中国计算机科学评论
Chinese Computer Sciences Review
2023 年第 01 卷第 02 期
Vol. 01, No. 02, 2023

基于单目视觉的人体姿态估计方法综述

闫　鑫，高　浩*，李昊伦

（南京邮电大学，南京 210023）

摘要：本文是一篇关于单目人体姿态估计技术方法和行业应用的综述，主要介绍近些年单目人体姿态估计的发展历程。随着计算机视觉和机器学习领域的快速发展，单目人体姿态估计已经成为一项备受关注的研究方向。本文首先介绍了单目人体姿态估计的相关概念和意义，并阐释了该领域研究的重要性。然后，我们详细介绍了单目人体姿态估计的技术方法，包括单目 2D 人体姿态估计以及单目 3D 人体姿态估计两种不同的技术路线。针对每种研究方法，我们讨论了其原理、发展及优缺点。接着，我们探索了单目人体姿态估计在人机交互，自动驾驶，医疗健康等领域的应用，并对其举例分析以突出人体姿态估计的重要性。最后，我们对当前的研究热点和挑战进行了分析，同时也展望了单目人体姿态估计在未来的发展方向。本文旨在为研究者和从业者提供一个全面的概述，促进单目人体姿态估计技术的应用和进一步研究。

关键词：单目人体姿态估计；深度学习；技术方法；行业应用

DOI：10.48014/ccsr.20230614001

引用格式：闫鑫，高浩，李昊伦．基于单目视觉的人体姿态估计方法综述[J]．中国计算机科学评论，2023，1(2)：13-30.

0　引言

单目人体姿态估计是计算机视觉的一项基本且富有挑战性的任务。它旨在从输入图像或者视频中预测人体的空间位置信息。随着深度学习的发展，单目人体姿态估计表现有着飞跃性的提升。其已经广泛应用于视觉任务，如人员重新识别，人体动作识别和人机交互等。基于单目的人体姿态估计深度学习方法与传统方法相比，不再需要复杂的穿戴式传感器或联合多摄像头进行拍摄。因此，在复杂场景下，单目人体姿态估计有着更广泛的应用，例如 3D 电影制作，虚拟现实以及自动驾驶等。

根据空间维度的不同，单目人体姿态估计可以划分为 2D 人体姿态估计与 3D 人体姿态估计。单目 2D 人体姿态估计旨在预测出图像中的关节点像素位置。不同于单目 2D 人体估计，单目 3D 人体姿态估计进一步预测了人体关节点的深度信息。在近些年的 3D 人体姿态估计中，常常将 2D 姿态作为 3D 姿态估计的中间表示，即借助 2D 姿态来恢复 3D 姿态。此外，由于 3D 姿态与 2D 姿态相比，提供了附加的深度信息，因此 3D 人体姿态估计能够实现更广泛的应用。

随着深度学习的不断发展以及大数据集的不断扩充，人体姿态估计继续取得巨大进展，特别是在 2D 图像上。然而，由于缺乏足够的 3D 户外数据集，3D 人体姿态估计的性能仍有继续提升的空间。最近，一些弱监督方法已经被提出来解决这个问题，并且在一定程度上取得了一些效果。

在本文中，我们首先介绍 2D 和 3D 人体姿态估计，包括从数据集角度，评价指标角度等。然后，我们详细介绍单目人体姿态估计的深度学习方法以及近年的主要成果。接着，我们会列举单目人体姿

* 通讯作者　Corresponding author：高浩，tsgaohao@gmail.com

收稿日期：2023-06-14；　录用日期：2023-11-27；　发表日期：2023-12-28

态估计在行业中的应用,以突出单目人体姿态研究的重要性。最后,我们分析了当前研究热点和挑战,并展望了单目人体姿态估计的未来发展方向。

1 背景

1.1 深度学习框架概述

深度学习技术显著提高了人体姿态估计的准确性和效率。与传统算法不同,深度学习技术可以从大数据中学习特点任务的特征。复杂的网络通常由多层网络或多个网络模块构成。例如,在姿态估计中,我们使用深度学习网络作为姿态编码器,紧接着是姿态解码器模块。姿态编码器(也称为"骨干")旨在通过逐渐减小的分辨率这一过程来学习图像高级特征。然后,姿态解码器模块以逐渐增加分辨率的方式来检测具有高级特征的关键点。近年来,研究人员广泛探索如何设计更有效的网络架构,以更好地学习数据特征。这包括使用不同变体的卷积神经网络(CNN)、循环神经网络(RNN)、图卷积网络(GCN)和变换器(Transformer)等方法。

深度学习技术仍然面临许多挑战。其中之一是稳定性,即深度学习网络对于输入的微小扰动可能会导致显著的失真。针对这一问题,研究人员致力于设计更稳定的网络架构。另一个挑战是深度学习方法的黑箱性质,导致结果难以解释。如何解释网络功能以及如何更好地在网络中融入领域知识仍然是个问题。同时,模型泛化能力也是一个具有挑战性的问题,因为尽管深度学习模型在训练数据上可以有出色的表现,但在面对新的、不同于训练数据分布的数据时,其性能可能下降。

尽管如此,深度学习已经成功应用到许多计算机视觉任务中,例如目标检测、图像分类、语义分割等。深度学习方法通过构建深层神经网络模型,可以自动地从大规模的数据中学习特征表示,并通过优化算法进行训练,从而实现高度准确的视觉任务。在目标检测中,深度学习模型可以有效地识别图像中的不同目标并进行定位。图像分类任务中,深度学习模型可以对图像进行分类,识别图像的类别。而在语义分割任务中,深度学习模型可以将图像的像素进行分类,实现对图像中每个像素的语义分割。

1.2 人体表示

人体表示,即用来描述复杂的人体姿态的方式。目前,对于人体姿态估计有两种表征形式:基于关键点的表示和基于模型的表示。本文主要介绍基于关键点的表示方法。

关键点:身体关键点可以通过 2D 或 3D 坐标来明确描述。为了形成骨架,关键点按照固有的身体结构进行连接。有很多方法直接通过回归预测人体的关键点坐标。

热图:高斯热图在对应的 2D/3D 坐标上具有高响应值,而在其他位置上具有低响应值。这种表示方式能够更好地捕捉关键点的位置信息,并为网络提供更准确的回归目标。通过使用高斯热图,我们可以在训练过程中让网络学习如何从输入图像中定位关键点,并预测它们的准确位置。这种方法在姿态估计等任务中取得了良好的效果,并且成为了深度学习中常用的关键点定位技术之一。

骨骼向量:在合成人体姿态时,可以利用骨骼向量组合成人体骨架。3D 人体骨架可以由一组骨骼向量表示,每个骨骼向量从父关键点指向子关键点。骨骼向量可以用球坐标表示。

1.3 数据集

数据集的快速发展推动了人体姿态估计方法的发展。公共数据集为不同方法提供了数据来源,具体的 2D 数据集如表 1 所示,3D 数据集如表 2 所示。以下将介绍一些目前使用最广泛的数据集。

MPII[3] 人体姿态数据集:MPII 数据集包含 28821 张用于训练的图片和 11701 张用于测试的图片。该数据集包含了 491 种人类活动,并涉及 4 千多名人员,骨架点为 16 个 2D 关键点。

COCO[4] 数据集:Microsoft COCO 是最常用的大规模数据集之一,其包含 33 万张图像,被注释图像超过 20 万张,常用于图像识别,目标检测,全景分割和姿态估计等视觉任务。对于 2D 人体姿态估计,该数据集包括 25 万个姿态注释和 20 万张标记图像。关于人体姿态估计,骨架为 17 个 2D 关键点。其已经成为评估模型性能的主要数据集

之一。

PoseTrack 数据集:PoseTrack 是一个大规模的公共数据集,用于人体姿态估计和关节跟踪的研究。该数据集包含了在拥挤环境中复杂运动和高度遮挡的情况。PoseTrack2017[8] 数据集包含 514 个视频片段,其中包括 16219 个姿势注释。而 PoseTrack2018[9] 数据集增加了视频片段的数量,达到 1138 个,总共包含 153615 个姿势注释。在训练视频中,为每个视频的 30 个中心帧提供注释。在验证视频中,每隔四帧进行一次人体姿态的注释。这两个数据集都标注了 15 个关节点,并提供了关节可见性的额外注释标签。

表 1 2D 人体姿态估计数据集总结

Table 1 Summary of 2D human posture estimation dataset

	数据集	年份	类型	关键点数	训练样本	验证样本	测试样本	评价指标
基于图像	LSP[1]	2010	单人	14	1000	/	1000	PCK
	Flic[2]	2013	单人	10	5000	/	1016	PCK
	MPII single[3]	2014	单人	16	28821	/	11701	PCK
	MPII Multi[3]	2014	多人	16	3800	/	1700	PCK
	COCO[4]	2017	多人	17	57000	5000	20000	mAP
	CrowdPose[5]	2019	多人	14	10000	2000	8000	mAP
基于视频	Penn Action[6]	2013	单人	13	1000	/	1000	mAP
	JHMDB[7]	2013	单人	15	600	/	300	mAP
	PoseTrack2017[8]	2017	多人	15	250	50	214	mAP
	PoseTrack2018[9]	2018	多人	15	592	170	375	mAP
	HiEve[10]	2020	多人	14	19	/	13	mAP

表 2 3D 人体姿态估计数据集总结

Table 2 Summary of 3D human posture estimation dataset

数据集	年份	帧数	受试者数量	相机视角
HumanEva-I[11]	2010	3.7 万	4	7
Human3.6M[12]	2013	360 万	11	4
CMU Panoptic[13]	2016	150 万	8	31
MPI-INF-3DHP[14]	2016	130 万	8	14
SURREAL[15]	2017	600 万	145	1
JTA[16]	2018	50 万	>21	1
3DPW[17]	2018	5.1 万	7	1

由于 3D 数据集的拍摄成本昂贵,且对具体的拍摄环境有着很高的要求,所以 3D 数据集比 2D 数据集更珍贵稀缺。

HumanEva-I[11]:该数据集在 60Hz 下从 7 个相机视角(4 个灰度和 3 个彩色)拍摄数据。视频分辨率为 659×494。该数据集有 4 个受试者,执行 6 个动作。通常对受试者 S1、S2 和 S3 的 3 个动作(步行、慢跑和拳击)进行模型性能验证。

Human3.6M[12]：该数据集是最广泛使用的单人姿态数据集，其在 4m×3m 的室内使用 4 个 RGB 相机，一个 time-of-flight 传感器，10 个运动相机捕获数据。它包含 360 万个 3D 人体姿势和 15 种场景下的视频。视频分辨率为 1000×1000 像素。目前，只有 7 名受试者数据可用。一般常用两种策略进行评估训练。第一种方案对五个受试者(S1，S5，S6，S7，S8)进行训练，对受试者 S9 和 S11 进行测试。第二种方案共享相同的训练/测试集，但仅在正面视角上进行评估。

MPI-INF-3DHP[14]：该数据集在 14 个摄像机工作室种使用商业无标记动作捕捉系统捕获数据。它由 8 名演员表演 8 项活动。测试集由 6 名受试者执行 7 个动作。动作范围从行走、坐到复杂的锻炼动作。动作类的数量多于 Human3.6M 数据集。为了增加数据的多样性，每个演员穿着日常服装和素色服装进行活动。

1.4 评价指标

2D 姿态估计的评估旨在测量预测的 2D 位置的准确性。广泛使用的评价指标包括：正确部位的百分比(PCP)，正确关键点的百分比(PCK)，平均精度(AP)。

PCP(Percentage of Correct Parts)：其常用来衡量身体部位预测的准确性。如果两个预测关节位置与真实肢体关节位置之间的距离几乎小于肢体长度的一半，则认为肢体被检测到。然而，有时对于较短的肢体，很容易被认为检测错误，例如下臂。

PCK(Percentage of Correct Keypoints)：PCK 被广泛用于测量 2D 关键点预测。用于计算预测的关键点与真值之间的归一化距离小于设定阈值的比例。PCKh@0.5 是 PCK 的轻微修改。采用阈值为被测人头段长度的 50%。通过使用头部尺寸作为参考，PCKh 使测量关节独立。通过改变阈值百分比，可以生成曲线下面积(AUC)以评估不同姿态估计算法的能力。

AP(Average Precision)：AP 用于多人姿势估计，通过测量关键点相似度(Object Keypoint Similarity，OKS)来计算。AP 正确地惩罚漏检和误检。如果检测到的关键点在与真实值的阈值内，则被视为真阳性。每个关键点都单独计算其与真值的对应关系。给定所有标记关键点的 OKS，可以计算平均精度(AP)和平均召回率(AR)。在不同的 OKS 下，AP 和 AR 可以全面反映测试算法的性能。

3D 姿态估计常用的评价指标有：MPJPE，P-MPJPE，MPJAE 等。

MPJPE(Mean Per Joint Position Error)：MPJPE 是 3D 姿态估计使用最广泛的评估度量。它以毫米为单位计算预测的 3D 关键点与真值之间的平均欧几里得距离。

P-MPJPE(Procrustes-Aligned MPJPE)：将预测点经过平移，旋转，缩放后，再计算预测点与真值之间的误差。这种对齐操作可以帮助消除预测姿态的整体偏移和旋转，从而更准确地评估预测结果的精确度。

MPJAE(Mean Per Joint Angle Error)：MPJAE 用于衡量预测的关节点之间的角度误差与真实姿态之间的差异。MPJAE 提供了一种评估算法在姿态估计任务中角度准确性的度量方式。较低的 MPJAE 值表示算法能够更准确地预测关节点之间的角度，而较高的 MPJAE 值则表示存在较大的角度误差。

2 单目 2D 人体姿态估计

单目 2D 人体姿态估计问题，包括单人姿态估计问题和多人姿态估计问题。但是，多人姿态估计也可以用于解决单人姿态估计问题，且目前对于 2D 姿态的研究多为多人问题，所以，本节我们将介绍单目 2D 多人人体姿态估计，表 3 列出了近几年的一些算法以及这些算法在 COCO 数据集中的表现。

关于多人 2D 人体姿态估计，可以划分为两类框架：自顶向下(top-down)和自底向上(bottom-up)。自顶向下将 2D 姿态估计分为两步，首先检测人类边界框，然后针对每个边界框提取人体姿态。如图 1 所示，边界框为人体检测算法检测结果，再根据人体框检测身体部位。自底向上的方法是首先找出图片种所有关键点，然后对关键点进行分组，从而得到一个人的完整骨架。如图 2 所示，先找关键点，再对关键点分组连接。通常来说，自顶向下方法具有更高的精度，而自底向上方法具有更快的推理速度，二者各有优缺。

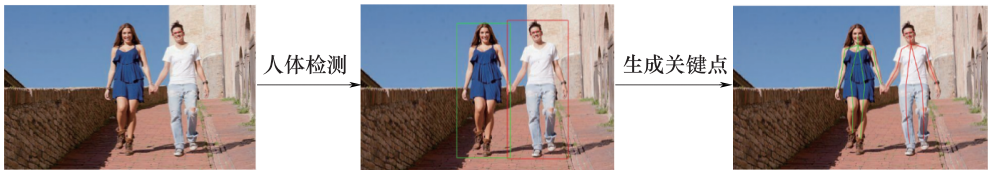

图 1 自顶向下框架流程

Fig. 1 Top-down framework flow

图 2 自底向上框架流程

Fig. 2 Bottom-up framework flow

表 3 单目 2D 多人姿态估计在 COCO 验证集中的表现

Table 3 Performance of monocular 2D multi-person attitude estimation in the COCO validation set

方法	出版时间	骨干网络	输入图像大小	mAP
自顶向下(top-down)方法				
SimpleBaseline[62]	ECCV'18	ResNet-152	384×288	73.7
HRNet-W48[32]	CVPR'19	HRNet-W48	256×192	75.1
DARK[70]	CVPR'20	HRNet-W48	384×288	76.8
TransPose-H/A6[81]	ICCV'21	HRNet-W48	256×192	75.8
TokenPose-L/D24[63]	ICCV'21	HRNet-W48	256×192	75.8
MSPN[64]	arXiv'19	4-stgMSPN	384×288	76.9
HRFormer-B[75]	arXiv'21	HRFormer-B	384×288	77.2
SimCC[80]	ECCV'22	HRNet-W48	384×288	76.9
PCT[66]	CVPR'23	Swin-Huge	256×256	79.3
ViTPose-H[71]	arXiv'22	ViTPose-H	256×192	79.1
自底向上(bottom-up)方法				
HigherHRNet[37]	CVPR'20	HRNet-W32	640×640	70.6
HigherHRNet	CVPR'20	HRNet-W48	640×640	72.1
DEKR[41]	CVPR'21	HRNet-W48	640×640	71.4
CenterGroup[42]	ICCV'21	HigherHRNet-W48	640×640	73.3
SWAHR[38]	CVPR'21	HigherHRNet	640×640	73.2

2.1 自顶向下

自顶向下的方法可以继续细分为回归方法,基于热图的方法等,我们将分别从这两个角度介绍自顶向下的人体姿态估计方法。

2.1.1 回归方法

早期的方法直接通过端到端的学习来实现从图像中直接预测关键点坐标,这种方法称为回归方法。

例如,DeepPose[18]网络是第一个利用深度神经

网络来解决姿态估计问题的网络。该网络首先使用级联的深度神经网络来提高关节点定位的准确性，最后再通过一个全连接层来回归出关键点坐标。该算法的提出，标志着姿态估计算法由传统算法向深度学习算法过渡。Sun 等[19]则利用堆叠沙漏网络（Hourglass Network[28]）作为网络骨干，提出了一种名为结构感知的回归方法。它使用骨骼向量而不是关节点坐标来表示人体姿态，并重新定义了损失函数来优化训练。

图卷积神经网络[20]近年来被广泛研究，它采用节点和边的方式来表示实体之间的关系。Qiu 等[21]提出了一个新颖的框架。该框架由两个阶段构成，第一个阶段产生初始姿态，第二个阶段利用 GCN 网络调整姿态。他们将人体结构转换为图形结构，其中节点表示人体关键点，边缘表示人体骨骼，并提出基于图像引导的图卷积模块，即关节点位置的图像特征作为图节点的输入，这种模块可以用于推理不可见关键点坐标。此外，相较于图卷积方法，Transformer[22]方法在图像领域有着更好的特征提取效果。其通过自注意力机制可以得到输入序列之间的关联性。Li 等[23]提出了级联 Transformer 网络来进行端到端的人体关键点检测，首先检测所有人的边界框，然后分别回归每个人的所有关键点坐标。

回归方法是一种比较高效的方法，特别适用于实时应用。然而，这种方法直接输出每个关节的单个 2D 坐标，却忽略了身体部位的大小差异。为了解决这个问题，出现了基于热图的方法，它使用概率热图而不是直接确定坐标来定位关键点。概率热图表示每个像素属于关键点的概率，从而提供了更丰富的信息，可以提高关键点定位的准确性。

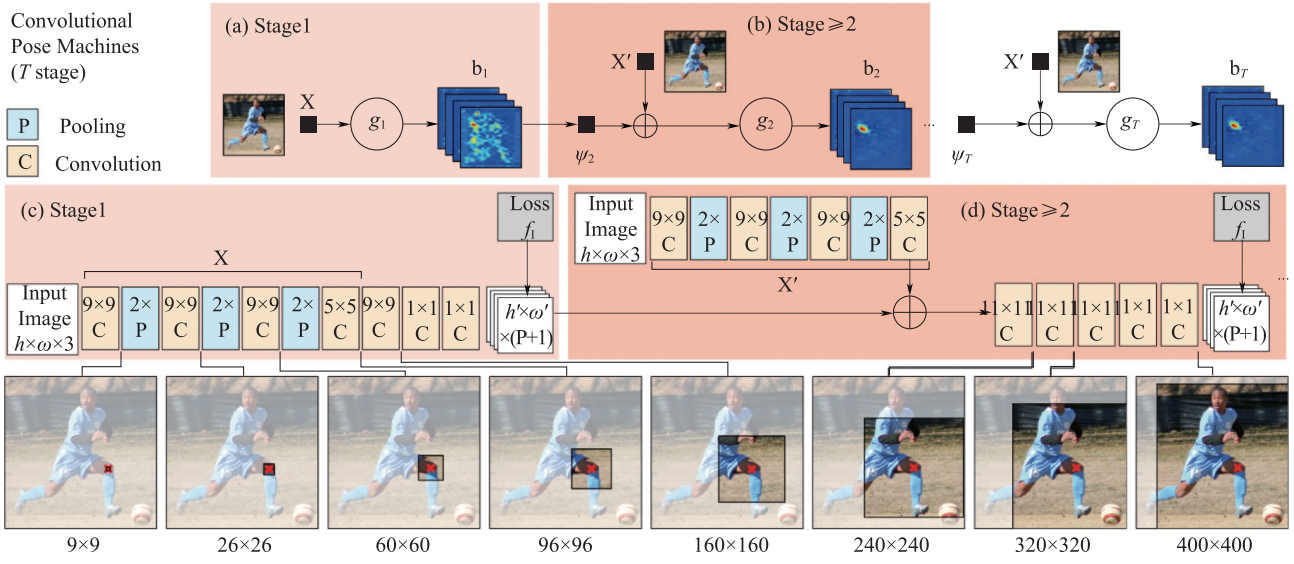

图 3　Wei 等[25]设计的可迭代化网络

Fig. 3　Iteratable network designed by Wei, et al.[25]

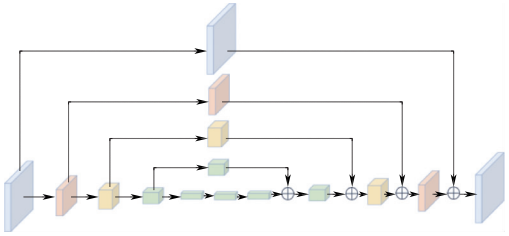

图 4　Hourglass 网络结构

Fig. 4　Hourglass network structure

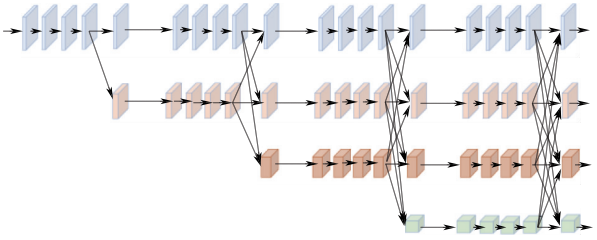

图 5　HRNet 网络结构

Fig. 5　HRNet network structure

2.1.2 基于热图

基于热图的方法的主要思想是将每个关键点表示为一个二维高斯分布的热图,其中峰值位置对应于关键点的位置。这样,对于每个像素,可以计算其属于每个关键点的概率,并根据最大概率确定关键点的位置。通过使用概率热图,能够考虑到关键点周围像素的信息,从而提高关键点的定位精度。这种方法也是目前2D姿态估计的主流方法。

一些模型采用可迭代的框架来逐步增强模型的精度,即先提取关键点特征然后再逐步细化微调,每一次迭代都会进一步改善输出结果。Ramakrishna 等[24]提出了一个推理模型,可以在多个阶段逐步推断并优化关节位置。Wei 等[25]进一步扩展了该结构。网络由多个阶段组成,每一阶段都有监督训练,可以减轻迭代过程中的梯度消失问题。其网络结构如图3所示,每个阶段的预测和图像特征被连接用于后续的细化微调阶段。随着残差网络(ResNet[26])的提出,梯度消失问题得到解决,它可以让更深层的网络进行反向传播,这极大地推动了2D人体姿态估计的进程,许多大模型被设计出来。

有一些深度网络采用对称的架构来提取人体特征,这类模型通常采用高分辨率到低分辨率(下采样)接着低分辨率到高分辨率(上采样)的方式。Newell 等[28]设计出了堆叠沙漏网络(Hourglass),如图4所示,该网络利用重复的上采样和下采样结构来融合不同尺度上的特征并捕获关节之间的各种空间关系以提高关节定位的精度。Chu 等[29]利用 Hourglass 网络整合了人体的整体热图和局部关节热图,整体热图关注的是人体的全局一致性,局部关节热图针对不同身体部位进行详细描述。然后,他们通过对自注意力模型热图的微调实现关键点定位。Ke 等[30]在 Hourglass 基础上提出了一种结构损失的损失函数,并利用多尺度监督方法来融合多尺度特征。Tang 等[31]采用 Hourglass 网络作为骨干网络,提出了采用多分支来预测相关性最高的关节点,而不是直接共享权值矩阵来一次性预测所有关键点坐标。这些方法都以 Hourglass 方法为基础,丰富了 Hourglass 方法,获得了比较不错的性能表现。

同时,也有一些网络采用非对称体系的结构。Chen 等[27]提出级联金字塔网络(CPN),该网络可以分为两步:全局网络和微调网络。全局网络对所有的关键点进行预测,但主要针对一些比较容易预测的点,如手、眼睛等,而微调网络是对全局网络预测的点进行微调,例如一些被身体遮住的点,该网络有效缓解了被遮挡点以及拥挤场景这样的问题。Sun 等[32]提出了 HRNet 网络,如图5所示,该网络能够在整个过程中保持高分辨率表示,通过融合低分辨率分支上的高等级特征,来达到多尺度特征融合的效果。这项工作证明了高分辨率表示对人体姿态估计的优越性。Liu 等[33]采用 HRNet 作为网络的骨干,并提出了一个新的多帧人体姿态估计框架。该框架利用前后两帧以及当前帧的时间与空间信息,来解决人体姿态估计中视频序列帧丢失等问题。Liu 等[34]则提出了一种更加精细的双重注意力机制,可以在通道和空间维度上保持高分辨率特征。这种双重注意力机制能够提供更准确的姿态估计结果,并进一步推动了姿态估计领域的发展。

在最近几年,Transformer[22]在图像领域大获成功,一些科研工作者也将 Transformer 应用到人体姿态估计领域。Xu 等[35]提出的 ViTPose 网络是目前2D人体姿态估计效果最好的网络。该网络主要分为编码(encoder)和解码(decoder)两个模块。编码模块主要将图片分块,然后送入自注意力机制提取特征以及相关性。在解码模块中,将特征矩阵解码为热图的形式得出关键点坐标。

总体而言,基于热图的方法比基于回归的方法有更高的准确性,所以,现有的方法大多在热图的基础上进行。但是,热图方法也带来了其他一些问题,例如需要足够大的计算力和热图的量化误差问题等。

2.2 自底向上

与自顶向下的方法相比,自底向上的方法不依赖于人体检测框,而是直接在原始图像上执行关键点估计,然后将关键点分组到每个人中,或者直接回归同一个人的关键点坐标。以下,我们将重点介绍一些近年的研究方法。

Cao 等[36]提出了 Openpose 方法,这是第一个在人体图像上实现实时多任务的系统。该系统不

仅可以检测人体的关键点，还可以检测手部和面部等其他关键点。这篇文章提出了部位亲和场(PAF)，主要用于编码关键点位置与方向信息，从而建模两个部位之间的相互关系，以此实现自底向上的人体姿态估计。这项工作对于人体姿态估计领域的研究具有重要意义，并为实时多任务系统的发展做出了重要贡献。

Cheng 等[37]为了解决由于尺度变化导致小人物骨架预测困难的问题，提出了一种新的 bottom-up 人体姿态估计方法：HigherHRNet，使用高分辨率特征金字塔学习多尺度特征。Luo 等[38]提出了尺度自适应的热图回归和权重自适应的热图回归方法，可以用来解决小尺度人体中利用大高斯核的热图问题。Jin 等[39]则利用图网络进行关键点聚类，为关节点分组方法提供了一种新思路。

在一些方法中，可以直接预测同一个人的关键点位置。例如，Wang 等[40]提出一个无须边界框检测和关键点分组的姿态推理网络。该网络不预测单个关键点，而是直接从一个人的可见身体部分预测出其完整的骨架。Geng 等[41]提出了解构式关键点回归(DEKR)方法，采用多分支结构让特征更加集中到关键点周围的区域。Bras'o 等[42]则利用 Transformer 进行关键点分组，可以实现端到端训练并加快模型推理速度。在拥挤图像中，该方法有更好的效果。

总体而言，自底向上的方法通过去除额外的目标检测技术来提高姿态检测的效率。由于自底向上方法具有较高的效率，在实际应用中具有很好的前景。例如，Openpose 已经在行业中被广泛应用。

3 单目 3D 人体姿态估计

对于单目 3D 人体姿态估计，我们可以将其分为两类：基于骨架的姿态估计和基于模型的姿态估计。基于模型的姿态估计，例如 SMPL[79]，通过建立人体姿态的模型来推断人体在 3D 空间中的姿态信息。基于骨架的姿态估计，只需要求出人体关键点在 3D 空间的位置。基于模型的姿态估计相较于骨架，更具挑战性，此处，我们主要介绍近些年基于骨架的 3D 姿态估计方法。

基于骨架的单目 3D 人体姿态估计，可以直接从图像中预测 3D 骨架，也可以借助 2D 骨架来实现 3D 骨架的预测，即 2D to 3D。

3.1 图像到 3D

3D 人体姿态估计最直接的方法就是直接设计端到端的网络来预测人体关键点。一些方法借助体积热图进行预测。例如，Luvizon 等[43]设计了一个多任务框架，可以直接从图像中预测人体 2D 或 3D 关键点坐标。Pavlakos 等[44]对三维空间进行划分，并训练 CNN 网络来预测每个关键点在三维网格中的可能性。此外，对于深度信息，他们采用二阶段算法，即首先预测 2D 关键点热图，然后在 3D 网格空间坐标上进行 3D 关节点坐标推理。Zhou 等[45]提出部位-中心-热图三元组，来构建空间体积，再用积分的方式实现端到端训练。这些方法都依赖于将热图坐标转换为关节坐标这个额外步骤，此外，预测的关键点的精度与热图分辨率的精度成比例，这缺乏固有的空间泛化性。为了达到高精度，预测的热图通常需要适当的空间分辨率，这会成倍地增加计算成本和内存消耗。

也有一些方法则直接从图像回归人体关键点坐标，而不采用热图的方式。例如，Dabral 等[46]提出了一个简单的时间网络，可以利用姿势序列中存在的时间和人体结构线索来暂时协调姿态估计，同时他们也提出了两个损失函数来与弱监督相结合。Sun 等[47]提出了一种结构感知方法，具体方式为利用骨骼而不是关键点来进行姿态估计。

总之，从图像到 3D 的端到端估计系统，会受到自遮挡和其他对象遮挡、深度模糊和训练数据不足的影响。此外，从图像到 3D 的网络更复杂，需要更大的计算力。Martinez 等[48]利用两个简单的全连接模块来实现 2D 骨架到 3D 骨架的升维，发现相对简单的深度前馈网络在性能上要比当时最好的从图像到 3D 结果好 30%，并提出从图像到 3D 姿态估计系统的很大一部分误差来源于其视觉分析即 2D 姿态理解出错。因此，2D 到 3D 的网络相较于基于图像到 3D 的网络，具有更好的性能表现。

表 4 2D to 3D 单目 3D 人体姿态估计在 Human3.6M 数据集(CPN 提取 2D 坐标)中的表现

Table 4 Performance of 2D to 3D monocular 3D human pose estimation in Human 3.6M dataset(CPN extracting 2D coordinates)

方法	出版时间	输入长度	MPJPE	P-MPJPE
Cai 等[54]	CVPR'19	7 帧	48.8	39.0
Pavllo 等[49]	CVPR'19	243 帧	46.8	36.5
Liu 等[82]	CVPR'20	243 帧	45.1	35.6
Zheng 等[56]	CVPR'21	81 帧	44.3	34.6
Chen 等[51]	TCSVT'21	243 帧	44.1	35.0
Li 等[57]	CVPR'21	351 帧	43.0	34.5
Zhang 等[58]	CVPR'22	243 帧	40.9	32.6
Shan 等[78]	arXiv'23	243 帧	39.5	31.6

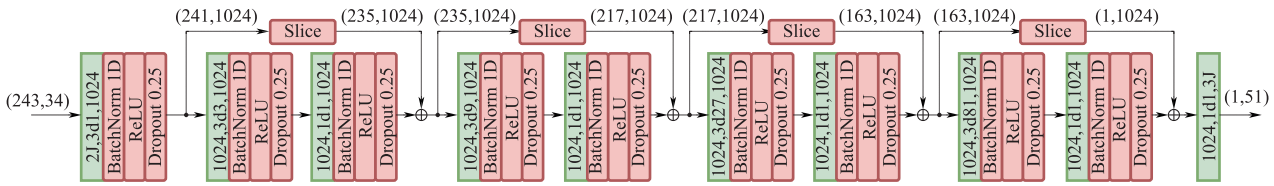

图 6 VideoPose 网络结构

Fig. 6 VideoPose network structure

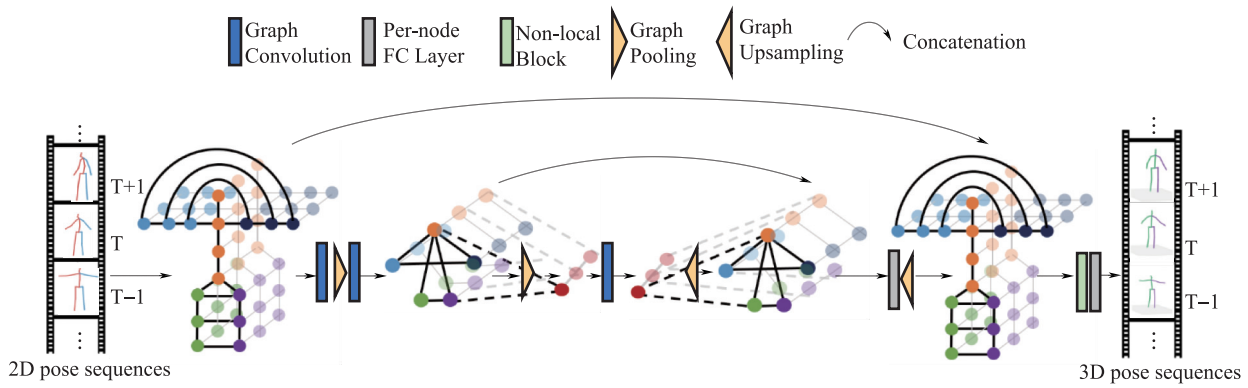

图 7 Cai 等[54]提出的图卷积网络结构

Fig. 7 Graph Convolution Network structure proposed by Cai,et al [54]

3.2 2D to 3D

对于 2D 到 3D 人体姿态估计,我们按照使用的网络模型将其分为卷积神经方法,图卷积方法和 transformer 方法。

卷积神经方法,即网络使用卷积神经网络(CNN)来提取 2D 人体骨架特征。Pavllo 等[49]提出 VideoPose,网络结构如图 6 所示,该网络利用卷积神经网络来捕获时间序列信息,具有较大的感受野。在此之前,大部分方法都是使用 LSTM 来提取视频序列的人体骨架特征,VideoPose 也是首个利用 CNN 提取视频序列信息的方法。此后的很多方法都是基于该方法实现。Zeng 等[50]将人体分为多个区域分别训练,在网络的最后,将多个分支的训练结果组合到一起。在 CNN 方法中,可以达到提高网络模型精度的效果。对于一些困难动作,可以有效降低深度模糊性。Chen 等[51]采用类似于 VideoPose 那样的网络,从人体解剖学出发,将任务

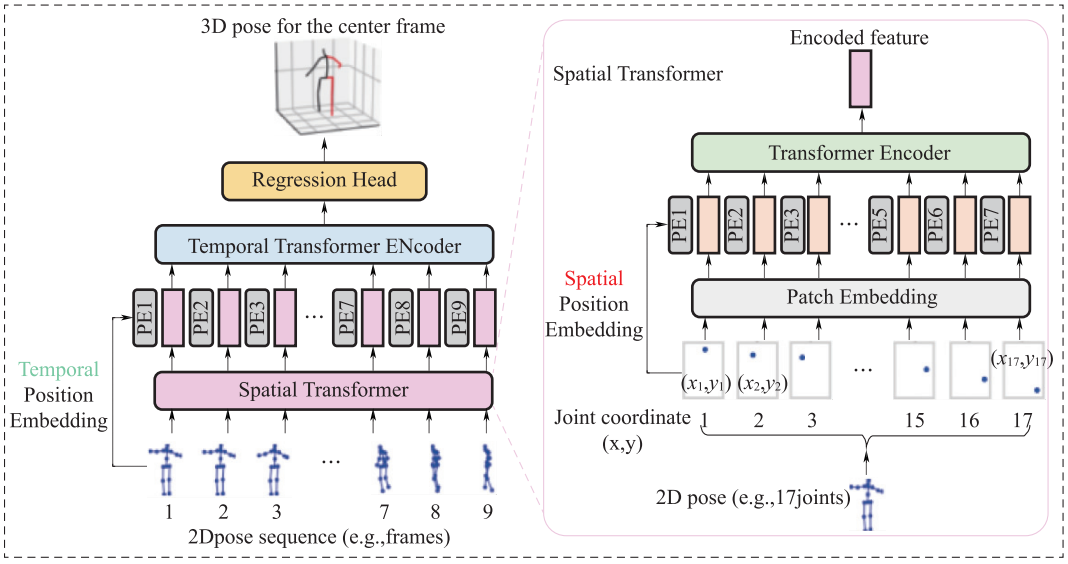

图 8　PoseFormer 网络结构

Fig. 8　PoseFormer network structure

分解为骨长预测和骨方向预测。Zhan 等[52]先将原始 2d 数据映射为 3d 射线，并进行归一化，然后进行训练，可以降低摄像机内外参对模型性能的影响。

图卷积方法指网络采用图卷积这类方法来实现 2D to 3D。Zhao 等[53]改进一般图卷积方法，不共享卷积核参数。同时他们对人体关键点分类，分为自身节点和相邻节点。图卷积可以自定义不同节点之间的连接关系。Cai 等[54]仿照 2D 姿态中的 Hourglass 网络，在图网络中增加了图池化层和图上采样，以实现不同尺度的特征提取，该网络结构如图 7 所示。同时，他们也对图卷积方法进行了改进，提出了一种基于一般图卷积运算的非均匀图卷积策略，即根据不同相邻节点的语义含义学习不同的卷积核权重。Zeng 等[55]提出了一种跨越感知的分层通道压缩融合层，可以有效地从相邻节点提取信息，同时抑制噪声，具体来说，挤压长距离上下文特征（即来自遥远节点的信息），并以分层的方式将它们与短距离特征（相邻或相近节点信息）融合。该网络可以较好地预测困难姿势。

近几年，Transformer 迅速发展，因此，2D to 3D 涌现出了很多 Transformer 算法。Zheng 等[56]提出 PoseFormer，以 Vision Transformer 为基础，搭建空间 transformer 提取人体骨架空间信息和时间 transformer 提取时间序列之间的关系。Li 等[57]提出 MHFormer，该算法可以学习多个合理姿势假设的时空表示，然后在假设之间建立关系。因为存在深度模糊性，所以对于单个 2d 骨架，存在多个可行解，因此，多解问题的目的为融合多个可行解，选出最接近的 3D 解。Zhang 等[58]在 PoseFormer 方法的基础上对时间 Transformer 的输入进行改进，同时采用输入骨架序列输出也为骨架序列的形式，以减少计算冗余，达到加快训练速度的目的。Zhu 等[59]提出了二阶段算法，即先采用大量数据预训练模型，主要学习人体运动学特征，接着再在 human3.6 数据集上精训练，以达到最大限度地学习人体特征的效果。

总体而言，2D to 3D 近年来发展迅速，尤其是在 Transformer 方法的基础上，很多研究者对其进行了大量改进。然而，2D to 3D 仍有大量改进空间，例如，扩充 3D 数据集，探索困难姿势的深度模糊性，减小网络参数实现轻量化等。

4　行业应用

3D 姿态估计相比 2D 姿态估计，多了一个深度信息，因此 3D 姿态估计能够实现更广泛的应用。为了更好地理解 2D/3D 姿态估计，我们列举了一些关于姿态估计的应用。

4.1　人机交互

如果机器人能够理解人的 3D 姿势、动作和情

感,它就能更好地服务和帮助用户。例如,当机器人检测到容易跌倒的人的 3D 姿势时,它可以及时采取行动。此外,辅助机器人可以更好地与人类用户进行社交互动,前提是它们可以感知 3D 人类姿势。姿态估计可以用于开发姿势识别游戏,例如舞蹈游戏或体感游戏。通过追踪用户的姿势动作,游戏可以根据用户的表现给予反馈、评分或奖励,增加游戏的互动性和娱乐性。例如,Microsoft[60] 的 Kinect 传感器可以让计算机直接感知玩家和环境的三维(深度)。它还能理解用户何时说话,并识别用户是谁,而且可以解释他们的动作。姿态估计也可以用于身体动作捕捉系统,将用户的身体动作转化为虚拟角色或机器人的动作。这在虚拟现实、动画制作、人机协作等领域具有重要应用,例如在图 9 的电影制作中,通过关键点追踪技术可以实时捕捉演员的动作并将其应用于虚拟角色。

图 9 骨架点用于电影制作

Fig. 9 Skeleton points for movie production

4.2 自动驾驶

自动驾驶需要利用传感器或者摄像头识别行人,以免发生碰撞,因此理解行人的姿势,运动是十分重要的。如图 10 所示,通过估计行人的姿态,可以分析其行为和意图,从而改善自动驾驶系统对行人的感知和预测能力。例如,通过识别行人的姿态,可以判断其是否准备要横穿马路,从而及时采取相应的驾驶决策来避免事故。在自动驾驶过程中,也需要监控驾驶员的状态,以便在需要时重新将驾驶任务交还给驾驶员。通过人体姿态估计,可以监测驾驶员的姿势、眼睛的注视方向、疲劳程度等指标,从而及时发出警示或采取相应的措施。自动驾驶车辆需要提供乘客舒适的乘坐体验。通过人体姿态估计,可以评估乘客的姿势、坐姿以及情绪状态,从而根据实时反馈对座椅、空调等系统进行调整,以提供更好的乘坐舒适度。

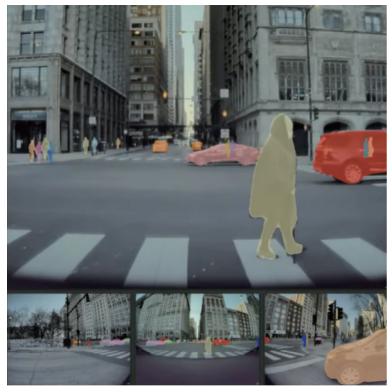

图 10 自动驾驶中的行人识别

Fig. 10 Pedestrian recognition in autonomous driving

4.3 医疗康复

人体姿态估计在医疗康复领域具有广泛的应用,例如运动评估与监控,姿势纠正和训练,功能恢复训练和运动损伤预防等。情景交互式康复训练和效果评估需要患者主要关节的三维空间位置,已有的人体姿态估计方法需要患者佩戴多种传感器或光学标志,可能影响运动和导致心理不适。同时,Kinect 自带的骨骼绑定算法在人体被部分遮挡时可能无法识别或误识别。唐心宇等[61] 提出一种基于 OpenPose 和 Kinect 的三维人体姿态估计方法,并创新应用到康复训练中。可以将 Openpose 得到的二维坐标与 Kinect 获得的深度数据融合得到三维坐标,然后用于训练并进行康复评估。

4.4 运动分析

姿态估计可以用于对人体进行动作识别和分类。如图 11 所示,通过监测关键点的位置和运动轨迹,系统可以识别不同的动作,例如跑步、跳跃、打球等。这对于运动研究、运动训练和运动表演等领域都非常有价值。姿态估计也可以用于检测和评估运动员或运动者的动作技术和执行质量。通过比较实际动作与理想动作的差异,可以识别和纠正不良的动作技术,提供实时反馈和指导,帮助改进运动表现和减少运动伤害。姿态估计还可以用于评估运动员的技能水平,并为他们提供个性化的训练方案。通过分析姿势数据和运动轨迹,可以评估

运动员的动作准确性、速度、力量等指标，并根据评估结果设计相应的训练计划，帮助运动员提高技能水平。

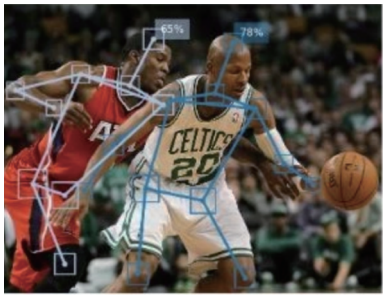

图 11　人体姿态估计用于姿态分析

Fig. 11　Human body pose estimation for posture analysis

5　未来发展趋势及挑战

5.1　未来发展趋势

姿态估计需要增强在困难姿势和复杂场景下的表现。例如，姿态估计可以用于评估运动员的技能水平，并为他们提供个性化的训练方案。通过分析姿势数据和运动轨迹，可以评估运动员的动作准确性、速度、力量等指标，并根据评估结果设计相应的训练计划，帮助运动员提高技能水平。一些研究已经尝试解决 2D 姿态估计中的人群和遮挡，但它们在真实的场景中仍然表现不佳。此外，对于 3D 姿态估计，遮挡将导致不合理的人体形状和姿态重建。由于复杂场景的上下文包含人与人和人与物之间交互的线索，进一步的工作可以利用场景和人的关系进行更好的推理。

随着基于骨架的姿态估计的发展，基于人体模型的姿态估计也会得到发展。3D 模型包含比 3D 骨架更多的信息，例如外表信息，丰富的表情等。随着 3D 人体模型发展，可以进一步促进人机交互，虚拟现实等的进展，例如生成具有情感，包含丰富表情的数字人等。

总之，单目人体姿态估计是一个具有挑战性和实际意义的问题。用于姿态估计的深度学习的发展是有前途和令人兴奋的。在未来，人体姿态估计的研究和应用都将面临许多机遇和挑战。单目人体姿态估计的未来将在很大程度上取决于算法，数据和应用场景的进展。

5.2　2D 姿态估计挑战

随着深度学习的快速发展，近年来有很多优秀的人体姿态估计方法被提出，并实现了较高的识别准确率。但是由于人体为一个非刚性物体，并没有固定的表现形式，所以难以进行定量研究，除此之外，各种内部外部因素也使人体姿态估计在研究过程中面临着巨大的挑战。

复杂背景：复杂的环境和图像中的背景信息有时会在颜色上与人体相似，这会增加人体姿态估计的困难，使模型难以准确识别关键点位置。

肢体运动灵活：人的肢体灵活性很高，具有很高的活动范围。这种灵活性会使人在不同的运动状态下产生不同的姿态。在图像中，会出现不同的肢体有相似的姿态，例如，跑步中，腿部和手臂会出现相似角度弯曲，造成训练困难。

遮挡问题：遮挡问题可分为自身遮挡和其他物体遮挡。不管是那种遮挡，都会给关键点预测带来困难。当图片中出现较大的遮挡时，仅根据局部区域很难预测完整姿态。

抖动问题：将 2D 人体姿态估计应用到视频上时，图像中人体关键点会有一定程度的抖动。如何降低视频中关键点的抖动也是一个全新的挑战。

热图量化误差问题：热图通常使用高斯分布或其他概率分布来表示关键点的位置，其中关键点处具有较高的响应值，而其他位置则具有较低的响应值。然而，由于热图的离散化和量化过程，会引入一定的误差。

关键点分组：当采用自底向上的姿态估计方案时，会面临关键点分组的问题。正确的关键点分组可以提高姿态估计的精度和鲁棒性，而错误的分组可能导致姿态估计的错误或不完整。因此，在设计自底向上的姿态估计算法时，需要考虑有效的关键点关联和姿态匹配策略，以解决关键点分组问题并获得准确的姿态估计结果。

5.3　3D 姿态估计挑战

目前的 3D 姿态估计主流方法都是建立在 2D 姿态估计基础之上，所以 3D 姿态估计最大的困难是 2D 姿态估计稳不稳的问题。当然，对于 3D 姿态估计，也有着一些其他的问题。

数据集稀缺:目前最大的数据集为 human3.6m,虽然是最大的数据集,但是公开的部分只有七名演员,这无疑会增加网络的过拟合问题。针对此问题,很多研究者采用弱监督或者无监督的方式进行训练,但是训练结果仍有很大提升空间。

深度模糊:对于同一个 2D 骨架,可能对应着多个 3D 骨架,其主要原因为深度不可知,这也是影响 3D 姿态估计精度的一个主要原因。很多科研人员通过精心设计网络,反复提取时间、空间特征来降低深度模糊性对预测精度的影响。

复杂背景、遮挡:对于 2D to 3D,没有这些问题。但是对于从图像到 3D,这些问题无疑会降低最终的预测精度,这也是为什么现在主流方法为 2D to 3D 的原因。

困难姿势预测:在绝大多数网络中,对于站立,走路这类简单动作,有着较好的预测结果,但是对于坐这类复杂动作,网络性能则会大幅度下降。因此,如何实现对困难姿势的预测也是一个重要问题。

算力需求:目前的 3D 姿态估计网络模型是越来越大,这无疑加剧了对算力的需求。所以,如何设计轻量化网络并不降低预测精度也是一大挑战。

6 结束语

随着计算机视觉以及深度学习的发展,人体姿态估计已经成为一项重要的研究课题。本文对最近几年的人体姿态估计项目进行并对行业应用进行阐述。此外,本文对人体姿态面临的挑战进行了总结,以及对未来发展趋势进行了展望。

利益冲突:作者声明无利益冲突。

参考文献(References)

[1] Johnson, Sam and Mark Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation[C]. British Machine Vision Conference, 2010.
DOI:10.5244/C.24.12

[2] Sapp, Benjamin and Ben Taskar. MODEC: Multimodal Decomposable Models for Human Pose Estimation[C]. 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013:3674-3681.
DOI:10.1109/CVPR.2013.471

[3] Andriluka, Mykhaylo, et al. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014:3686-3693.
DOI:10.1109/CVPR.2014.471

[4] Lin Tsung-Yi, et al. Microsoft COCO: Common Objects in Context[C]. European Conference on Computer Vision, 2014.
DOI:10.1007/978-3-319-10602-1_48

[5] Li, Jiefeng, et al. CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2018:10855-10864.
DOI:10.1109/CVPR.2019.01112

[6] Zhang, Weiyu, et al. From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding[C]. 2013 IEEE International Conference on Computer Vision, 2013:2248-2255.
DOI:10.1109/ICCV.2013.280

[7] Jhuang, Hueihan, et al. Towards Understanding Action Recognition[C]. 2013 IEEE International Conference on Computer Vision, 2013:3192-3199.
DOI:10.1109/ICCV.2013.396

[8] Iqbal, Umar, et al. Pose for Action-Action for Pose[C]. 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition(FG 2017), 2016:438-445.
DOI:10.1109/FG.2017.61

[9] Andriluka, Mykhaylo, et al. PoseTrack: A Benchmark for Human Pose Estimation and Tracking[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017:5167-5176.
DOI:10.1109/CVPR.2018.00542

[10] Lin, Weiyao, et al. Human in Events: A Large-Scale Benchmark for Human-centric Video Analysis in Complex Events. ArXiv abs/2005.04490, 2020:n. pag.
DOI:10.48550/arXiv.2005.04490

[11] Sigal, Leonid, et al. HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion[J]. International Journal of Computer Vision, 2010, 87:4-27.
DOI:10.1007/s11263-009-0273-6

[12] Ionescu, Catalin, et al. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments[C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36:

1325-1339.
DOI:10.1109/TPAMI.2013.248

[13] Joo,Hanbyul,et al. Panoptic Studio:A Massively Multiview System for Social Motion Capture[C]. 2015 IEEE International Conference on Computer Vision (ICCV),2015:3334-3342.
DOI:10.1109/ICCV.2015.381

[14] Mehta,Dushyant,et al. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision [C]. 2017 International Conference on 3D Vision (3DV),2016:506-516.
DOI:10.1109/3DV.2017.00064

[15] Varol,Gül,et al. Learning from Synthetic Humans[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2017:4627-4635.
DOI:10.1109/CVPR.2017.492

[16] Fabbri,Matteo,et al. Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World[C]. European Conference on Computer Vision,2018.
DOI:10.1007/978-3-030-01225-0_27

[17] Marcard,Timo von,et al. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera[C]. European Conference on Computer Vision,2018.
DOI:10.1007/978-3-030-01249-6_37

[18] Toshev, Alexander and Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks [C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition,2013:1653-1660.
DOI:10.1109/CVPR.2014.214

[19] Sun,Xiao,et al. Compositional Human Pose Regression [C]. 2017 IEEE International Conference on Computer Vision(ICCV),2017:2621-2630.
DOI:10.1109/ICCV.2017.284

[20] Kipf,Thomas and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. ArXiv abs/1609.02907,2016:n. pag.
DOI:10.48550/arXiv.1609.02907

[21] Qiu,Lingteng,et al. Peeking into occluded joints:A novel framework for crowd pose estimation. ArXiv abs/2003.10506,2020:n. pag.
DOI:10.1007/978-3-030-58529-7_29

[22] Vaswani, Ashish, et al. Attention is All you Need. NIPS,2017.
DOI:10.48550/arXiv.1706.03762

[23] Li, Ke, et al. Pose Recognition with Cascade Transformers[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 1944-1953.
DOI:10.1109/CVPR46437.2021.00198

[24] Ramakrishna, Varun, et al. Pose Machines: Articulated Pose Estimation via Inference Machines[C]. European Conference on Computer Vision,2014.
DOI:10.1007/978-3-319-10605-2_3

[25] Wei, Shih-En, et al. Convolutional Pose Machines[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2016:4724-4732.
DOI:10.1109/CVPR.2016.511

[26] He, Kaiming, et al. Deep Residual Learning for Image Recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 770-778.
DOI:10.1109/cvpr.2016.90

[27] Chen, Yilun, et al. Cascaded Pyramid Network for Multi-person Pose Estimation[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,2017:7103-7112.
DOI:10.1109/CVPR.2018.00742

[28] Newell, Alejandro, et al. Stacked Hourglass Networks for Human Pose Estimation[C]. European Conference on Computer Vision,2016.
DOI:10.1007/978-3-319-46484-8_29

[29] Chu, Xiao, et al. Multi-context Attention for Human Pose Estimation[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2017: 5669-5678.
DOI:10.1109/CVPR.2017.601

[30] Ke,Lipeng,et al. Multi-Scale Structure-Aware Network for Human Pose Estimation[C]. European Conference on Computer Vision,2018.
DOI:10.1007/978-3-030-01216-8_44

[31] Tang, Weixian and Ying Wu. Does Learning Specific Features for Related Parts Help Human Pose Estimation?" [C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 1107-1116.
DOI:10.1109/CVPR.2019.00120

[32] Sun, Ke, et al. Deep High-Resolution Representation Learning for Human Pose Estimation[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2019:5686-5696.

DOI:10.1109/CVPR.2019.00584

[33] Liu,Zhenguang,et al. Deep Dual Consecutive Network for Human Pose Estimation[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2021:525-534.
DOI:10.1109/CVPR46437.2021.00059

[34] Liu, Huajun, et al. Polarized Self-Attention: Towards High-quality Pixel-wise Regression. ArXiv abs/2107.00782,2021:n. pag.
DOI:10.1016/j.neucom.2022.07.054

[35] Xu, Yufei, et al. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. ArXiv abs/2204.12484,2022:n. pag.
DOI:10.48550/arXiv.2204.12484

[36] Cao,Zhe,et al. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016:1302-1310.
DOI:10.1109/CVPR.2017.143

[37] Cheng,Bowen,et al. HigherHRNet:Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019:5385-5394.
DOI:10.1109/cvpr42600.2020.00543

[38] Luo, Zhengxiong, et al. Rethinking the Heatmap Regression for Bottom-up Human Pose Estimation[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2020:13259-13268.
DOI:10.1109/CVPR46437.2021.01306

[39] Jin, Sheng, et al. Differentiable Hierarchical Graph Grouping for Multi-Person Pose Estimation. ArXiv abs/2007.11864,2020:n. pag.
DOI:10.1007/978-3-030-58571-6_42

[40] Wang, Dongkai, et al. Robust Pose Estimation in Crowded Scenes with Direct Pose-Level Inference [R]. Neural Information Processing Systems,2021.
DOI:10.24963/ijcai.2021/5271

[41] Geng, Zigang, et al. Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2021:14671-14681.
DOI:10.1109/CVPR46437.2021.01444

[42] Bras'o,Guillem,et al. The Center of Attention:Center-Keypoint Grouping via Attention for Multi-Person Pose Estimation[C]. 2021 IEEE/CVF International Conference on Computer Vision(ICCV),2021:11833-11843.
DOI:10.1109/ICCV48922.2021.01164

[43] Luvizon,Diogo Carbonera,et al. 2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,2018:5137-5146.
DOI:10.1109/CVPR.2018.00539

[44] Pavlakos, Georgios, et al. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2016:1263-1272.
DOI:10.1109/CVPR.2017.139

[45] Zhou,Kun,et al. HEMlets Pose:Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation[C]. 2019 IEEE/CVF International Conference on Computer Vision(ICCV),2019:2344-2353.
DOI:10.1109/ICCV.2019.00243

[46] Dabral,Rishabh,et al. Learning 3D Human Pose from Structure and Motion [C]. European Conference on Computer Vision,2017.
DOI:10.1007/978-3-030-01240-3_41

[47] Sun,Xiao,et al. Compositional Human Pose Regression [C]. 2017 IEEE International Conference on Computer Vision(ICCV),2017:2621-2630.
DOI:10.1109/ICCV.2017.284

[48] Martinez,Julieta,et al. A Simple Yet Effective Baseline for 3d Human Pose Estimation[C]. 2017 IEEE International Conference on Computer Vision(ICCV),2017:2659-2668.
DOI:10.1109/ICCV.2017.288

[49] Pavllo,Dario,et al. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018:7745-7754.
DOI:10.1109/CVPR.2019.00794

[50] Zeng, Ailing, et al. SRNet: Improving Generalization in 3D Human Pose Estimation with a Split-and-Recombine Approach. ArXiv abs/2007.09389,2020:n. pag.
DOI:10.1007/978-3-030-58568-6_30

[51] Chen,Tianlang,et al. Anatomy-Aware 3D Human Pose Estimation With Bone-Based Pose Decomposition[C]. IEEE Transactions on Circuits and Systems for Video Technology 32,2021:198-209.

DOI:10.1109/TCSVT.2021.3057267

[52] Zhan, Yu-Wei, et al. Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2022:13106-13115.
DOI:10.1109/CVPR52688.2022.01277

[53] Zhao, Long, et al. Semantic Graph Convolutional Networks for 3D Human Pose Regression[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2019:3420-3430.
DOI:10.1109/CVPR.2019.00354

[54] Cai, Yujun, et al. Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks[C]. 2019 IEEE/CVF International Conference on Computer Vision(ICCV),2019:2272-2281.
DOI:10.1109/ICCV.2019.00236

[55] Zeng, Ailing, et al. Learning Skeletal Graph Neural Networks for Hard 3D Pose Estimation[C]. 2021 IEEE/CVF International Conference on Computer Vision(ICCV),2021:11416-11425.
DOI:10.1109/ICCV48922.2021.01124

[56] Zheng, Ce, et al. 3D Human Pose Estimation with Spatial and Temporal Transformers[C]. 2021 IEEE/CVF International Conference on Computer Vision(ICCV),2021:11636-11645.
DOI:10.1109/ICCV48922.2021.01145

[57] Li, Wenhao, et al. MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2021:13137-13146.
DOI:10.1109/CVPR52688.2022.01280

[58] Zhang, Jinlu, et al. MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022:13222-13232.
DOI:10.1109/CVPR52688.2022.01288

[59] Zhu, Wenjie, et al. MotionBERT: Unified Pretraining for Human Motion Analysis. ArXiv abs/2210.06551, 2022:n. pag.
DOI:10.48550/arXiv.2210.06551

[60] Zhang, Zhengyou. Microsoft Kinect Sensor and Its Effect[J]. IEEE Multim,2012,19:4-10.
DOI:10.1109/MMUL.2012.24

[61] 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.
DOI:10.19650/j.cnki.cjsi.J1803879

[62] Xiao, Bin, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking[C]. Proceedings of the European conference on computer vision(ECCV). 2018.
DOI:10.1007/978-3-030-01231-1_29

[63] Li, Yanjie, et al. Tokenpose: Learning keypoint tokens for human pose estimation[C]. Proceedings of the IEEE/CVF International conference on computer vision. 2021.
DOI:10.1109/ICCV48922.2021.01112

[64] Li, Wenbo, et al. Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148,2019.
DOI:10.1109/TPAMI.2019.2958916

[65] Zhang, Feng, et al. Distribution-aware coordinate representation for human pose estimation[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
DOI:10.1109/cvpr42600.2020.00712

[66] Geng, Zigang, et al. Human Pose as Compositional Tokens[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
DOI:10.1109/CVPR52729.2023.00071

[67] Liu, Ze, et al. Swin transformer v2: Scaling up capacity and resolution[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
DOI:10.1109/CVPR52688.2022.01170

[68] Liu, Huajun, et al. Polarized self-attention: Towards high-quality pixel-wise regression. arXiv preprint arXiv:2107.00782,2021.
DOI:10.48550/arXiv.2107.00782

[69] Zhang, Jing, Zhe Chen, and Dacheng Tao. Towards high performance human keypoint detection[J]. International Journal of Computer Vision 129.9,2021:2639-2662.
DOI:10.1007/s11263-021-01482-8

[70] Zhang, Feng, et al. Distribution-Aware Coordinate Representation for Human Pose Estimation[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2019:7091-7100.
DOI:10.1109/cvpr42600.2020.00712

[71] Xu, Yufei, et al. Vitpose: Simple vision transformer baselines for human pose estimation. arXiv preprint arXiv:2204.12484,2022.

DOI:10.48550/arXiv.2204.12484

[72] Dosovitskiy, Alexey, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929,2020.
DOI:10.48550/arXiv.2010.11929

[73] He, Kaiming, et al. Mask r-cnn. Proceedings of the IEEE international conference on computer vision. 2017.
DOI:10.1109/ICCV.2017.322

[74] Papandreou, George, et al. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model[C]. Proceedings of the European conference on computer vision (ECCV). 2018.
DOI:10.1007/978-3-030-01264-9_17

[75] Yuan, Yuhui, et al. Hrformer: High-resolution transformer for dense prediction. arXiv preprint arXiv: 2110.09408,2021.
DOI:10.1109/CVPR.2021.01300

[76] McNally, William, et al. Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation[C]. Computer Vision-ECCV 2022:17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI. Cham: Springer Nature Switzerland,2022.
DOI:10.1007/978-3-031-20068-7_3

[77] Li, Jiefeng, et al. Human pose regression with residual log-likelihood estimation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
DOI:10.1109/ICCV48922.2021.01084

[78] Shan, Wenkang, et al. Diffusion-Based 3D Human Pose Estimation with Multi-Hypothesis Aggregation. ArXiv abs/2303.11579,2023:n. pag.
DOI:10.48550/arXiv.2303.11579

[79] Loper, Matthew, et al. SMPL: A skinned multi-person linear model[J]. ACM transactions on graphics(TOG) 34.6,2015:1-16.
DOI:10.1145/3596711.3596800

[80] Li, Yanjie, et al. SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation[C]. European Conference on Computer Vision,2021.
DOI:10.1007/978-3-031-20068-7_6

[81] Yang, Sen, et al. TransPose: Keypoint Localization via Transformer[C]. 2021 IEEE/CVF International Conference on Computer Vision(ICCV),2020:11782-11792.
DOI:10.1109/ICCV48922.2021.01159

[82] Liu, Ruixu, et al. Attention Mechanism Exploits Temporal Contexts: Real-Time 3D Human Pose Reconstruction[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 5063-5072.
DOI:10.1109/cvpr42600.2020.00511

A Review of Human Pose Estimation Methods Based on Monocular Vision

YAN Xin, GAO Hao*, LI Haolun

(Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: This article is a review of the technology methods and industry applications of monocular human pose estimation, which focuses on the recent development of monocular human pose estimation. With the rapid development of computer vision and machine learning, monocular human pose estimation has become a research direction that has attracted much attention. In this article, we first introduce the relevant concepts and significance of monocular human pose estimation, explain the importance of research in this field. Then, we provide a detailed overview of the technical methods of monocular human pose estimation, including two different approaches: monocular 2D human pose estimation and monocular 3D human pose estimation. For each research method, we discuss its principles, development, and pros and cons. Next, we explore the applications of monocular human pose estimation in various fields such as human-computer interaction, autonomous driving, healthcare, and provide examples to highlight the significance of human pose estimation. Finally, we analyze the current research hotspots and challenges and also look forward to the future directions of monocular human pose estimation. This article aims to provide researchers and practitioners with a comprehensive overview and promote the application and further research of monocular human pose estimation technology.

Keywords: Monocular human pose estimation; deep learning; technical approaches; industry applications

DOI: 10.48014/ccsr.20230614001

Citation: YAN Xin, GAO Hao, LI Haolun. A review of human pose estimation methods based on monocular vision [J]. Chinese Computer Sciences Review, 2023, 1(2): 13-30.