in the wild using improved cnn supervision[C]//2017 international conference on 3D vision (3DV). Qingdao:IEEE,2017:506-516.

[21] IONESCU C,PAPAVA D,OLARU V,et al. Human3.6m:large scale datasets and predictive methods for 3d human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,36(7):1325-1339.

[22] ANDRILUKA M,IQBAL U,INSAFUTDINOV E,et al. PoseTrack:a benchmark for human pose estimation and tracking[J]. arXiv:1710.10000,2017.

[23] KANAZAWA A,ZHANG J Y,FELSEN P,et al. Learning 3d human dynamics from video[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach:IEEE,2019:5614-5623.

[24] LUO Z,GOLESTANEH S A,KITANI K M. 3D human motion estimation via motion compression and refinement[J]. arXiv:2008.03789,2020.

[25] REDMON J,FARHADI A. Yolov3:an incremental improvement[J]. arXiv:1804.02767,2018.