计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(4):187−193 [doi: 10.15888/j.cnki.csa.009467]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进的三维人体姿态估计算法①

陈荣桂, 贾振堂

(上海电力大学 电子与信息工程学院, 上海 201306)
通信作者: 贾振堂, E-mail: 462458081@qq.com

摘　要: 针对目前三维人体姿态由于遮挡、姿态复杂等预测不准确的问题, 提出了一种改进的三维人体姿态估计算法以获得准确的三维人体姿态, 提高人体姿态估计性能. 本文采用时空图注意力卷积网络中的图注意力块来构建整个网络, 在此基础上对全局多头图注意力部分的网络结构进行改进, 使节点间更好传播和融合信息, 捕获图中没有显式表示的语义信息. 同时引入运动学约束, 在 *MPJPE* 损失的基础上, 加上骨骼长度损失. 通过对局部和全局的空间节点信息建模, 实现对局部运动学连接、对称性和全局姿态的人体骨骼运动学约束的学习. 通过实验证明, 本文改进后的模型有效地提高了人体姿态估计性能, 在 Human3.6M 数据集上相较于原始模型, 实现了 1.8% 的平均关节位置误差 (*MPJPE*) 提升和 1.3% 的预测关节与真值关节刚性对齐后的平均关节位置误差 (*P-MPJPE*) 提升.

关键词: 三维人体姿态估计; 图注意力卷积; 骨骼长度损失; 深度学习

引用格式: 陈荣桂,贾振堂.改进的三维人体姿态估计算法.计算机系统应用,2024,33(4):187–193. http://www.c-s-a.org.cn/1003-3254/9467.html

Improved Algorithm for 3D Human Pose Estimation

CHEN Rong-Gui, JIA Zhen-Tang

(College of Electronics and Information Engineering, Shanghai University of Electric Power, Shanghai 201306, China)

Abstract: Aiming at the current inaccurate predictions in 3D human pose due to factors such as occlusion and complexity of poses, this paper proposes an improved 3D human pose estimation algorithm to obtain accurate 3D human pose and enhance the performance of human pose estimation. Meanwhile, it adopts the graph attention block from the spatio-temporal graph attention convolutional network to construct the entire network. On this basis, the network structure of the global multi-head graph attention part is improved to facilitate better information propagation and fusion among nodes and capture semantic information not explicitly represented in the graph. Kinematic constraints are introduced as well, and a bone length loss is added based on the *MPJPE* loss. By the modeling of local and global spatial node information, the learning of kinematic constraints of human skeletal movements is achieved, including local kinematic connections, symmetry, and global poses. Empirical results show that the improved model effectively enhances the performance of human pose estimation. Compared to the original model on the Human3.6M dataset, a 1.8% improvement in mean per joint position error (*MPJPE*) and a 1.3% improvement in the Procrustes aligned *MPJPE* (*P-MPJPE*) after rigid alignment of predicted and true joints have been realized.

Key words: 3D human pose estimation; graph attention convolution; bone length loss; deep learning (DL)

近年来, 人体姿态估计作为计算机视觉领域的热门研究课题, 受到了广泛的瞩目. 三维人体姿态估计的目标是通过分析二维图像中信息, 推断出对应的三维空间中人体关键点的位置和姿态[1], 它本质上是一个回

① 基金项目: 国家自然科学基金 (62105196)
收稿时间: 2023-09-12; 修改时间: 2023-10-09, 2023-11-24; 采用时间: 2023-12-07; csa 在线出版时间: 2024-01-30
CNKI 网络首发时间: 2024-02-01

归问题. 这项任务对于实现自然的人机交互[2]、逼真的虚拟角色生成[3]、行为分析[4]、医疗辅助[5]具有重要意义, 它也可以作为其他算法的辅助环节, 为其他计算机视觉任务 (如动作识别[6]) 提供人体骨骼等方面的信息.

单目三维人体姿态估计的研究方法主要是两种. 一种是从图片直接回归得到 3D 坐标[7–9]. 另一类是先获取 2D 信息再提升到 3D 姿态[10,11], 分两个阶段完成. 第 2 种方法又分为两类. 一类是端到端的网络, 输入图片后经过 2D 和 3D 姿态网络共同训练. 还有一类是直接用预训练的 2D 姿态网络, 把获取到的 2D 关键点坐标作为 3D 姿态估计网络的输入来估计相应 3D 姿态, 由于 2D 人体姿态估计技术的成熟和发展, 此种方法得到很好的流行并且取得了不错的效果[12], 常采用的 2D 姿态估计器有: Hourglass[13]、CPN[14]和 HRNet[15].

Martinez 等[7]提出基于深度学习的全连接网络, 大大提高了三维姿态估计的精度, 由于网络中的密集连接和缺乏人体先验知识, 导致对二维信息较敏感. Kocabas 等[16]利用多个视角的二维姿态作为输入, 通过多视角几何学对三维姿态进行自我监督学习, 缓解了遮挡问题. Wandt 等[17]用全连接层构建了弱监督的生成式对抗网络, 学习 3D 坐标和相机参数, 解决了映射约束的忽略及模型过拟合的问题, 对其他数据也拥有较好泛化能力. Zhao 等[11]提出语义图卷积网络, 把卷积神经网络推广到更多非欧几里得结构, 通过捕获节点间语义关系对人体关节的拓扑关系进行建模, 增设可学习的权重矩阵, 用骨长损失和关键点损失共同训练网络. Liu 等[18]基于语义图卷积, 用邻接矩阵和对称矩阵来表示图结构, 结合多头注意力对所有关节关系进行约束, 构建了时空图注意力卷积网络, 降低了空间上不存在的 3D 姿态出现的可能性, 在一定程度上缓解了遮挡的问题.

本文旨在探索有效的方法, 从二维关键点数据中恢复出精确的三维人体姿态. 选用较好鲁棒性的 CPN (级联金字塔网络) 作为 2D 姿态预测方法, 从单目图片获取到 2D 关键点, 将其作为网络的输入然后预测对应的 3D 姿态. 受 Zhao 等[11]和 Liu 等[18]使用图卷积预测三维姿态的启发, 本文采用时空图注意力卷积网络中的图注意力块来构建整个网络, 引入骨骼长度损失的人体运动学约束, 并对全局多头注意力部分的网络结构进行改进, 最后在 Human3.6M 数据集上验证了本文改进方法的有效性.

1 相关工作

1.1 局部图注意力网络

人体姿态的关节点可以被视为图中的节点, 它们之间的关联可以被表示为图中的边, 这种图结构天然契合了人体姿态的特点. 传统的卷积神经网络对于非规则的图结构表现较为困难, 而图卷积网络则可以很好地捕捉节点之间的连接关系. 用$G=(V,E)$来表示一个图, V代表节点的集合, E代表边的集合, 每个节点i对应的特征用x_i表示, 其相邻节点为$j\in N(i)$, 若有N节点, 每个节点特征数是D, 则可用X_{N*D}作为节点特征表示. 图卷积公式定义为:

$$X^{(l+1)}=\sigma(WX^{(l)}\tilde{A}) \tag{1}$$

其中, $\tilde{A}$表示邻接矩阵A对称归一化, $X^{(l)}\in R^{D_l\times K_l}$表示第$l$层特征向量, $X^{(l+1)}$表示第$l+1$层的特征向量, $W\in R^{D_{l+1}\times D_l}$是可学习的参数矩阵, σ是一个非线性激活函数 ReLU.

由于图卷积 (GCN) 的局限性, 对所有边都共享参数矩阵W, 使图的节点以及内部结构的关系没有得到很好利用, 且只考虑到了一阶邻接节点的信息不利于全局特征的学习. 语义图卷积 (SemGCN) 的提出一定程度上解决了这些问题, 该方法将 CNN 中单个卷积核提取多维特征的思想类比到 GCN 上, 使 GCN 对每个节点的特征学习独立的权重向量, 然后通过共享的转换矩阵来结合这些特征向量. 这样使得 GCN 在节点间更好传播并融合信息, 学习捕获在图中没有显式表示的语义信息. SemGCN 在 GCN 基础上增加了一个可学习的加权矩阵, 其公式定义为:

$$X^{(l+1)}=\sigma(WX^{(l)}\rho_i(M\odot A)) \tag{2}$$

其中, ρ_i是 Softmax 非线性函数; $M\in R^{K\times K}$是可学习的加权矩阵; $\odot$是一个元素运算; A为掩码, 强制对节点i进行计算, 仅计算其相邻节点$j\in N(i)$的权重.

基于 SemGCN 的思想, Liu 等[18]提出了时空图注意力卷积网络. 在该网络的局部图注意力模块中, 对输入的数据增设两个图矩阵, 将关联的节点用矩阵表达. 利用 SemGCN 中定义的 SemCHGraphConv 网络框架, 以这个框架为基础输入 adj_sym 和 adj_con 得到局部图注意力架构, 其中 adj_sym 是对左右对称的关节点进行关联的矩阵, adj_con 是将一阶和二阶邻接矩阵相加得到的一阶二阶联合邻接矩阵. 由于关键点之间存在着很强的依赖关系, 局部图注意力模块不仅考虑到

以躯干为中心的人体的对称结构, 且加入了二阶相邻关节的约束, 建模这种依赖关系对缓解歧义有重要作用, 尤其是发生严重遮挡时. 用邻接矩阵A和单位矩阵I来初始化图的结构, $\tilde{A}=(A+I)$, 给定第l层的节点特征, 通过卷积后得到$l+1$层节点的特征:

$$X^{(l+1)}=\mathop{\Big\|}_{c=1}^{C_{l+1}}\rho(M_c\odot\tilde{A})X^{(l)}W_c \tag{3}$$

其中, ρ是 Softmax 非线性函数, $M_c\in R^{K\times K}$是可学习掩码矩阵, $\odot$是逐元素乘法运算, W_c是矩阵W的第c行.

1.2 全局多头图注意力网络

除了直接连接的关节点间的重要关系, 没有相连的关节 (如手腕和踝关节) 之间的关系在编码全局姿势和约束信息中发挥着重要作用, 捕捉节点之间的全局依赖, 建立不同关节点之间的关系, 能在一定程度上有效地解决卷积核感受野有限和遮挡问题.

Liu 等[18]将全局注意力机制与 GCN 相结合, 提出了一种多头注意力机制的图卷积网络, 能有效自适应地编码非局部关节点的关系, 通过注意力机制对全局空间信息进行建模, 多头注意力通过并行计算单个注意力机制来增强模型的表达能力, 改进了对人体全局姿势的学习. 在 GCN 中, 每个顶点的新特征是其邻居特征的加权平均, 而全局注意力机制则可以用来确定这些权重. 全局多头图注意力网络对式 (3) 进行了拓展:

$$X^{(l+1)}=\mathop{\Big\|}_{k=1}^{K}(B_k+C_k)X^{(l)}W_k \tag{4}$$

其中, K是注意力头的数量且K为 4, $B_k\in R^{N\times N}$是自适应全局邻接矩阵, $C_k\in R^{N\times N}$是可学习全局邻接矩阵; $W_k\in R^{C_l\times(C_l/K)}$是变换矩阵, B_k表示一个数据相关矩阵, 为每个节点学习一个唯一的图. 通过注意力系数函数[19]来确定节点之间是否存在连接以及连接强度的多少. C_k是一个可学习邻接矩阵, C_k的元素是任意的, 在训练过程中随着整个网络一同训练. C_k与式 (3) 的掩码矩阵M_c所执行的注意力机制类似但M_c比M_c更灵活.

2 改进的三维人体姿态估计算法

2.1 网络结构

本文改进的三维姿态估计网络建立在先进的二维姿态估计算法的基础上, 用 CPN (级联金字塔网络) 从输入的图片中获取二维关键点, 用二维关键点作为本网络的输入获取对应的三维关键点. 图 1 所示是本文改进后的三维人体姿态估计算法的网络结构图, 基于文献[18]的图注意力块搭建而成, 原始图注意力块由局部图注意力网络和全局多头图注意力网络两个模块的结果按通道拼接而成, 在此基础上改进了图注意块中的全局多头图注意力网络的内部结构, 同时引入残差连接.

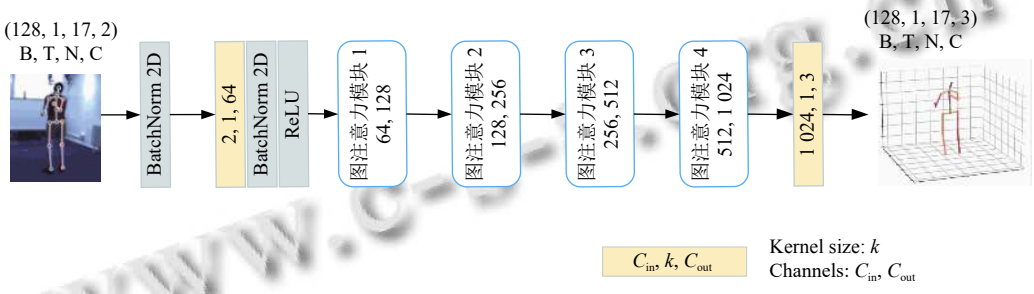


图 1 网络结构图

本文方法中的图注意力块如图 2 所示, 左侧为全局多头图注意力网络, 右侧为局部图注意力网络. 局部图注意力网络分为两个分支, 使用语义图卷积中的 SemCHGraphConv 网络框架, 分别输入 adj_sym 和 adj_con, 此时通道数为输入通道数的 2 倍, 将得到的结果拼接在一起, 然后通过卷积核为 1 的卷积使输入输出通道数保持一致, 对局部空间信息进行建模, 利用局部运动学连接和对称性信息, 提高模型的预测能力. 在全局多头图注意力网络中, 引入了 4 个独立的图注意力头, 每个图注意力头将通道数下采样到其 1/4, 然后将 4 个图注意力头的结果拼接在一起, 同样通过卷积核为 1 的卷积使输入输出通道数保持一致, 这允许模型在不同的表示子空间中分别关注不同的信息, 并行地学习不同的表示子空间, 从而捕捉不同层次和角度的信息, 有效学习到人体全局姿势信息. 将全局和局部这两部分的结果按通道拼接在一起, 再加上原本的输入, 通过卷积核为 1 的卷积使输入输出通道数保持一致. 整个图注意力块的输出通道变为输入通道的 2 倍.

网络中的卷积之后有批量归一化层和 ReLU 线性修正单元, 用于加速深度神经网络训练, 帮助缓解梯度消失和爆炸问题, 有助于训练更深的网络. 每个图注意力块中加入了残差连接, 允许原始输入信息直接传递到下一层网络, 这有助于网络保留更多的细节和信息, 学习到更有用的特征表示.

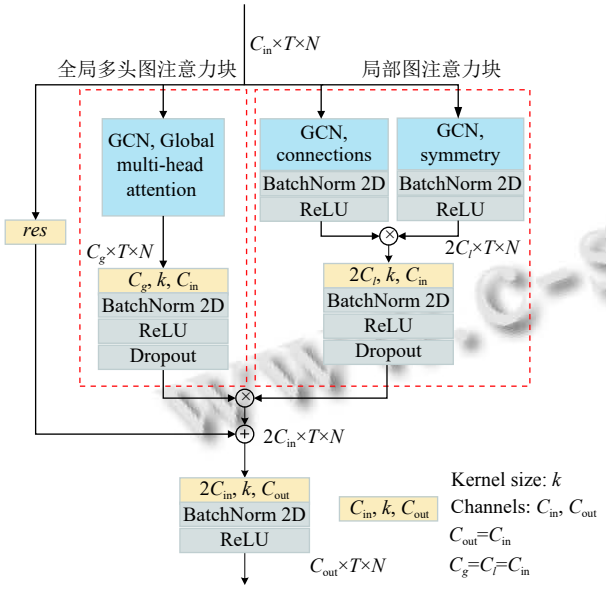


图 2 图注意力块

2.2 网络结构的改进

受文献[20]的启发, 本文对全局多头图注意力网络进行改进, 图 3 是改进后的全局多头图注意力网络结构图.

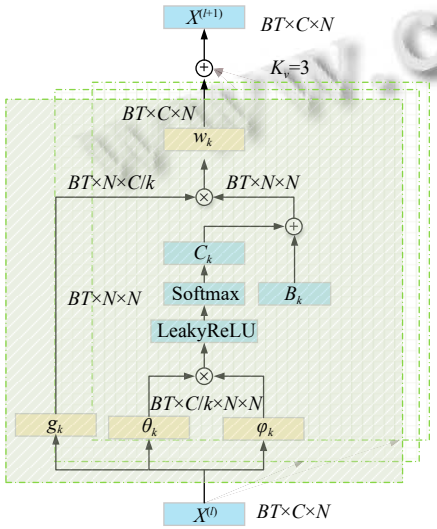


图 3 改进后的全局多头图注意力网络结构图

在原始全局多头图注意力基础上, 增设 3 个子集空间, 再将 3 个子集空间的全局多头图注意力网络的输出结果相加. 改进后的全局多头图注意力网络和局部图注意力网络两个模块的结果按通道拼接组成新的图注意力块, 并引入残差连接. 与式 (4) 一样设置为 4 个图注意力头, 并行地计算每一个头将结果按通道数拼接在一起, 组成 4 个头即多头图注意力. 对式 (4) 进一步改进, 设置K为 3 代表子集个数, 将 3 个子集的结果累加作为每个头的输出. 改进的图卷积公式定义为:

$$X^{(l+1)}=\Big\Vert_{i=1}^{I}\sum_{k=1}^{K_v}W_kX^{(l)}(B_k+C_k) \tag{5}$$

其中, I表示图注意力头为 4, K_v表示子集个数为 3. B_k是一个 17×17 的矩阵, 和式 (4) 中B_k一样, 是一个数据依赖图, 为每个节点学习生成唯一的图, 通过注意系数函数捕捉关节之间的相似性, 来确定节点之间是否存在连接以及连接强度的多少. 具体来说, 将输入特征向量x分别经过 3 个 1×1 的卷积$\theta(x)$、$\phi(x)$和$g(x)$运算, 其中$\theta(x)$和$\phi(x)$用于计算两个节点间的相似度, $g(x)$则用于生成加权和. 注意力系数公式定义为:

$$\alpha_{ij}=\frac{e^{\sigma(\omega_f\cdot[\theta(x_i)\Vert\phi(x_i)])}}{\sum_{k=1}^{N}e^{\sigma(\omega_f\cdot[\theta(x_i)\Vert\phi(x_k)])}} \tag{6}$$

其中, θ和ϕ是核大小为 1 的卷积; $[\cdot\Vert\cdot]$表示级联; σ表示非线性激活函数 LeakyReLU, 负输入斜率$\alpha=0.2$.

C_k也是一个17×17的矩阵, 和式 (4) 中C_k一样是一个可训练权重矩阵, 其元素可为任意值, 完全由训练数据学习而得到.

2.3 损失函数的改进

每关节的平均位置误差 (*MPJPE*) 是一种用于评估三维人体姿态估计算法性能的常用损失函数之一, 它用于衡量预测三维关键点的位置坐标与真实三维关键点标签的位置坐标的平均距离, 其定义为:

$$MPJPE=\frac{1}{N}\sum_{i=1}^{N}\left\Vert J_i-J_i^*\right\Vert_2 \tag{7}$$

其中, N表示关节数量为 17, J_i是真实关节位置, J_i^*是预测的关节位置, $\Vert\cdot\Vert$表示欧几里得距离.

整个网络只考虑到了距离误差的损失, 为了确保生成的姿态在现实世界中是合理和自然的, 因此额外引入骨骼长度运动学约束, 确保生成的姿态中骨骼长度和和实际人体骨骼长度保持一致, 以保持人体的生物力学一致性. 将骨骼长度损失这个惩罚项添加到姿

态估计损失函数中, 以帮助网络更好地学习并生成更准确的三维人体姿态. 骨骼长度损失和改进后网络总的损失函数分别为:

$$loss\_bone = \sum_{i,j} \left| \left\| B^*_{i,j} \right\|_2 - \left\| B_{i,j} \right\|_2 \right| \tag{8}$$

$$Loss = 0.0025 \times loss\_bone + MPJPE \tag{9}$$

其中, (i, j)是身体部位的骨骼对应点, B^*表示真实三维关节点间的骨骼长度, B表示预测的三维关节点间的骨骼长度.

3 实验

3.1 数据集及训练测试

使用公开可用的数据集 Human3.6M[21]对本文的模型进行训练评估. Human3.6M 是一个包含约 360 万个视频帧数据集, 由 11 名受试者的 15 项日常活动 (如吃饭、散步、坐着等) 图片构成, 且有对应姿态的 2D 和 3D 的关键点真值. 同其他方法一样, 将 1、5、6、7、8 编号的受试者动作图片作为训练集, 9 和 11 编号受试者的动作图片作为测试集来对算法模型进行评估分析.

使用 PyTorch 框架来实现本文方法, 选用 RTX 2080Ti, PyTorch 1.11.0, Python 3.8, CUDA 11.3, 内存为 43 GB 的 GPU. 设置批次数为 128, 训练轮数为 60. 学习率从 0.001 开始, 每一轮的衰减因子为 0.95, 随机失活参数设置为 0.05.

在实验中使用两种常见的用于人体姿态估计的评估指标. 指标 1, 计算真值和预测的 3D 坐标之间的每个关节的平均位置误差 (*MPJPE*). 指标 2, 将预测关节与真值关节刚性对齐后, 再计算每个关节的平均位置误差 (*P-MPJPE*).

3.2 实验结果及分析

对于三维人体姿态估计, 选取通用的两个指标: *MPJPE* 和 *P-MPJPE*, 单位为 mm. 这两个指标越小, 表示模型预测的三维关节位置与真实位置越接近, 模型的性能越好. 这里对数据集中的 15 个动作分别计算平均关节位置误差 (*MPJPE*) 和刚性对齐后的平均关节位置误 (*P-MPJPE*), 以及对 15 个动作的指标结果求平均.

表 1 是用 CPN 作为 2D 检测器和使用真实的 2D 关键点, 以及其他方法得到 3D 关键点的指标 1 结果. 表 2 是用 CPN 作为 2D 检测器和使用真实的 2D 关键点, 以及其他方法得到 3D 关键点的指标 2 结果. 原始方法是由原本的图注意力块搭建而成, 改进 1 是在原始方法基础上加上改进的损失函数, 使预测的姿态中骨骼长度和和实际的保持一致, 以帮助网络更好地学习并生成更准确的三维人体姿态. 改进 2 是在原始方法基础上对网络进行改进, 获取到更多的特征得到更准确的预测结果. 而本文方法是原始方法加上两个改进, GT 表示真实 2D 关键点. 最好的结果用粗体标出, 次好的结果用下划线标出.

表 1 指标 1 结果 (mm)

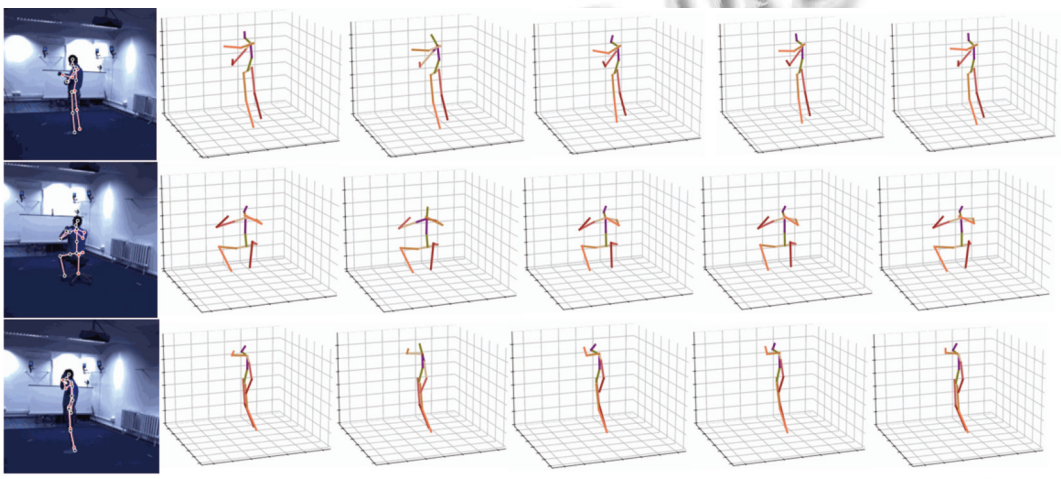
Protocol #1 Error	Directions	Discuss	Eating	Greet	Phone	Photo	Pose	Purchases	Sitting	SittingDown	Smoke	Wait	WalkDog	Walk	WalkPair	Avg.
Ci等[22]	46.8	52.3	44.7	50.4	52.9	68.9	49.6	46.4	60.2	78.9	51.2	50.0	54.8	40.4	43.3	52.7
Liu等[23]	46.3	52.2	47.3	50.7	55.5	67.1	49.2	46.0	60.4	71.1	51.5	50.1	54.5	40.3	43.7	52.4
Xu等[24]	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
原始方法	47.4	51.1	47.9	51.0	53.1	60.0	48.5	48.2	58.6	65.7	51.5	48.0	55.2	40.0	42.3	51.2
改进1	47.2	50.0	47.8	51.0	53.9	60.2	48.4	48.0	59.4	65.3	51.9	48.0	54.6	39.3	41.7	51.1
改进2	46.1	**49.5**	47.1	50.4	53.1	**59.2**	48.5	47.1	**57.9**	**64.7**	50.7	**47.1**	54.4	40.2	42.6	50.6
本文方法	**44.8**	49.6	**46.0**	**49.6**	**51.7**	59.7	**46.6**	**46.7**	58.5	69.9	**50.5**	47.6	**53.8**	**38.1**	**41.0**	**50.3**
本文方法 (GT)	35.4	38.9	29.9	34.5	35.0	38.8	39.4	33.3	36.8	41.8	35.5	35.4	33.9	28.3	31.1	35.2

表 2 指标 2 结果 (mm)

Protocol #2 Error	Directions	Discuss	Eating	Greeting	Phone	Photo	Pose	Purchases	Sitting	SittingDown	Smoke	Wait	WalkDog	Walk	WalkPair	Avg.
Ci等[22]	36.9	41.6	38.0	41.0	41.9	51.1	38.2	37.6	49.1	62.1	43.1	39.9	43.5	32.2	37.0	42.2
Liu等[23]	35.9	40.0	38.0	41.5	42.5	51.4	37.8	36.0	48.6	56.6	41.8	38.3	42.7	31.7	36.2	41.2
原始方法	36.9	39.3	37.0	41.7	40.7	45.8	37.2	35.7	46.0	53.3	40.7	36.7	42.6	31.1	35.4	40.0
改进1	36.0	39.0	37.7	41.1	40.9	45.7	37.0	36.3	47.1	**52.2**	40.9	36.1	42.7	30.7	34.6	39.9
改进2	36.1	38.9	37.1	41.2	40.6	44.7	37.1	35.7	46.9	53.5	40.6	**36.0**	42.6	31.1	34.6	39.8
本文方法	**35.1**	**38.5**	**36.1**	**40.0**	**39.7**	**45.4**	**35.9**	**35.4**	**46.5**	57.9	**40.0**	**36.0**	**42.4**	**30.1**	**33.8**	**39.5**
本文方法 (GT)	24.5	29.2	24.5	26.0	26.6	29.9	29.6	24.7	29.3	32.4	27.9	26.9	26.3	22.7	25.2	27.1

从表 1、表 2 可以看出, 使用 CPN 作为 2D 检测器, 本文方法相较于原来方法, 大部分动作的指标 1 和指标 2 结果都有了提升, 对于 15 个动作的指标 1 和指标 2 的平均值, 分别提升了 1.8% 和 1.3%, 同时与其他几种经典方法比较, 在大部分动作上都有较好的提升, 获得了更为准确的 3D 关键点. 使用真实 2D 值作为输入也分别取得 35.2 和 27.1 的指标 1 和指标 2 的结果. 结果表明, 使用 Human3.6M 数据集, 改进后的本文方法明显取得更好的指标效果, 且优于其他几种方法.

图 4 是本文方法可视化的三维人体姿态, 采用 CPN 得到的 2D 姿态作为网络输入得到指标 1 平均位置误差. 图 4(a) 是 2D 姿态, 图 4(b) 方法 1 是指文献[23]方法得到的 3D 人体姿态, 图 4(c) 方法 2 是指文献[24]方法得到的 3D 人体姿态, 图 4(d) 是指原始方法得到的 3D 人体姿态, 图 4(e) 是指改进原始方法得到的本文方法的 3D 人体姿态, 图 4(f) 表示真实的 3D 姿态. 选取了 3 种不同动作进行展示, 第 1 行是 Directions 动作, 第 2 行是 Eating 动作, 第 3 行是 Greeting 动作. 由可视化的图片结果可以看出, 在肘部和膝盖等关节的位置, 本文方法预测的更准确, 改进后的本文方法预测出的整体姿态更好, 更接近真实的 3D 姿态.



(a) Input (b) 方法 1 (c) 方法 2 (d) 原始方法 (e) 本文方法 (f) GT

图 4 可视化效果

4 结论

本文基于时空图注意力卷积网络的图注意力块构建人体三维姿态估计网络, 改进了全局多头图注意力网络, 重新设计了损失函数, 通过对局部和全局的空间节点信息建模, 实现局部运动学连接、对称性和全局姿态的人体骨骼运动学约束的学习, 有效地缓解了深度模糊和自遮挡, 提高了人体姿态估计性能, 获得了较为精确的三维人体关键点. 同时, 与同样的对单张图片进行三维人体姿态估计的一些算法相比, 使用 2D 姿态检测器 CPN 或者使用真实 2D 标签作为输入, 本文改进的算法在评价指标上均要更好一些.

参考文献

1 王仕宸, 黄凯, 陈志刚, 等. 深度学习的三维人体姿态估计综述. 计算机科学与探索, 2023, 17(1): 74–87.

2 Erol A, Bebis G, Nicolescu M, *et al*. Vision-based hand pose estimation: A review. Computer Vision and Image Understanding, 2007, 108(1–2): 52–73. [doi: 10.1016/j.cviu.2006.10.012]

3 Zhang HT, Sciutto C, Agrawala M, *et al*. Vid2Player: Controllable video sprites that behave and appear like professional tennis players. ACM Transactions on Graphics, 2021, 40(3): 24.

4 朱凌飞, 万旺根. 基于骨架模型的人体行为分析. 电子测量技术, 2021, 42(8): 68–73. [doi: 10.19651/j.cnki.emt.1802315]

5 Chen WM, Jiang ZJ, Guo HL, *et al*. Fall detection based on key points of human-skeleton using OpenPose. Symmetry, 2020, 12(5): 744. [doi: 10.3390/sym12050744]

6 Chen YC, Tian YL, He MY. Monocular human pose estimation: A survey of deep learning-based methods. Computer Vision and Image Understanding, 2020, 192: 102897. [doi: 10.1016/j.cviu.2019.102897]

7 Martinez J, Hossain R, Romero J, *et al*. A simple yet effective baseline for 3D human pose estimation.

Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2659–2668.
8 Fang HS, Xu YL, Wang WG, *et al*. Learning pose grammar to encode human body configuration for 3D pose estimation. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 6821–6828.
9 Hossain MRI, Little JJ. Exploiting temporal information for 3D human pose estimation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 69–86.
10 Pavllo D, Feichtenhofer C, Grangier D, *et al*. 3D human pose estimation in video with temporal convolutions and semi-supervised training. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7745–7754.
11 Zhao L, Peng X, Tian Y, *et al*. Semantic graph convolutional networks for 3D human pose regression. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3420–3430.
12 马双双, 王佳, 曹少中, 等. 基于深度学习的二维人体姿态估计算法综述. 计算机系统应用, 2022, 31(10): 36–43. [doi: 10.15888/j.cnki.csa.008711]
13 Newell A, Yang KY, Deng J. Stacked hourglass networks for human pose estimation. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 483–499.
14 Chen YL, Wang ZC, Peng YX, *et al*. Cascaded pyramid network for multi-person pose estimation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7103–7112.
15 Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5686–5796.
16 Kocabas M, Karagoz S, Akbas E. Self-supervised learning of 3D human pose using multi-view geometry. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1077–1086.
17 Wandt B, Rosenhahn B. RepNet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7774–7783.
18 Liu JF, Rojas J, Li YH, *et al*. A graph attention spatio-temporal convolutional network for 3D human pose estimation in video. Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an: IEEE, 2021. 3374–3380.
19 Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR, 2017.
20 Shi L, Zhang YF, Cheng J, *et al*. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12018–12027.
21 Ionescu C, Papava D, Olaru V, *et al*. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1325–1339.
22 Ci H, Wang CY, Ma XX, *et al*. Optimizing network structure for 3D human pose estimation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 2262–2271.
23 Liu KK, Ding RQ, Zou ZM, *et al*. A comprehensive study of weight sharing in graph networks for 3D human pose estimation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 318–334.
24 Xu TH, Takano W. Graph stacked hourglass networks for 3D human pose estimation. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 16100–16109.

(校对责编: 牛欣悦)