基于序列多尺度特征融合表示的层级舞蹈动作姿态估计方法

杨红红[1],王刘丽[1],张玉梅[1,2],吴晓军[1,2],党允彤[3]
(1. 陕西师范大学现代教学技术教育部重点实验室,陕西西安710062; 2. 陕西师范大学计算机科学学院,陕西西安710062;
3. 陕西师范大学音乐学院,陕西西安710062)

摘 要: 人体姿态估计是计算机视觉研究领域的热点研究问题之一,但其在传统民间舞蹈动作姿态估计方面的应用研究尚处于起步阶段.由于舞蹈图像中人体动作复杂多变、舞蹈动作连贯性强、舞蹈者存在严重遮挡不易检测等特点,传统人体姿态估计方法难以准确估计舞蹈者的动作变化,导致舞蹈动作姿态估计准确率较低.针对此问题,本文提出一种基于序列多尺度特征融合表示的层级舞蹈动作姿态估计方法,该方法针对舞蹈动作骨骼关节点尺度变化剧烈的问题,构建基于序列多尺度特征融合表示的关节点估计模型.并且,针对舞蹈姿态形变较大,遮挡严重的问题,设计基于关节点几何关系的层级姿态估计模型,提高舞蹈动作姿态估计的效果.实验结果表明,本文方法在标准人体姿态估计数据集及自建舞蹈数据集上取得较好的姿态估计结果.

关键词: 舞蹈动作姿态估计; 序列多尺度特征融合; 关节点几何关系; 层级姿态估计
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2021)12-2428-09
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200637

Hierarchical Dance Pose Estimation Algorithm Based on Sequential Multi-Scale Feature Fusion

YANG Hong-hong[1], WANG Liu-li[1], ZHANG Yu-mei[1,2], WU Xiao-jun[1,2], DANG Yun-tong[3]
(1. *Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, Xi'an, Shaanxi* 710062, *China*;
2. *School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi* 710062, *China*;
3. *School of Journalism and Communication, Shaanxi Normal University, Xi'an, Shaanxi* 710062, *China*)

Abstract: Human pose estimation is one of the hot research topics in the field of computer vision, but its application in traditional dance pose estimation is still in its infancy. Due to the complexity of dance pose, the strong coherence of dance movements, and difficulty in detecting of dancers' poses caused by serious occlusion in dance images, the traditional human pose estimation methods are difficult to accurately estimate the pose changes of dancers, thus resulting in low accuracy in estimating dance pose. We propose a hierarchical dance pose estimation method based on sequential multi-scale feature fusion. To address the problems of the drastic scale changes of the dancer pose, a keypoint estimation model based on sequential multi-scale feature fusion is constructed. Furthermore, aiming to solve the issues that the large deformation and serious occlusion of dance pose, a hierarchical pose estimation model based on the geometric relationship between human keypoints is designed to improve the accuracy of dance pose estimation. The experimental results show that the proposed method can achieve good pose estimation results on the standard human pose estimation dataset and the self-collected dance dataset.

Key words: dance pose estimation; sequential multi-scale feature fusion; geometry relationship among keypoints; hierarchical pose estimation

收稿日期:2020-06-29;修回日期:2021-07-27;责任编辑:梅志强
基金项目:国家自然科学基金(No.61907028, No.11872036,);陕西省青年科技新星项目(No.2021KJXX-91);陕西省高校科学人才托举项目(No.20200105);中央高校(No. GK202103114, No.GK202101004);陕西省自然科学基金(No.2019JQ-574)

1 引言

舞蹈是文化的重要表现形式之一,我国舞蹈课堂人数通常较多,教师只能粗略地通过学生的肢体动作及面部表情获取学生的动作变化及情感变化,难以精确地了解到学生对舞蹈动作实时掌握的情况. 因此,应用信息技术实时对舞者的动作姿态进行估计,及时获得课堂舞蹈教学状态信息,将极大促进因材施教的实施[1].

随着科技与文化深度融合的开展,对舞蹈图像中的动作姿态进行估计将成为计算机视觉技术的一个重要应用领域,其不仅可以用于专业舞蹈者动作纠正,舞蹈自助教学等应用场景,还可以用于运动员运动分析、比赛仲裁、动作识别、影视娱乐、辅助游戏设计、增强现实(Augmented Reality, AR)、虚拟现实(Virtual Reality, VR)等多个人机交互现实场景[2-4].

目前,多人姿态估计方法可以分为自顶向下(Top-down)和自底向上(Bottom-up)两类,前者主要是先通过目标检测器检测出图像中的人体检测框,然后对每一个人体检测框进行单人姿态估计产生人体关节点,最后对关节点进行连接形成人体姿态估计结果. Chen 等人提出 CPN(Cascaded Pyramid Network)方法[5],通过特征金字塔和 RefineNet 网络实现关节点的估计. Simple-Baseline[6]是一种用于多人姿态估计和跟踪的简单有效网络. Fang 等人提出 RMPE(Regional Multi-person Pose Estimaion)方法[7]实现单人姿态估计. Newell 等人提出 Hourgalss[8]网络来结合不同尺度的特征. Sun 等人提出 HRNet[9]网络利用高分辨率特征进行人体姿态估计. 自底向上的方法主要分为关节点检测和关节点聚类两部分,其利用单人姿态估计算法将图像中所有的关节点检测出来,然后对不同人体的关节点聚类,将属于同一个人体的关节点聚合到一起实现多人姿态估计. 该类人体姿态估计算法的典型代表主要有 Cao 等人提出的 OpenPose[10], Insafutdinov 等人提出 Deepcut[11]网络, HigherHRNet[12]和 Pifpaf[13]等.

上述两类多人姿态估计方法各有优缺点, Top-down 方法将人体姿态估计分为人体目标检测和单人姿态估计两步. 由于其依赖于性能较好的目标检测算法及单人姿态估计算法,人体姿态估计的准确率较高. 但是,该类方法性能受目标检测框质量影响严重,即使最为先进的目标检测器也会存在检测误差,造成人体检测框冗余、漏检和误检等现象[14]. 而 Bottom-up 方法不依赖于目标检测器进行人体框的检测,因此其检测速度较快,但是对不同关节点进行聚合时受遮挡影响严重,当多人距离较近时,很容易造成同一人体关节点聚类歧义问题,因此其人体姿态估计准确率较低.

此外,现有的人体姿态估计方法主要针对传统的数据集,如 MSCOCO, MPII、LSP 等,其包含简单的人体姿态,如站立,走路等. 但是,舞蹈动作姿态估计中存在舞蹈动作复杂多变,连贯性强,严重遮挡,舞蹈课堂场景中多存在光照变化及相机视角变化等干扰因素,极大地增加了舞蹈动作姿态估计的难度. 因此,传统人体姿态估计方法存在难以准确估计舞蹈者动作变化的问题.

针对上述问题,本文提出一种基于序列多尺度特征融合的层级舞蹈动作姿态估计方法,其算法流程图如图 1 所示. 该算法采用 Top-down 框架,首先利用 YOLOv3 进行舞蹈者人体框的检测;然后,针对舞蹈动作骨骼关节点尺度变化剧烈的问题,以 HRNet 网络为骨干网络,提出一种序列多尺度特征融合方法,通过对高、低层多尺度特征进行融合,提高姿态估计对尺度变化的鲁棒性. 接着,针对舞蹈姿态存在较大的形变及遮挡严重的问题,通过分析人体骨骼关节点之间的几何关系,设计基于关节点几何关系的层级姿态估计模型,进行多层次的关节点估计,提高舞蹈者关节点位置的

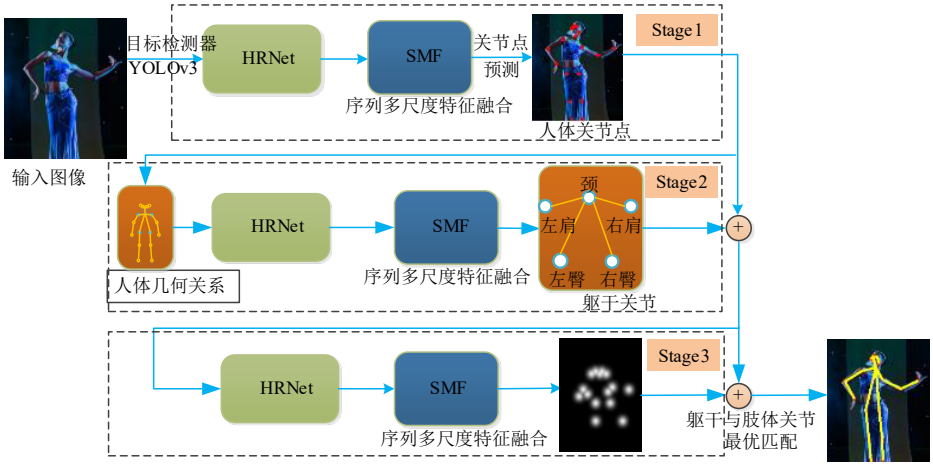

图 1　本文所提出算法的流程图

准确估计. 最后,在公共数据集及自建舞蹈数据集上验证所提算法的有效性.

2 基于序列多尺度特征融合表示的层级舞蹈动作姿态估计方法

2.1 基于YOLOv3的人体框检测

本文在人体目标检测阶段采用端到端的算法,基于YOLOv3检测器[15]进行舞蹈者人体检测框(human proposal)的提取. 将RGB图像输入YOLOv3模型,获得相应的人体检测框用于人体姿态估计.

2.2 多尺度特征融合表示

由于姿态估计任务是像素(pixel-wise)级关节点估计问题,其需要利用低层和高层特征对不同尺度大小的关节点进行定位,高层特征有利于大尺度关节点的定位,而低层特征对小尺度关节点的定位非常重要. 针对舞蹈动作骨骼关节尺度变化剧烈的问题,本文构建一种序列多尺度特征融合模型,提高姿态估计对尺度变化的鲁棒性.

2.2.1 HRNet网络

本文以HRNet网络为骨干网络[9],如图2所示,其由4个并行的多分辨率子网构成,每个子网络采用ResNet模块设计原则,由4个残差单元组成.

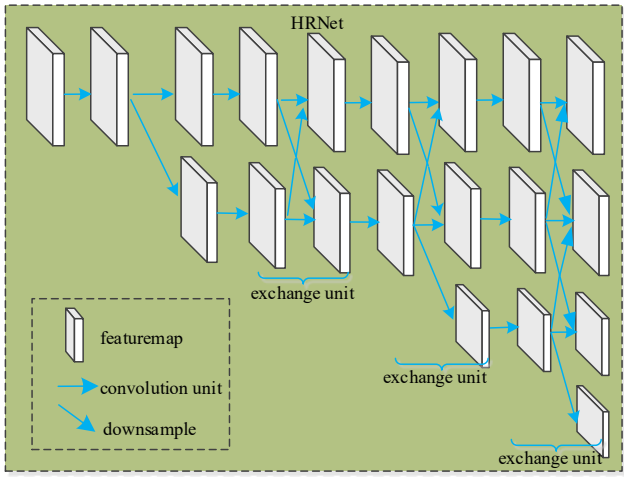

图2 HRNet骨干网络

HRNet网络由于能够较好提取输入图像的多分辨率特征,其具有较强的特征表示能力,在目标检测、识别、图像分割以及人体关节点估计任务中获得较好的结果,但是HRNet网络在人体姿态关节点估计过程中并没有充分利用其提取的多分辨率特征,仅使用其中的高分辨率特征进行关节点热图(heatmap)估计,丢弃其他中、低分辨率特征,从而造成特征表示过程中的信息损失,影响关节点估计的准确性. 因此,针对上述问题,本文提出构建序列多尺度特征融合模型,提高姿态估计特征表示的能力.

2.2.2 序列多尺度特征融合

在特征表示中,低分辨率的高层特征具有丰富的语义信息而位置信息相对粗糙,而高分辨率的低层特征虽然语义信息相对较弱但包含准确的位置信息. 因此,本文提出序列多尺度特征融合方法(Sequential Multi-scale Feature fusion,SMF),对高、低分辨率特征进行有序融合,增强网络特征表示的能力. 如图3所示,该序列多尺度特征融合方法对HRNet网络[9]最后一个聚合单元输出的4个不同分辨率特征图经过卷积(convolution)、插值(interpolation)和反卷积(deconvolution)操作进行由高分辨率到低分辨率的序列多特征融合.

文中以HRNet网络最后一个聚合单元输出的4个特征图作为序列多尺度特征融合模块的输入特征 $\tilde{\boldsymbol{X}}=(\tilde{\boldsymbol{X}}_1,\tilde{\boldsymbol{X}}_2,\cdots,\tilde{\boldsymbol{X}}_m)$,其中,$m$ 表示输入特征对应的分辨率(本文 $m=4$). 文中以HRNet-W32骨干网络作为骨干网络,其4个平行子网络涉及的多分辨率特征的通道数分别为32,6,4,128和256. 对于任意的第 ith 分辨率的特征,首先进行conv(3×3)卷积操作,然后进行插值和反卷积操作使 ith 分辨率的特征 $\tilde{\boldsymbol{X}}_i$ 上采样变成修正后的 $(i-1)$ th分辨率特征 $\hat{\boldsymbol{X}}_{i-1}$:

$$\hat{\boldsymbol{X}}_{i-1}=\mathrm{Int}(\mathrm{conv}(\tilde{\boldsymbol{X}}_i))+\mathrm{Dec}(\mathrm{conv}(\tilde{\boldsymbol{X}}_i)) \tag{1}$$

其中,conv表示卷积操作,Int和Dec分别表示插值和反卷积操作.

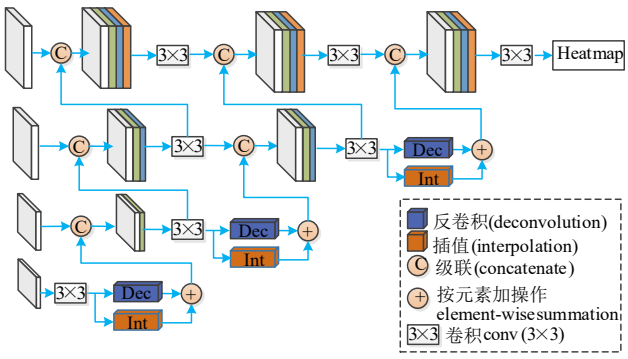

图3 序列多尺度特征融合模块(SMF)

接着,级联上采样获得的修正后的 $(i-1)$ th分辨率特征 $\hat{\boldsymbol{X}}_{i-1}$ 和第 $(i-1)$ th分辨率特征 $\tilde{\boldsymbol{X}}_{i-1}$,得到融合后的第 $(i-1)$ th分辨率特征 $\boldsymbol{X}'_{i-1}$:

$$\boldsymbol{X}'_{i-1}=\mathrm{concat}(\hat{\boldsymbol{X}}_{i-1},\tilde{\boldsymbol{X}}_{i-1}) \tag{2}$$

其中,concat表示级联特征 $\hat{\boldsymbol{X}}_{i-1}$ 和 $\tilde{\boldsymbol{X}}_{i-1}$.

经过反复执行式(1)和式(2)实现高、低分辨率特征的序列融合,如式(2)所示,最终获得融合多分辨率多尺度信息的特征 $\boldsymbol{X}'_1$.

最后,在最终的特征 $\boldsymbol{X}'_1$ 上使用Softmax函数获得关节点热图,由热图估算获取各关节点的位置信息.

2.3 基于关节点几何关系的层级姿态估计

由于舞蹈姿态存在大的形变及严重遮挡的问题,本文利用2.2节所估计得到的人体骨骼关节点进行关节点几何关系关联性预测,通过分析关节点之间的几何关系,构建基于关节点几何关系的层级姿态估计模型(Hierarchical Pose Estimation, HPE),进行多层次的关节点估计,提高舞蹈者身体关节点位置的准确估计.

首先,根据人体结构将2.2节所获得的关节点划分为两类:第一类是形变较小的连接人体各关节的躯干关节(k^{trunk}),如肩,臀、颈部;第二类是形变明显的肢体关节(k^{limb}),如手腕、手肘、膝盖及脚踝等铰链关节.然后,根据所划分的两类关节,设计层级姿态估计模型,将人体所有关节点聚合为如图4所示的5部分,颈、左肩、右肩、左臂、右臂,进行基于人体关节几何关系的关节点预测.

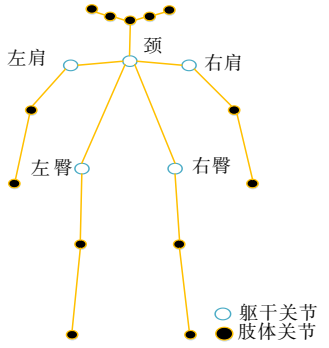

图4　人体关节点几何关系

如图1所示,本文所设计的层级网络由三个阶段组成.网络的第一阶段为根据2.2节所设计的SMF模型进行人体所有关节点的热图预测,并计算相应的坐标位置.然后将第一阶段所获得的关节点热图作为第二阶段网络的输入,鉴于人体躯干关节的形变较小及肢体关节的形变较大的特点,本文利用SMF模型从第一阶段所获得的所有人体关节中预测形变较稳定的躯干关节(k^{trunk}),将人体关节划分为以躯干关节为主的5部分,也称为5类(颈、左肩、右肩、左臂、右臂).接着,将网络第一阶段所有的关节点及第二阶段预测获得的5类躯干关节作为输入,构建第三阶段网络.同时,考虑人体结构的几何相关性,将人体所有的关节点进行类内关联划分到5类躯干关节中,实现肢体关节与躯干关节的连接.

由于每一类躯干关节,可以有多个候选肢体关节与其连接,同样,每一个肢体关节也可能与任意一类躯干关节相连接.因此,对于任意一类的躯干关节与肢体关节,设N_1、N_2分别为第c部分躯干关节k_1^{trunk}和肢体关节k_2^{limb}的候选关节点集合,则所有候选关节点类内连接集合的最优匹配问题为:

$$Z_c=\{z_{k_1,k_2}^{m,n}:k_1,k_2\in\{1,\cdots,K\},m\in\{1,\cdots,N_1\},\ n\in\{1,\cdots,N_2\},c\in\{1,\cdots,5\}\}\tag{3}$$

其中,$z_{k_1,k_2}^{m,n}\in\{0,1\}$表示关节点$k_1$和$k_2$是否连接.

对于相互连接的成对关节(k_1,k_2),根据图模型中两条边共享共同节点的方式将关节点之间的连接匹配问题转化为偶图匹配子问题[16].通过求解所有类内候选关节点连接集合的最优匹配问题,得到躯干关节与肢体关节之间连接的最优匹配,表示为:

$$z_{k_1,k_2}=\max E_{mn}z_{k_1,k_2}^{m,n}\tag{4}$$

$$\sum_{N_1}z_{k_1,k_2}^{m,n}\le 1\wedge\sum_{N_2}z_{k_1,k_2}^{m,n}\le 1\tag{5}$$

其中,E_{mn}由PAFs(Part Affinity Fields for Part Association)方法(文献[10]中式(11))计算关节点之间的关联概率.

最后,连接所有躯干关节与肢体关节的最优匹配组成人体的最终姿态估计结果.

2.4 损失函数

损失函数由3部分组成,分别为网络的第一阶段所估计的所有关节点与真值的差,第二阶段网络所估计的躯干关节与真值的差以及第三阶段网络所估计的肢体关节与真值的差.

$$l=\frac{1}{3m'n'}\sum_{i=1}^{n'}\sum_{j=1}^{m'}\left(\left\|H(k_{ij})-\text{GT}(k_{ij})\right\|^2+\left\|H(k_{ij}^{\text{trunk}})-\text{GT}(k_{ij}^{\text{trunk}})\right\|^2+\left\|H(k_{ij}^{\text{limb}})-\text{GT}(k_{ij}^{\text{limb}})\right\|^2\right)\tag{6}$$

其中,n'为训练样本数,m'为关节点个数.$H(k_{ij})$为预测的第i个样本第j个关键点的热图,其对应的真值为$\text{GT}(k_{ij})$.

3 实验结果与分析

3.1 实验数据、对比算法及评价指标

为了验证本文所提出算法的有效性,文中设计的实验包含公共数据集MSCOCO2017[17]及舞蹈数据集(包含自摄舞蹈数据及网络下载舞蹈数据)上的单人及多人舞蹈动作姿态估计对比分析.MSCOCO2017数据集包含57k训练数据COCO train2017,5k的验证数据COCO val 2017以及20k的测试数据COCO test-dev2017.舞蹈数据集包含20k舞蹈数据,其中,10k用于训练,10k用于测试.本文将所提出的人体姿态估计算法与目前主流的人体姿态估计算法进行对比分析,其中包含5种Top-down算法(CPN[5], SimpleBaseline[6], Mask-RCNN[18], RMPE[7],HRNet[11])和4种Bottom-up算法(Openpose[10],Pifpaf[13], Hourglass[8], HigherHRnet[12]),采用OKS(Object Keypoint Similarity)作为定量评价指标.

3.2 实验设置

本文实验在 Ubuntu 16.04，4 个 NVIDIA 1080Ti GPU 组成的服务器上运行，选用 python 语言及 Pytorch 深度网络框架. 以 HRNet-W32 网络为骨干网络，输入图片大小为 256×192. 对输入图像进行图像增强操作，包括随机翻转、±45°的旋转以及 ±35% 的尺度缩放. 网络使用 Adam[19]优化方式，其初始学习率为 0.001，在 100 epoch 到 130 epoch 下降到 0.00001，网络总共训练 210 epoch.

3.3 实验结果分析

3.3.1 消融实验分析

为了更好地对本文设计的网络架构进行分析，揭示网络组成各部分的性能，本文将构成算法的每一部分分别从整体算法中剥离出来，通过与未使用各个模块的算法进行对比实验来分析各模块对姿态估计的影响. 消融实验在 MSCOCO 验证集上进行，其度量指标结果如表 1 所示，使用 HRNet-W32 作为主干模型，网络输入图片大小为 256×192.

表 1 分别揭示了本文网络两个重要组成部分对姿态估计的影响，分别为(1)序列多尺度融合模块(SMF)，(2)基于关节点几何关系的层级优化模型(HPE)对网络的影响.

如表 1 所示，当所提出模型的两个重要组成模块同时使用时，如表 1 模型 D 所示，所提出算法获得最好的 mAP 值，其在热力图回归 HRNet-W32 模型的基础上带来了 1.83%（76.29~74.46）的 mAP 关节点定位精度提升. 在消融实验中，如表 1 模型 B 和模型 C 所示，在 HRNet-W32 模型的基础上分别仅添加 SMF 模型和 HPE 模型，其分别获得 0.75%（75.21~74.46）和 1.3%（75.76~74.46）的 mAP 提升. 从表 1 可以看出，模型 B 中 SMF 模型由于将 HRNet 网络的多分辨率多尺度特征进行有序融合，提高了多尺度特征表示的能力，有利于关节点的估计. 模型 C 中 HPE 模型依据关节点几何关系，设计三阶段层级网络模型进行关节点估计，灵活处理不同类的躯干关节和肢体关节，通过求解所有类内候选关节点连接集合的最优匹配问题，得到躯干关节与肢体关节之间连接的最优匹配，对遮挡关节点进行推理，从而提高姿态估计的准确性.

表 1 消融实验结果分析

模型	HRNet-W32	SMF	HPE	AP
A	√			74.46
B	√	√		75.21
C	√		√	75.76
D	√	√	√	76.29

3.3.2 实验 1 标准数据集实验结果

本文首先在 COCO test-dev 2017 数据集上，将所提出的算法与 9 种主流的人体姿态估计算法进行多人姿态对比分析. 为了与其他对比算法公平比较，本文在 COCO 数据集上采用文献[6]中使用的目标检测器进行人体框的提取，获得人体检测框之后再根据本文所提出的算法进行人体姿态估计，实验结果如表 2 所示，本文算法在 COCO test-dev 2017 数据集上姿态估计结果的部分可视化图如图 5 所示.

表 2 本文算法及对比算法在标准数据集上的人体姿态估计结果.

对比算法	Bottom-up methods				Top-down methods					
	Openpose[10]	Hourglass[8]	Pifpaf[13]	HigherHRNet-W48[12]	Mask-RCNN[18]	CPN[5]	RMPE[7]	SimpleBaseline[6]	HRNet-W32[9]	本文算法
Backbone	–	Hourglass	–	HRNet-w48	ResNet-50-FPN	ResNet-Inception	PyraNet	ResNet-152	HRNet-W32	HRNet-W32
输入大小	–	512	–	640	–	384×288	320×256	384×288	256×192	256×192
AP	61.8	65.5	66.7	70.5	63.1	72.1	72.3	73.7	74.4	75.7
AP.5	84.9	86.8	–	89.3	87.3	91.4	89.2	91.9	90.5	92.8
AP.75	67.5	72.3	–	77.2	68.7	80	79.1	81.1	81.9	82.8
AP(M)	57.1	60.6	62.4	66.6	57.8	68.7	68	70.3	70.8	71.9
AP(L)	68.2	72.6	72.9	75.8	71.4	77.2	78.6	80	81.0	81.4
AR	66.5	70.2	–	74.9	–	78.5	–	79	79.8	79.7

通过表 2 得出，本文算法在输入图像大小为 256×192，HRNet-W32 为骨干网络的框架上，获得 75.7AP，而 HRNet W32 方法获得 74.4 AP，SimpleBaseline 获得 73.7AP、Mask-RCNN 获得 63.1 的 AP . 其中，AP 到 AP(L)为各种对比方法在数据集上使用相应的目标检测器所获得的多人姿态估计的平均准确率. 如表 2 所示，本文所提出的方法相比 4 种 Bottom-up 方法而言，获得高于 Openpose、Hourglass、Pifpaf 以及 HigherHRNet-W48 分别 13.9、10.2、9、5.2 个 point 的 AP 值；与 5 种 Top-down 方法对比，获得高于 Mask-RCNN、CPN、RM-

PE、SimpleBaseline 以及 HRNet-W32 分别 12.6 、3.6、3.4、2.0、1.3 个 point 的 AP 值. 综上所述,本文算法在多数评价指标上都取得较好的多人姿态估计结果,这是由于本文方法针对人体尺度变化的问题,构建序列多尺度特征融合模型,提高算法对尺度变化的鲁棒性,同时,针对遮挡及姿态形变问题,设计基于关节点几何关系的层级姿态估计模型,提高算法姿态估计的效果,因此,如图5所示,本文算法在遮挡,尺度变化较大、复杂背景等场景中能较好实现姿态估计,获得较好的姿态估计结果.

图5　本文算法在标准数据集上的部分人体姿态估计可视化图

3.3.3　实验2 自建舞蹈数据集实验结果

为了验证本文算法的普适性,在自建舞蹈数据集上对所提出的算法进行人体姿态估计. 由于自建数据集缺少真值,采用自己标定的方法与其他对比方法进行定量对比分析缺乏公平性,因此,在自建舞蹈数据集上本文仅使用定性分析,部分实验结果如图6、图7所示. 在此,本文选取5类特色鲜明的单人、多人民族舞蹈数据集作为可视化结果分析,分别为藏族舞蹈、傣族舞蹈、汉族秧歌、蒙古族舞蹈、维吾尔族舞蹈.

3.3.3.1　自建单人舞蹈数据集实验结果

图6为本文算法在5类单人民族舞蹈数据集上的部分可视化结果,如图6所示,在藏族舞蹈中,由于舞者所穿黑色长裙的遮挡,即使人眼也很难准确定位出腿部关节的位置,此外,舞者白色水袖对胳膊关节及腿部关节的遮挡,加大了姿态估计的难度. 本文方法在此情况下,根据人体关节的几何关系,通过层级姿态估计模型进行偶图匹配子问题求解,进行遮挡关节点的预测,从而获得较为准确的关节点估计结果. 在傣族舞蹈中,舞者姿态存在剧烈的形变、舞者身体自遮挡及服饰遮挡现象严重,加大了关节点估计的难度;在汉族秧歌舞蹈中,如图6所示,道具扇子对舞者姿态存在严重的遮挡,同时,舞者存在快速的运动导致运动模糊. 在蒙古族舞蹈中,舞者姿态存在严重的自遮挡及服饰遮挡,即使人眼也很难准确定位出长裙遮挡腿部关节的位置. 在维吾尔族舞蹈中,由于灯光的影响以及舞者服饰和快速的运动,增加了舞者关节点估计的难度. 综上所述,本文方法针对舞蹈动作复杂多变,姿态形变剧烈等问题构建序列多尺度特征融合模型,提高姿态估计特征表示的能力,同时,对人体关节点几何关系进行分析,设计层级姿态估计模型,对遮挡关节点进行推理. 如图6所示,所提出的算法在舞者存在遮挡、剧烈形变、灯光干扰及快速运动等情况下均能较好实现舞者姿态的估计,验证了本文所提出算法的有效性.

3.3.3.2　自建多人舞蹈数据集实验结果

图7为本文算法在5类多人民族舞蹈数据集上的部分可视化结果. 多人舞蹈动作姿态估计相对单人舞蹈姿态估计更具挑战性,其不仅包含单人姿态估计中舞者服饰变化、复杂背景、自遮挡及视角变化等问题,还需要处理人数未知,多人之间的互遮挡等问题. 在藏族舞蹈中,由于舞者所穿长裙的遮挡、摄像头视角变化以及舞者剧烈的动作变化,即使人眼也很难同时准确定位出多个舞者的身体关节点. 在傣族舞蹈中,由于舞台灯光昏暗、舞者动作复杂多变、舞者姿态变化剧烈、统一着装服饰的干扰及多人互遮挡、自遮挡的影响,加

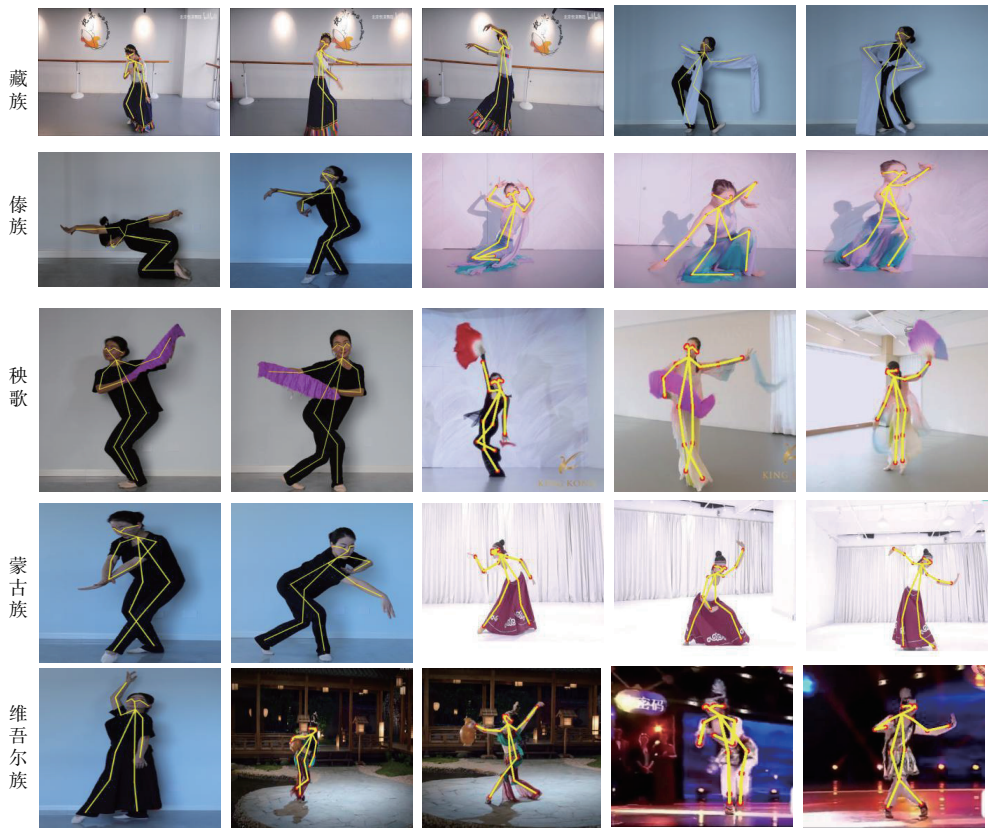

图6 本文算法在自建单人舞蹈数据集上的部分舞蹈姿态估计可视化图

图7 本文算法在自建多人舞蹈数据集上的部分舞蹈姿态估计可视化图

剧了舞者关节点估计的难度;在汉族秧歌舞蹈中,舞者存在较大的姿态变化、舞者动作复杂多变以及摄像头视角的变化,增加了姿态估计的难度;在蒙古族舞蹈中,舞者姿态尺度变化较大、存在较为严重的遮挡及服饰遮挡,增加了舞者关节点估计的难度. 在维吾尔族舞蹈中,由于舞者服饰严重的遮挡、舞蹈动作的复杂变化

以及快速的运动,即使人眼也很难准确定位出遮挡部位关节点的位置. 而本文方法在上述舞者动作复杂多变、姿态形变剧烈,遮挡严重、服饰及舞台灯光干扰等情况下,通过所构建的序列多尺度特征融合的层级舞蹈动作姿态估计模型,提高姿态估计特征表示的能力及关节点估计的准确性,从而较好地实现了舞者姿态的估计,验证了本文所提出算法的有效性.

综上所述,本文所提出的姿态估计算法虽然在标准人体姿态估计数据集和自建的单人、多人舞蹈数据集上取得较好的姿态估计效果. 但是本文算法仅针对舞蹈动作的普遍现象,如舞蹈者动作复杂多变、舞蹈动作连贯性强、舞蹈者存在严重遮挡不易检测的问题进行研究的,未能充分考虑不同民族舞蹈动作体态的多样性特点,这将是本文下一步需要深入研究的问题.

4　结论

本文提出一种基于序列多尺度特征融合的层级舞蹈动作姿态估计方法,该方法针对舞蹈动作复杂多变,姿态尺度变化较大等问题,构建序列多尺度特征融合模型,提高姿态估计算法对舞蹈动作骨骼关节点剧烈尺度变化的鲁棒性,同时,针对舞蹈姿态形变较大,遮挡严重的问题,设计基于关节点几何关系的层级姿态估计模型,提高舞蹈动作姿态估计的效果. 实验结果表明,该姿态估计算法在标准人体姿态估计数据集和自建的单人、多人舞蹈数据集上取得较好的姿态估计效果. 后续可以结合实际应用场景任务需求,将其应用于舞蹈教学中,实现舞蹈动作的实时纠正,辅助舞蹈者的教学训练,同时对于传承中华文化具有重要意义.

参考文献

[1] 杨丹妮. 传统文化传承视角下的民族民间舞蹈发展问题探讨[J]. 北方音乐, 2019, 39(13): 241 – 242.

[2] 彭学艳. 多媒体技术在高校舞蹈教学改革中的地位与作用[J]. 戏剧之家, 2018, 277(13): 167 – 168.

[3] 罗会兰, 童康,孔繁胜. 基于深度学习的视频中人体动作识别进展综述[J]. 电子学报, 2019, 47(5): 1162 – 1173.
Luo H L, Tong K, Kong F S. The progress of human action recognition in videos based on Deep learning: A review [J]. Acta Electronica Sinica, 2019, 47(5): 1162 – 1173. (in Chinese)

[4] 李康, 李亚敏, 胡学敏,等. 基于卷积神经网络的鲁棒高精度目标跟踪算法[J]. 电子学报, 2018, 46(9): 2087 – 2093.
Li K, Li Y M, Hu X M, et al. A robust and accurate object tracking algorithm based on convolutional neural network [J]. Acta Electronica Sinica, 2018, 46(9): 2087 – 2093. (in Chinese)

[5] Chen Y L, Wang Z C, Peng Y X, et al. Cascaded pyramid network for multi-person pose estimation[A]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition [C]. Salt Lake City, UT, USA: IEEE. 7103 – 7112

[6] Xiao B, Wu H P, Wei Y C. Simple baselines for human pose estimation and tracking[A]. Computer Vision—ECCV 2018[C]. Munich, Germany: Springer. 472 – 487.

[7] Fang H S, Xie S Q, Tai Y W, et al. RMPE: regional multi-person pose estimation[A].2017 IEEE International Conference on Computer Vision (ICCV)[C]. Venice, Italy: IEEE, 2017. 2353 – 2362

[8] Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation[A]. Computer Vision—ECCV 2016[C]. Amsterdam, Netherlands: Springer. 483 – 499.

[9] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[A]. 2019 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Long Beach, USA: IEEE, 2019. 5693 – 5703.

[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, 2021, 43(1): 172 – 186.

[11] Insafutdinov E, Pishchulin L, Andres B, et al. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model[A]. Computer Vision—ECCV 2016[C]. Amsterdam, Netherlands: Springer. 34 – 50.

[12] Cheng B W, Xiao B, Wang J D, et al. Bottom-up higher-resolution networks for multi-person pose estimation [A].2020 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Seattle, USA: IEEE, 2020. 1 – 10.

[13] Kreiss S, Bertoni L, Alahi A. PifPaf: Composite fields for human pose estimation[A]. 2019 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) [C]. Long Beach, USA: IEEE, 2019. 11969 – 11978

[14] 罗会兰, 陈鸿坤. 基于深度学习的目标检测研究综述[J]. 电子学报, 2020, 48(6): 1230 – 1239.
Luo H L, Chen H K. Survey of object detection based on deep learning[J]. Acta Electronica Sinica, 2020, 48, (6): 1230 – 1239. (in Chinese)

[15] Redmon J, Farhadi A. YOLOv3: An incremental improvement [A].2018 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) [C]. Salt Lake City, USA: IEEE, 2018. 1 – 6.

[16] West D B. Introduction To Graph Theory(Second Edition) [EB/OL]. http://bayanbox. ir/download/13403877125 4298 574/West-2nd-Edition-Solution-Manual.pdf, 2001.
[17] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[A]. Computer Vision—ECCV 2014[C]. Zurich, Switzerlan:Springer. 740 – 755.
[18] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN [A].2017 IEEE International Conference on Computer Vision (ICCV)[C].Venice, Italy: IEEE, 2017. 2980 – 2988.
[19] Kingma Diederik P, Ba Jimmy. Adam: A Method for Stochastic Optimization [EB/OL]. https://arxiv. org/pdf/1412. 6980v8.pdf, 2014.

作者简介

杨红红 女,1988年6月生,甘肃陇西县人.现为陕西师范大学现代教学技术教育部重点实验室副研究员,主要从事人工智能,深度学习与计算机视觉等领域的研究.
E-mail:yanghonghong0615@163.com

王刘丽 女,1997年1月生,山西晋城人.现为陕西师范大学现代教学技术教育部重点实验室硕士研究生,主要研究方向是知识工程与智能教学系统.
E-mail:1136946628@qq.com

张玉梅(通信作者) 女,1977年10月生,陕西榆林人.现为陕西师范大学计算机学院教授,主要从事信号处理与分析相关领域研究工作.
E-mail:zym0910@snnu.edu.cn

吴晓军 男,1970年12月生,陕西凤翔人.现为陕西师范大学计算机学院教授,主要从事模式识别,智能系统与复杂系统相关研究工作.
E-mail:xjwu@snnu.edu.cn

党允彤 女,1984年3月生,陕西西安人.现为陕西师范大学音乐学院副教授,新闻与传播学院在读博士生,主要从事数字技术与文化传播,科技艺术融合等领域的研究.
E-mail:dyt2011@snnu.edu.cn