DOI: 10.19650/j.cnki.cjsi.J2312272

基于 Transformer 的三维人体姿态估计及其动作达成度评估*

杨傲雷[1,2],周应宏[1],杨帮华[1],徐昱琳[1]
(1. 上海大学机电工程与自动化学院 上海 200444; 2. 上海市电站自动化技术重点实验室 上海 200444)

摘 要:针对人机交互、医疗康复等领域存在的人体姿态分析与评估问题,本文提出了一种基于 Transformer 的三维人体姿态估计及其动作达成度评估方法。首先,本文定义了人体姿态的关键点及关节角,并在深度位姿估计网络(DPEN)的基础上,提出并构建了一个基于 Transformer 的三维人体姿态估计模型(TPEM),Transformer 的引入能够更好的提取人体姿态的长时序特征;其次,利用 TPEM 模型对三维人体姿态估计结果,设计了基于加权 3D 关节角的动态时间规整算法,在时序上对不同人物同一动作的姿态进行姿态关键帧的规整匹配,并据此提出了动作达成度评估方法,用于给出动作的达成度分数;最后,通过在不同数据集上进行实验验证,TPEM 在 Human3.6 M 数据集上实现了平均关节点误差为 37.3 mm,而基于加权 3D 关节角的动态时间规整算法在 Fit3D 数据集上的平均误差帧数为 5.08,展现了本文所提方法在三维人体姿态估计与动作达成度评估方面的可行性和有效性。

关键词:三维人体姿态估计;深度学习;动态时间规整;动作评估
中图分类号:TP391 TH86 **文献标识码**:A **国家标准学科分类代码**:510.4050

Transformer-based 3D Human pose estimation and action achievement evaluation

Yang Aolei[1,2], Zhou Yinghong[1], Yang Banghua[1], Xu Yulin[1]
(*1. School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China*;
2. Shanghai Key Laboratory of Power Station Automation Technology, Shanghai 200444, China)

Abstract: According to the challenges of human pose analysis and assessment in domains such as human-computer interaction and medical rehabilitation, this paper introduces a Transformer-based methodology for 3D human pose estimation and the evaluation of action achievement. Firstly, key points of human pose and their joint angles were defined, and based on the deep pose estimation network (DPEN), a Transformer-based 3D human pose estimation model (TPEM) is proposed and constructed, the incorporation of Transformer facilitates better enhanced extraction of long-term sequential features of human pose. Secondly, the TPEM model's outcomes in 3D human pose estimation are utilized to formulate a dynamic time warping algorithm, which focuses on weighted 3D joint angles. This algorithm temporally aligns pose keyframes for different individuals performing the same action and subsequently introduces an assessment method for action accomplishment to provide scores for the degree of action fulfillment. Finally, through experimental validation across various datasets, TPEM achieves an average joint point error of 37.3 mm on the Human3.6 M dataset, while the dynamic time warping algorithm based on weighted 3D joint angles yields an average error of 5.08 frames on the Fit3D dataset. These results demonstrate the feasibility and effectiveness of the proposed approach for 3D human pose estimation and action accomplishment assessment.

Keywords: 3D human pose estimation; deep learning; dynamic time wrapping; action evaluation

收稿日期:2023-12-13 Received Date: 2023-12-13
*基金项目:国家重点研发计划项目资助(2023YFF1203503)、上海市自然科学基金(22ZR1424200)项目资助

0　引　　言

三维人体姿态估计,作为计算机视觉领域的焦点任务之一,专注于通过处理图像或视频流来捕捉人体各个关键骨骼节点的姿态。这项技术在人机交互[1]、医疗康复[2]。等领域中扮演着重要角色。然而,获取三维人体姿态通常依赖于昂贵的动作捕捉设备,其成本高昂且设备笨重,限制了其应用范围。随着深度学习方法在人体姿态估计领域的迅速发展,现在可以通过普通的 RGB 相机实现对三维人体姿态的估计,降低了成本并提高了便利性。

单目三维人体姿态估计作为深度学习领域的一个重点研究方向,目前主要分为两种主流方法。第一种方法采取端到端的策略,直接从单目相机获取的人体姿态图像作为输入,通过网络模型或深度信息实现三维人体姿态的估计。例如,文献[3]中提出的专注于运动中人体姿态连续性的捕捉(motion pose and shape network,MPS-Net),该模型利用残差网络 ResNet-50 提取输入视频流中图像的静态特征,通过时序编码器捕获运动姿态的时间序列特征,最终通过参数化人体模型回归得到三维姿态。文献[4]将人体图像与深度相机的深度图进行对齐匹配,然后依据人体主要关节尺寸和身高将关节像素值转换为 3D 坐标。上述端到端的方法思路简单、应用便捷,但神经网络估计的三维人体姿态误差相对较大。

第二种方法则是在现有的 2D 人体姿态估计模型的基础上先提取出图像或视频中的人体二维姿态关键点,然后进一步估计人体的三维姿态。例如,文献[5]通过图卷积网络来提取 2D 人体姿态关键点中的空间结构,使用神经网络学习到的图邻接矩阵,以加强网络对 3D 人体姿态空间信息的理解。文献[6]则将 Vision Transformer 模型迁移到三维人体姿态估计中,不仅解决了 Transformer 模型训练参数量大的问题,还验证了 Transformer 在三维人体姿态估计中的有效性。虽然 Transformer 能够有效提取时序特征,但其在姿态空间特征提取方面效果一般。文献[7]采用了二阶段法,首先根据 2D 人体姿态的空间特征通过 DPEN 估计得到三维人体姿态,然后使用门控循环单元结构(gate recurrent unit, GRU)模型对第一阶段的三维人体姿态进行时序上的平滑处理。由于现有的 2D 人体姿态估计方法已相对成熟,因此利用 2D 人体姿态进行 3D 人体姿态估计,相比直接使用端到端的三维人体姿态估计模型,其精度通常更高。而基于 2D 人体姿态的估计方法往往关注人体的结构特征或动作的时序的特征。因此本文采用二阶段法将结合人体结构与运动时序进行三维姿态估计。

动作评估作为三维人体姿态估计的一个重要应用,正在逐渐发展成为运动科学和康复训练的关键技术。在这方面,文献[8]专注于高尔夫击球动作,创建了名为“DHU-Golf”的数据集。研究者们利用 VideoPose 模型估计高尔夫动作的三维人体姿态,并通过动态时间规整算法处理。这一算法以三维人体姿态中的关节位置作为特征,在时间序列上对不同人执行的高尔夫动作姿态进行对齐。基于这一过程,研究最终能够为高尔夫击球动作提供一个客观的评级系统。这种方法不仅为高尔夫运动员提供了改进技术的参考,也为运动科学研究提供了新的视角和工具。文献[9]将人体姿态估计与计算机辅助康复环境相结合并对于这一场景中的性能限制提出了多视角三维人体姿态估计模型 EVLT-Net。该模型能够辅助病人进行详细和准确的步态分析以及为患者提供更有针对性的康复训练指导。

本文针对三维人体姿态时序估计与动作评估的问题,提出了一种基于 Transformer 的三维人体姿态估计方法和动作达成度评估框架。其主要贡献如下:提出了一种 TPEM 模型,该模型结合了 Transformer 的强大能力,有效地捕捉了长时序特征中的姿态时序信息;设计了一种基于加权 3D 关节角的动态时间规整算法,它能够有效地将不同人的姿态与标准姿态在时间序列上进行对齐,以支撑人体动作达成度的评估。

1　问题描述及方法架构

1.1　三维人体姿态估计与动作达成度评估问题

为了描述三维人体姿态估计(human pose estimation, HPE)的场景,相关坐标系如图 1 所示。从左到右分别为相机坐标系 $\{C\}$,图像坐标系 $\{I\}$ 和骨架坐标系 $\{B\}$。$\{C\}$ 的原点为相机的光心位置 O_c,光轴为 Z_c 指向相机的前方,X_c 与 Y_c 分别是水平方向与竖直方向;$\{I\}$ 中有水平坐标轴 U 和竖直坐标轴 V 来描述像素的位置;$\{B\}$ 以人体的髋骨节点作为原点,人体的正向朝向为 Z_B,X_B 和 Y_B 分别为人体的水平方向和竖直方向。

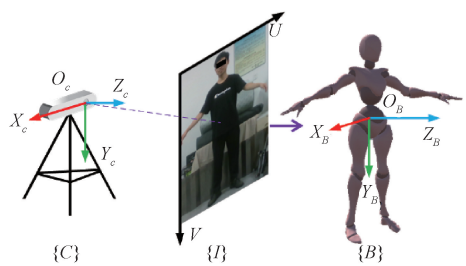


图 1　三维人体姿态估计的简单场景

Fig. 1　A simple scenario of 3D HPE

本文提出的方法是使用单目相机从视频流中估计出单人的三维人体姿态,并根据具体的动作类型将标准姿态与待评估姿态进行时序一致性匹配,以便评估动作达成度。根据问题描述,t 时刻图像坐标系的 2D 人体姿态为 ${}^{I}P_t$,t 时刻人体坐标系的 3D 人体姿态是 ${}^{B}P_t$,其定义如下:

$$ {}^{I}P_t = [{}^{I}u_t, {}^{I}v_t] \tag{1} $$

$$ {}^{B}P_t = [{}^{B}x_t, {}^{B}y_t, {}^{B}z_t] \tag{2} $$

三维人体姿态估计问题可采用式(3)表示:

$$ {}^{B}\hat{P}_t = \pi({}^{I}P_t, \omega) \tag{3} $$

其中,π 表示三维人体姿态估计模型,ω 是模型的可训练参数。为了训练 ω,约束函数为:

$$ \min_{\omega} L(\pi({}^{I}P_t, \omega), {}^{B}_{gt}P_t) \tag{4} $$

其中,${}^{B}_{gt}P_t$ 为三维人体姿态的真值。

为了方便进行统一的动作姿态评价,本文提出动作达成度指标 φ 对动作进行评估。

$$ \varphi({}^{B}_{gt}P_0, \cdots, {}^{B}_{gt}P_t, {}^{B}\hat{P}_0, \cdots {}^{B}\hat{P}_t) \tag{5} $$

其中,$\varphi(\cdot) \in [0,1]$,${}^{B}_{gt}P_t$ 与 ${}^{B}\hat{P}_t$ 分别为 t 时刻人体姿态的标准值与人体姿态的估计值,$\varphi(\cdot)$ 计算的值越大表明动作达成度越高。

1.2 整体方法与架构

本文的方法是以单目相机采集的人体 RGB 视频流为基础,提出的三维姿态估计与动作达成度评估方法架构如图 2 所示。

整体方法架构分为 3 个阶段。首先是 2D 人体姿态获取,该阶段旨在从图像中提取人体的 2D 姿态信息。在获得 2D 人体姿态数据后,经过预处理步骤,消除个体差异性,然后进入第 2 阶段 3D 人体姿态估计。在这个阶段,深度学习模型被应用于获取人体的 3D 姿态。最终,在第 3 阶段,通过对比标准姿态与模型估计的姿态,进行姿态关键帧匹配与动作达成度评估。

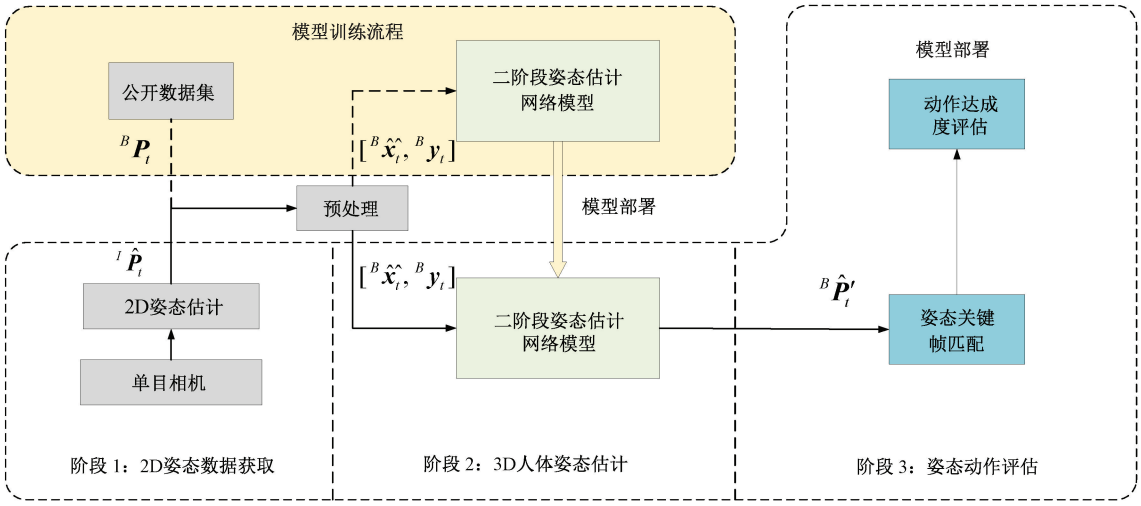


图 2 整体方法架构

Fig. 2 Architecture of the proposed method

在第一阶段的场景中,单目相机将实时采集人体的 RGB 图像,2D 人体姿态估计方法估计人体的 2D 关节点坐标。在经过数据预处理后,经过二阶段 3D 人体姿态估计网络将获得 $\{B\}$ 下的三维人体姿态。其中第一阶段的三维人体姿态为单帧的三维姿态估计,第二阶段的三维姿态估计是在时序上对第一阶段的三维人体姿态进行优化。在第三阶段,采用基于人体 3D 关节角的动态规整算法,将当前动作姿态与标准动作姿态的关键帧进行对比,分析并估计人体动作的达成度。

2 三维人体姿态估计模型构建

2.1 人体骨架与关节角定义

表 1 给出了人体关节的索引顺序表,本文定义的人体关键点共有 14 个人体姿态关键点。人体骨架图如图 3 所示。

表 1 人体关节索引表

Table 1 Skeleton joint mapping

关节索引号	关节名称	关节索引号	关节名称
0	左髋	7	头部
1	左膝	8	左肩
2	左踝	9	左肘
3	右髋	10	左腕
4	右膝	11	右肩
5	右踝	12	右肘
6	脖子	13	右腕

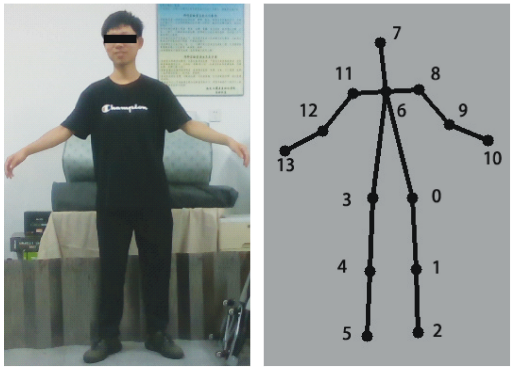


图 3　人体骨架示意图

Fig. 3　Human skeleton definition

关节角为 3 个连续相邻关节的夹角。本文根据人体骨架的定义和人体的关节活动情况定义了 8 个主要的人体关节角来描述人体的运动信息。人体关节角 φ 定义如表 2 所示。

表 2　人体关节角索引表

Table 2　Skeleton joint angle mapping

关节角索引	关节向量	关节角索引	关节向量
φ_1	7-6-8	φ_5	6-11-12
φ_2	6-8-9	φ_6	11-12-13
φ_3	8-9-10	φ_7	0-1-2
φ_4	7-6-11	φ_8	3-4-5

2.2　2D 人体姿态数据预处理

常用的 2D 人体姿态估计方法的性能指标如表 3 所示,其采用的软硬件平台与后续实验验证部分相同。

表 3　2D 人体姿态估计方法的性能

Table 3　Performance of 2D pose estimation methods

2D 姿态估计方法	准确度(AP)	实时性(FPS)
OpenPose[10]	61.8	11.5
Lightweight-OpenPose[11]	42.8	23.4
HR-Net[12]	64.1	9.3
RTM-Pose[13]	59.1	52.6

可以看到,OpenPose 和 HR-Net 能够在准确度上取得较好的效果,但实际应用中实时性不高。RTM-Pose 虽然准确度比 HR-Net 略低,但实时性比 HR-Net 有显著提升。因此,本文采用 RTM-Pose 作为 2D 人体姿态估计方法。同时,考虑到不同相机的图像分辨率和内部参数各不相同,有必要对获取的 2D 人体姿态数据进行姿态标准化处理。

在 $\{I\}$ 坐标系中的 2D 人体姿态标准化定义为:

$$ {}^{I}\hat{P}_t = [{}^{I}\hat{u}_t, {}^{I}\hat{v}_t] $$

$$ \begin{cases} {}^{I}\hat{u}_t = \dfrac{{}^{I}u_t - {}^{I}\bar{u}_t}{\sigma({}^{I}u_t) + \sigma({}^{I}v_t)} \\ {}^{I}\hat{v}_t = \dfrac{{}^{I}v_t - {}^{I}\bar{v}_t}{\sigma({}^{I}u_t) + \sigma({}^{I}v_t)} \end{cases} \tag{6} $$

其中,σ 是标准差运算符,${}^{I}\bar{u}_t$ 与 ${}^{I}\bar{v}$ 为 t 时刻 u 与 v 的平均值。同样,对于数据集中的三维人体姿态也需要进行标准化处理,${}^{B}\hat{P}_t$ 是标准化后的三维人体姿态:

$$ {}^{B}\hat{P}_t = [{}^{B}\hat{x}_t, {}^{B}\hat{y}_t, {}^{B}\hat{z}_t] $$

$$ \begin{cases} {}^{B}\hat{x}_t = \dfrac{{}^{B}x_t - {}^{B}\bar{x}_t}{\sigma({}^{B}x_t) + \sigma({}^{B}y_t)} \\ {}^{B}\hat{y}_t = \dfrac{{}^{B}y_t - {}^{B}\bar{y}_t}{\sigma({}^{B}x_t) + \sigma({}^{B}y_t)} \\ {}^{B}\hat{z}_t = \dfrac{{}^{B}z_t - {}^{B}\bar{z}_t}{\sigma({}^{B}x_t) + \sigma({}^{B}y_t)} \end{cases} \tag{7} $$

2.3　基于 Transformer 的二阶段 3D 人体姿态估计模型

视频流中的单帧图像可以承载信息,但是其中的语义信息则通过连续帧表达。一阶段模型仅仅考虑了单帧的人体空间特征却忽略了人体姿态在时序中的变化,得到的结果存在姿态抖动的情况。PFN[8] 在二阶段中使用双向 GRU 模型在时序上滤除一阶段 3D 姿态的噪声,但由于 GRU 模型在长时序中无法捕捉到有效的时序信息,会导致模型的臃肿。本文在二阶段中采用基 Transformer 的模型对一阶段单帧 3D 人体姿态在时序上进行优化,模型结构如图 4 所示。

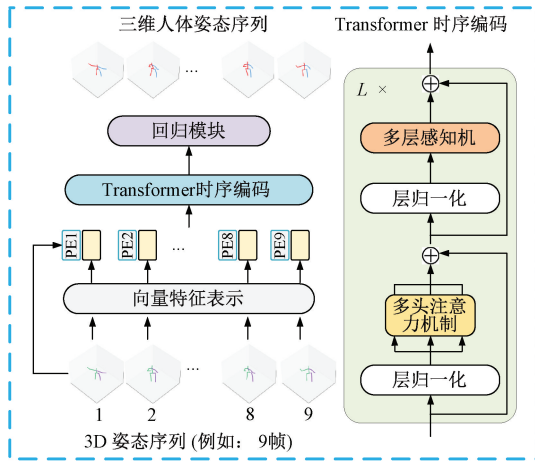


图 4　基于 Transformer 的时序优化模型

Fig. 4　TPEM Model

本文提出的基于 Transformer 的 3D 人体姿态估计模

型 TPEM 仅平滑坐标系 $\{B\}$ 深度通道的估计值${}^{B}\hat{z}_t$。模型输入一个序列矩阵(T,N,C),T 为选取的时序长度,N 为输入人体关节点的数量,C 则是每个关节的维度。在经过向量特征表示(patch embedding)后,共有 T 个 Patch 将作为 Transformer 时序编码的输入。Transformer 使用多头注意力机制在不同的长时序中提取时序特征。同时 Transformer 的输入维度与输出维度是一致的,可以堆叠 L 个 Transformer 编码模块提升模型的时序特征提取能力。在提取了高维时间语义特征后,回归模块(由两层全连接层组成)输出最后 T 帧 3D 人体姿态序列。

姿态时序优化网络模型需要学习在时间维度的分布特征,本文采用的损失函数如下:

$$\begin{cases}\mathcal{L}=\lambda_1\mathcal{L}_\Delta+\lambda_2\mathcal{L}_z\\ \mathcal{L}_\Delta=\dfrac{1}{T-1}\sum_{t=1}^{T-1}\left(\left\|{}^{B}\hat{z}_t-{}^{B}\hat{z}_{t-1}\right\|^2-\left\|{}^{B}\hat{z}_t^{gt}-{}^{B}\hat{z}_{t-1}^{gt}\right\|^2\right)^2\\ \mathcal{L}_Z=\dfrac{1}{T-1}\dfrac{1}{N}\sum_{t=1}^{T}\sum_{j=0}^{N-1}\left\|{}^{B}\hat{z}_{t,j}-{}^{B}\hat{z}_{t,j}^{gt}\right\|^2\end{cases} \tag{8}$$

其中,λ_1 与 λ_2 为超参数,$\mathcal{L}_\Delta$ 约束估计姿态时序与真值姿态时序的变化一致性,$\mathcal{L}_Z$ 约束人体姿态的深度值。

3 动态时间规整算法与动作达成度评估

3.1 基于加权 3D 关节角的动态时间规整算法

不同人做同一种动作在时序上与动作的达成度均会出现不一致的情况,为统一这种不一致性,可以采用关键帧匹配的方式进行时间归整。但是,动态时间规整算法一般根据传统的欧式距离进行时序的相似性比较,并不适用人体姿态的时间归整。因此,本文提出基于加权的 3D 人体关节角的动态时间规整算法(dynamic time warping,DTW,简称 ADTW)。由于 3D 人体姿态估计模型在不同关节角呈现出不同范围的误差分布,为了抑制由于模型精度带来的影响,采用加权的融合方式。

一个标准的运动序列 $A=\{a_1,\cdots a_i,\cdots,a_M\mid i\in[1,M]\}$ 由 M 个姿态组成,其中 a_i 为运动的分解姿态。同理,$B=\{b_1,\cdots b_j,\cdots,b_N\mid j\in[1,N]\}$ 为同一运动的不同时序姿态共有 N 个姿态组成 $N\neq M$,其中 b_j 为 B 的分解姿态。构造大小为 $M\times N$ 的距离矩阵 $\boldsymbol{D}$:

$$\boldsymbol{D}=\begin{bmatrix}d(a_1,b_n)&\cdots&d(a_m,b_n)\\ \vdots&\ddots&\vdots\\ d(a_1,b_1)&\cdots&d(a_m,b_1)\end{bmatrix} \tag{9}$$

其中,$d(a_m,b_n)$ 表示序列 A 的第 m 帧和序列 B 第 n 帧之间 3D 关节角之间的距离,表达式如下:

$$d(a_m,b_n)=\lambda_1\left|(\varphi_1^m-\varphi_1^n)\right|+\cdots+\lambda_8\left|(\varphi_8^m-\varphi_8^n)\right| \tag{10}$$

式中,φ_i^m 为 A 序列的第 i 个 3D 空间关节角,其下标最大值为 8,与表 2 中关节角索引对应;φ_j^n 为 B 序列的第 j 个 3D 空间关节角;λ_i 为每组 3D 关节角之间的权重系数。

代价矩阵 $\boldsymbol{C}$ 如式(11)所示,其大小为$(M+1)\times(N+1)$。除了 $\boldsymbol{C}(0,0)=0$,$\boldsymbol{C}$ 的第 0 行和第 0 列中其它所有位置初始化为 ∞。这样,从矩阵 $\boldsymbol{C}(M,N)$ 到 $\boldsymbol{C}(0,0)$ 回溯找到一条最短路径,该路径表示了两个动作序列的姿态对应关系。

$$\boldsymbol{C}_{i,j}=\boldsymbol{D}(i,j)+\min\begin{cases}\boldsymbol{C}_{i-1,j-1}\\ \boldsymbol{C}_{i-1,j}\\ \boldsymbol{C}_{i,j-1}\end{cases} \tag{11}$$

3.2 动作达成度评估

在经过 ADTW 算法后,分解后的标准姿态与 TPEM 模型估计的姿态完成关键帧对齐。对于标准姿态而言,其中动作关键帧姿态为 M 个,整个动作的达成度 δ 计算公式如下:

$$\delta=\frac{1}{8M}\sum_{i=1}^{M}\sum_{j=1}^{8}k_{i,j} \tag{12}$$

$$k_{i,j}=\begin{cases}1-\dfrac{\left\|\varphi_{i,j}^{A}-\varphi_{i,j}^{B'}\right\|}{90}, & \left\|\varphi_{i,j}^{A}-\varphi_{i,j}^{B'}\right\|<=90\\ 0, & \left\|\varphi_{i,j}^{A}-\varphi_{i,j}^{B'}\right\|>90\end{cases} \tag{13}$$

其中,$k_{i,j}$ 为第 i 个关键帧中第 j 个角的分值,$k_{i,j}\in[0,1]$。$\varphi_{i,j}^{A}$ 与 $\varphi_{i,j}^{B'}$ 分别是标准姿态第 i 个关键帧中的第 j 个角与模型估计并在时序对齐后的第 i 个关键帧中的第 j 个角。$k_{i,j}$ 取值随着误差角的变化而不同。当 3D 误差角越小,则 $k_{i,j}$ 的取值越大。当 3D 误差角大于 90°时,则 $k_{i,j}$ 的值为 0,这意味着该姿态为错误姿态。

4 模型训练评估与实验

4.1 数据集与实验平台

本文采用 Human3.6M[14] 数据集进行模型训练和测试,选取编号为 S1, S5, S6, S7, S8 的表演者作为训练集。为了验证 ADTW 算法的有效性,本文选取 Fit3D[15] 中编号为 S3 与 S4 教练员的姿态作为分析与验证数据。该数据集与 Human3.6 M 采集人体姿态的场景一致,不同之处在于 Fit3D 是由 11 位专业健身教练分别完成不同的 47 种健身动作。

硬件环境:NVIDIA GTX 2080 8 G 显存,Intel Core i5-10200H,相机 Real-Sense D435i,通道模式 RGB,分辨率为 640×480。软件环境:PyTorch 平台。

4.2 TPEM 模型评估

本文使用平均关节位置误差(MPJPE)和普氏分析平

均关节位置误差(PA-MPJPE)作为 TPEM 模型的评价指标。公式如下:

$$E_{MPJPE}=\frac{1}{N}\sum_{j=0}^{N-1}\| {}^{B}\hat{p}_j-{}^{B}\hat{p}_j^{gt}\| \tag{14}$$

其中,${}^{B}\hat{p}_j$ 为标准化后 TPEM 模型估计的第 j 个人体姿态关节点,${}^{B}\hat{p}_j^{gt}$ 为标准化后第 j 个人三维人体关节真值。PA-MPJPE 则是将预测关节与真值关节刚性对齐后的 MPJPE。

TPEM 超参数中时序长度为 25 帧,L 的层数为 4 层,多头注意力机制为 8 个。表 4 列出了测试集中 13 种动作的 MPJPE 结果,表 5 列出了测试集中 13 种动作的 PA-MPJPE 结果。图 5 将部分动作的三维姿态进行可视化,图 5(a)为原 RGB 图像,图 5(b)为三维人体姿态估计的真值,图 5(c)为 TPEM 估计的三维人体姿态结果。

表 4　TPEM 在 Human3.6M 数据集上的 MPJPE 定量分析

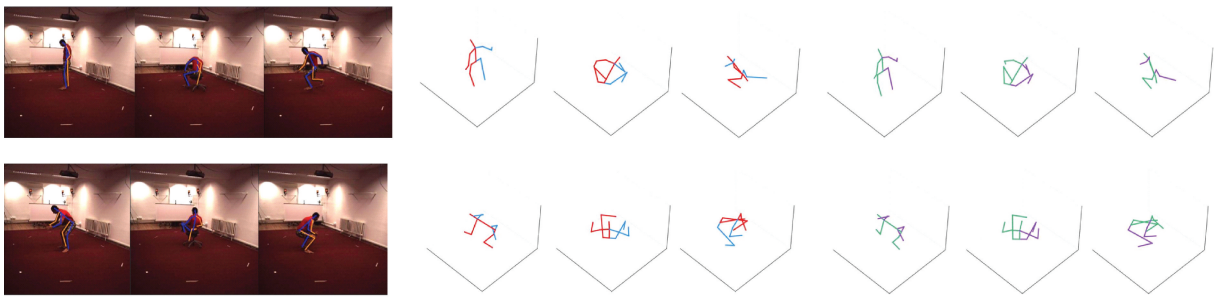
Table 4　Quantitative analysis of TPEM on the Human3.6M dataset with MPJPE

MPJPE	指向	讨论	吃饭	问候	电话	照相	姿势	购物	坐着	蹲下	吸烟	走路	等待	平均
SemGCN[16]	37.8	49.4	37.6	40.9	45.1	41.1	40.1	48.3	50.1	42.2	53.5	42.26	44.3	44.1
VideoPose[17]	45.1	47.4	42	46	49.1	56.7	44.5	44.4	57.2	66.1	47.5	38.6	44.8	48.4
GraFormer	32	38	30.4	34.4	34.7	43.3	35.2	31.4	38	46.2	34.2	31.4	35.7	35.8
PFN	37	42.6	34.9	38.9	39.6	46.2	41.2	33	46	52.9	40.3	31.7	42.4	40.5
TPEM	32.6	37.0	33.5	37.0	38.5	45.6	37.6	30.9	42.2	47.5	36.5	28.7	37.6	37.3

表 5　TPEM 在 Human3.6M 数据集上的 PA-MPJPE 定量分析

Table 5　Quantitative analysis of TPEM on the Human3.6M dataset with PA-MPJPE

PA-MPJPE	指向	讨论	吃饭	问候	电话	照相	姿势	购物	坐着	蹲下	吸烟	走路	等待	平均
VideoPose	34.2	36.8	33.9	37.5	37.1	43.2	34.4	33.5	45.3	52.7	37.7	30.5	34.1	37.8
文献[18]	38.5	42.7	43.9	46.1	49.1	53.2	41.0	39.8	53.9	63.8	48.1	37.6	43.9	46.3
文献[19]	38.2	41.3	43.5	44.4	45.4	54.7	39.3	38.0	53.2	59.2	45.0	33.0	40.7	44.3
PFN	30.5	42.4	34.5	30.6	41.2	47.1	41.2	31.4	47.1	53.2	31.3	32.6	30	37.9
TPEM	27.5	32.6	28.3	32.4	33.9	39.0	31.9	27.8	36.8	43.4	33.1	25.4	32.4	32.7



(a) 二维人体姿态
(a) 2D human pose

(b) 三维人体姿态真实值
(b) Ground truth of 3D human pose

(c) 三维人体姿态估计值
(c) Estimation of 3D human pose

图 5　TPEM 回归模型效果图

Fig. 5　TPEM regression model renderings

可以发现,SemGCN[16] 利用图卷积神经网络进行端到端的 3D 人体姿态估计,虽然图卷积能够有效利用人体骨架特征,但是 SemGCN 忽略了人体姿态的时序特征。VideoPose[17] 利用时空卷积模型在时序上取得了一定的效果,但是在模型架构中忽略了人体的空间特征。DPEN 由多个残差块组成,结构简单,GraFormer 利用注意力机制与图卷积提取人体关节之间的关联特征具有较好的效果。本文以 DPEN 为基础,在时序上对人体姿态进行优化,具有一定的新意,并将在后续工作中引入图卷积神经网络,以期进一步优化三维人体姿态估计。

此外,PFN 方法先利用人体姿态的空间特征估计

出待优化的三维人体姿态,随后利用 GRU 网络在时序上对一阶段估计的三维人体姿态进行优化,但是 GRU 网络在长时序的表现不佳,在时序长度大于 25 帧时,MPJPE 的误差开始逐步增大。而 TPEM 利用 Transformer 中多头注意力机制,能够在长时序中有效提取 3D 人体姿态的高维时序特征。图 6 展示了 PFN 与 TPEM 在不同时序长度的表现。表 6 列出了 TPEM 模型在一阶段之后的姿态优化效果。仅使用一阶段的网络模型 DPEN 的 MPJPE=46.9 mm,使用 PFN 进行姿态优化后误差减少了 13.7%,而使用 TPEM 进行姿态优化后误差进一步减少到 20.5%。结果表明,基于 Transformer 的 3D 姿态优化模型 TPEM 能够较好的对一阶段模型估计的结果进行优化。

本文在实际场景中对 TPEM 模型进行了自建数据的测试,实验效果如图 7 所示。可以看到,即使图像中的人物出现了部分关节的遮挡或是人物有一些复杂动作,TEMP 仍能估计出人物的三维人体姿态。

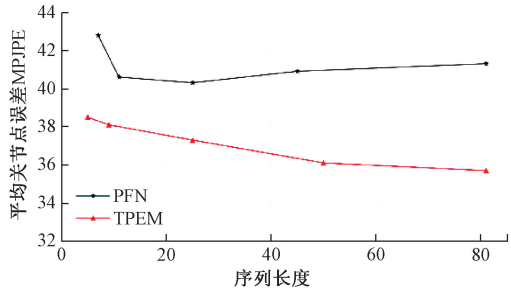


图 6 序列长度对 MPJPE 的影响

Fig. 6 Impact of sequence length on MPJPE

表 6 消融实验

Table 6 Ablation study

方法	MPJPE	Δ/%
DPEN	46.9	-
DPEN+PFN	40.5	-13.7
DPEN+TPEM	37.3	-20.5

(a) 实际场景
(a) Practical scenario

(b) TPEM模型验证效果
(b) Validation results of the TPEM

图 7 TPEM 模型的实际场景效果图

Fig. 7 Practical scenario validation results of the TPEM

4.3 ADTW 方法评估

为验证 ADTW 算法的有效性,本文将数据集 Fit3D 的健身动作拆解为一系列姿态组合的形式,通过真值匹配与 ADTW 的匹配进行分析。TPEM 模型的平均关节角误差和 ADTW 关节角权重系数如表 7 所示。表中的平均关节角误差为 TPEM 模型在 Human3.6M 整个数据集上的分析结果。

表 7 平均关节角误差与 ADTW 权重系数

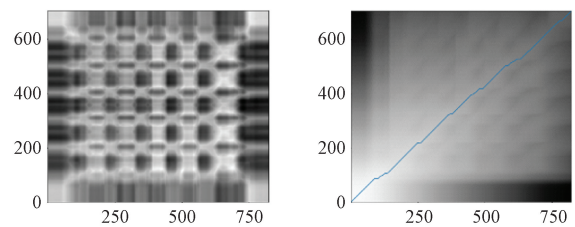
Table 7 Average joint angle error and ADTW weighting coefficients

关节角	平均关节角误差	ADTW 权重系数
φ_1	5.97	0.11
φ_2	4.05	0.16
φ_3	6.45	0.10
φ_4	5.22	0.12
φ_5	5.24	0.12
φ_6	3.60	0.18
φ_7	6.84	0.09
φ_8	5.38	0.12

误差指标 E_{mt} 的定义如下:

$$E_{mt}=\frac{1}{N}\sum_{i=0}^{N}\|f_i^{gt}-f_i\| \tag{15}$$

其中,N 为动作分解出的姿态序列,f_i^{gt} 为姿态匹配关键帧的真值,f_i 为 ADTW 算法得到的匹配关键帧的估计值。Fit3D 的摄像机拍摄的 FPS 为 50 帧/s,ADTW 方法计算所得的 E_{mt} 为 5.08 帧,取得了良好的匹配效果。以扩胸运动为例,图 8 展示了 S03 与 S04 的距离矩阵 $\boldsymbol{D}$ 与代价矩阵 $\boldsymbol{C}$,其中纵坐标为 S03 的姿态帧,横坐标为 S04 的姿态帧,代价矩阵中的曲线为最短匹配路径。表 8



(a) 距离矩阵
(a) Distance matrix

(b) 代价矩阵
(b) Cost matrix

图 8 距离矩阵与代价矩阵

Fig. 8 Distance matrix and cost matrix

展示了 S03 与 S04 扩胸运动的达成度评估结果。根据式(12)与式(13),单帧评分即为 $k_{i,j}$,最终该动作达成度的分数为 92.46 分。

表 8　S04 扩胸运动达成度评分

Table 8　S04 band pull apart achievement score

关键帧名称	S03 关键帧	S04 匹配真值	S04 估计值	单帧评分
准备动作	0	0	0	93.28
手臂举起	108	130	138	93.53
第一次扩胸	156	179	186	90.71
第一次扩胸结束	203	238	233	93.57
第二次扩胸	260	290	302	90.68
第二次扩胸结束	311	345	353	92.26
第三次扩胸	360	415	413	90.10
第三次扩胸结束	405	465	466	92.32
第四次扩胸	454	530	528	93.47
第四次扩胸结束	503	573	586	92.02
第五次扩胸	552	655	658	93.65
完成手放下	705	823	823	93.96

5　结　　论

本文的研究工作主要集中于提出了一种基于 Transformer 的神经网络模型 TPEM,用于三维人体姿态估计,以及一种基于加权 3D 关节角的动态时间规整算法,用于动作达成度的评估。通过在 Human3.6M 数据集上的实验,证实了 TPEM 模型在三维人体姿态估计方面的有效性,显示出较低的平均关节点误差。同时,在 Fit3D 数据集上的实验表明,提出的基于加权 3D 关节角的动态时间规整算法能够成功地在不同人执行同一动作时进行姿态关键帧的匹配。此外,动作达成度评估部分通过对时序对齐后的姿态结果进行分析,为每个动作提供了详细的达成度评分,这对于运动训练、医疗康复等领域具有重要的实用价值。尽管取得了一定的成果,但仍存在一些需要改进的地方。目前的 TPEM 模型在 Human3.6M 和 Fit3D 上表现良好,但其泛化能力在复杂环境下尚未得到充分验证。Transformer 模型通常需要较大的计算资源,这可能限制模型在资源受限环境中的应用。

参考文献

[1] 杨傲雷, 陈燕玲, 徐昱琳. 基于强化学习的机器人手臂仿人运动规划方法[J]. 仪器仪表学报, 2021, 42(12):136-145.
YANG AO L, CHEN Y L, XU Y L. Humanoid motion planning of robotic arm based on reinforcement learning[J]. Chinese Journal of Scientific Instrument, 2021, 42(12):136-145.

[2] 唐心宇, 宋爱国. 人体姿态估计及在康复训练情景交互中的应用[J]. 仪器仪表学报, 2018, 39(11):195-203.
TANG X Y, SONG AI G. Human pose estimation and its implementation in scenario interaction system of rehabilitation training[J]. Chinese Journal of Scientific Instrument, 2018, 39(11):195-203.

[3] 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 Computer Society Conference on Computer Vision and Pattern Recognition, 2022:13201-13210.

[4] 刘今越, 刘彦开, 贾晓辉, 等. 基于模型约束的人体姿态视觉识别算法研究[J]. 仪器仪表学报, 2020, 41(4):208-217.
LIU J Y, LIU Y K, JIA X H, et al. Research on human pose visual recognition algorithm based on model constraints[J]. Chinese Journal of Scientific Instrument, 2020, 41(4):208-217.

[5] ZHAO W, WANG W Q, TIAN Y J. GraFormer: Graph-oriented transformer for 3D pose estimation[C]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2022:20406-20415.

[6] ZHENG C, ZHU S, MENDIETA M, et al. 3D Human pose estimation with spatial and temporal transformers[C]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2021:11636-11645.

[7] YANG A L, LIU G C, NAEEM W, et al. A monocular 3D human pose estimation approach for virtual character skeleton retargeting[J]. Journal of Ambient Intelligence and Humanized Computing, 2023, 4:9563-9574.

[8] ZHANG Y T, WANG Q A, TU F Y, et al. Automatic moving pose grading for golf swing in sports[C]. IEEE International Conference on Image Processing, 2022: 41-45.

[9] XU W, XIANG D H, WANG G T, et al. Multiview video-based 3-D pose estimation of patients in computer-assisted rehabilitation environment[J]. IEEE Transactions on Human-Machine Systems, 2022, 52(2): 196-206.

[10] CAO Z, HIDALGO G, SIMON T, et al. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis

and Machine Intelligence, 2022, 43(1):172-186.
[11] OSOKIN D. Real-time 2D multi-person pose estimation on CPU: Lightweight OpenPose[C]. Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, 2019:744-748.
[12] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:5686-5696.
[13] JIANG T, LU P, ZHANG L, et al. RTMPose: Real-time multi-person pose estimation based on MMPose[J]. ArXiv Preprint,2023, ArXiv:2303.07399.
[14] 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, 2014, 36(7):1325-1339.
[15] FIERARU M, ZANFIR M, PIRLEA S C, et al. AIFit: Automatic 3D human-interpretable feedback models for fitness training[C]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2021:9914-9923.
[16] ZHAO L, PENG X, TIAN Y, et al. Semantic graph convolutional networks for 3D human pose regression[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:3420-3430.
[17] PAVLLO D, FEICHTENHOFER C, GRANGIER D, et al. 3D human pose estimation in video with temporal convolutions and semi-supervised training[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:7753-7762.
[18] CHEN Y, SHEN C, CHEN H, et al. Adversarial learning of structure aware fully convolutional networks for landmark localization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(7):1654-1669.
[19] LI Y, LI K, JIANG S, et al. Geometry-driven self-supervised method for 3D human pose estimation[C]. AAAI 2020-34th AAAI Conference on Artificial Intelligence, 2020: 11442-11449.

作者简介

杨傲雷(通信作者),2004 年于湖北工业大学获得学士学位,2009 年于上海大学获得硕士学位,2012 年于英国女王大学获得博士学位,现为上海大学副教授,主要研究方向为机器人与视觉控制、计算机视觉与感知定位等。

E-mail: aolei@shu.edu.cn

Yang Aolei (Corresponding author) received his B.Sc. degree in 2004 from Hubei University of Technology, M.Sc. degree in 2009 from Shanghai University and Ph.D. degree in 2012 from Queen's University Belfast, UK. Now he is an associate professor in Shanghai University. His main research interests include robotics and vision control, computer vision and perception localization, etc.

周应宏,2021 年于安徽农业大学获得学士学位,现为上海大学硕士研究生,主要研究方向为计算机视觉。

E-mail: zhouyinghon@shu.edu.cn

Zhou Yinghong received his B.Sc. degree in 2021 from Anhui Agriculture University. Now, he is currently a M.Sc. Candidate at Shanghai University. His main research interest is computer vision.

杨帮华,1993 年于西安工程大学获得学士学位,1996 年于西安工程大学获得硕士学位,2006 年于上海交通大学获得博士学位,现为上海大学机电工程与自动化学院教授,主要研究方向为脑机接口、生物信号处理和深度学习。

E-mail: yangbanghua@shu.edu.cn

Yang Banghua received her B.Sc. degree in 1993 year from Xi'an Polytechnic University, received her M.Sc. degree in 1996 year from Xi'an Polytechnic University, received her Ph.D. degree in 2006 year from Shanghai Jiao Tong University. Now she is a professor in School of Mechatronic Engineering and Automation, Shanghai University. Her main research interests include Brain Computer Interface, Biological Signal Processing and Deep Learning.