基于深度学习的单目人体姿态估计方法综述-I

原文：Monocular Human Pose Estimation: A Survey of Deep Learning-based Methods

摘要：

基于视觉的单目人体姿态估计是计算机视觉中最基本、最具挑战性的问题之一，其目的是从输入的图像或视频序列中获取人体的姿态。近年来，深度学习技术在人体姿态估计领域取得了重大进展和突破。本调查广泛回顾了自 2014 年以来发表的基于深度学习的 2D 和 3D 人体姿势估计方法。

一、Introduction

人体姿态估计(HPE)任务已经发展了几十年，其目标是从给定的传感器输入中获取人体的姿态。基于视觉的方法经常被用来通过使用摄像机来提供这样的解决方案。近年来，随着深度学习在图像分类、目标检测、语义分割等许多计算机版本任务中表现出良好的性能，人体姿态估计也通过采用深度学习技术获得了快速的发展。主要发展包括设计良好的网络，具有很强的估计能力，更丰富的数据集和更实际的身体模型探索。虽然已有一些关于人体姿态估计的综述，但是仍然缺乏一个综述来总结基于深度学习的最新成果。

本文综述了基于深度学习的 2D/三维人体姿态估计方法。依赖于其他传感器的算法，如深度、红外光源、射频信号和多视图输入不包括在本次调查中。

作为计算机视觉的基础任务之一，人体姿态估计是一个非常重要的研究领域，可以应用于许多应用领域，如动作/活动识别、动作检测、人体跟踪、电影和动画、虚拟现实、人机交互、视频监控、医疗救护、自动驾驶、运动分析等。

电影和动画：各种生动形象的数字人物的产生离不开对人类动作的捕捉。廉价准确的人体运动捕捉系统可以更好地促进数字娱乐产业的发展。

虚拟现实：虚拟现实是一种非常有前途的技术，可以应用于教育和娱乐。通过对人体姿态的估计，可以进一步明确人与虚拟现实世界的关系，增强交互体验。

人机交互：人体姿态估计对于计算机和机器人更好地理解人的身份、位置和行为是非常重要的。以人类的姿势（例如。手势），计算机和机器人可以以一种简单的方式执行指令，而且更加智能。

视频监控：视频监控是早期采用人体姿态估计技术对特定范围内的人进行跟踪、动作识别、再识别的应用之一。

医疗救助：在医疗救助的应用中，人体姿态估计可以为医生提供定量的人体运动信息，特别是康复训练和体能训练治疗。

自动驾驶：先进的自动驾驶技术发展迅速。有了人体姿态估计，自动驾驶汽车可以对行人做出更恰当的反应，并与交通协调员进行更全面的互动。

运动分析：通过对运动员在运动视频中的姿势进行估计，可以进一步得到运动员各项指标（如跑步距离、跳跃次数）的统计数据。在训练过程中，人体姿态估计可以提供动作细节的定量分析。

Fig. 1. Typical challenges of HPE in monocular images or videos. Example images are from Max Planck Institute for Informatics (MPII) dataset (Andriluka et al., 2014).

单目人体姿态估计具有一些独特的特点和挑战。如图 1 所示，人体姿态估计面临的挑战主要有三个方面：1.人类灵活的身体意味着关键点之间有着更复杂的内在关联和更高自由度的肢体动作，这对模型训练提出了更高的挑战；2.人体的着装意味着各式各样的身体外形；3.复杂的环境可能会导致前景信息难以提取(隐藏在背景中的人)，或者是进行多人检测时，不同个体间的相互遮挡会导致检测难度激增；同样地，相机的拍摄位置和角度，都会增加单目估计的难度。

人体姿态估计的文献可以分为不同的类型。根据是否使用设计的人体模型，可以将这些方法分为生成方法（基于模型）和判别方法（无模型）。根据从哪个级别（高级抽象或低级像素）开始处理，它们可以分为自上而下（top-down）的方法和自下而上（bottom-up）的方法。

二、人体姿态估计方法与人体模型的分类

2.1. 人体姿态估计方法分类

本节根据不同的特点总结了基于深度学习的人体姿态估计方法的不同分类：1）生成方法（基于人体模型）和判别方法（无人体模型）；2）自上而下（从高级抽象到低级像素证据）和自下而上（从低级像素证据到高级抽象）；3）基于回归（从输入图像直接映射到身体关节位置）和基于检测（生成关节位置的中间图像块或热图）；4)单阶段（端到端培训）和多阶段（分阶段培训）。

(1)生成方法 V.S. 判别方法

生成方法和判别方法之间的主要区别是方法是否使用人体模型。根据人体模型的不同表示，可以以不同的方式处理生成方法，例如关于人体模型结构的先验知识，从不同视图到 2D 或 3D 空间的几何投影，高维参数化空间回归方式的优化。

判别方法直接学习从输入源到人体姿势空间的映射（基于学习）或搜索不存在的示例（基于示例），而无需使用人体模型。判别方法通常比生成方法要快，但对于从未受过训练的姿势而言，判别方法的鲁棒性较差。

(2)自顶向下 V.S.自底向上：

对于多人姿态估计，人体姿态估计方法通常可以根据预测的出发点分为自顶向下和自底向上两种：高层抽象或低层像素。自顶向下的方法从高层抽象开始，首先检测人并在边界框中生成人的位置。然后对每个人进行姿态估计。相反，自底向上的方法首先预测输入图像中

每个人的所有身体部位，然后通过人体模型拟合或其他算法对它们进行分组。请注意，根据不同的方法，身体部位可以是关节、四肢或小模板贴片。随着图像中人数的增加，自顶向下方法的计算量显著增加，而自底向上方法的计算量保持稳定。然而，如果有一些人有一个很大的重叠，自下而上的方法面临的挑战，分组相应的身体部位。

(3) 基于回归 V.S. 基于检测

基于不同问题的表述，基于深度学习的人体姿态估计方法可以分为基于回归的方法和基于检测的方法。基于回归的方法直接将输入图像映射到人体关节坐标或人体模型参数。基于检测的方法将人体各部位作为检测目标，基于两种常用的表示方法：图像块和关节位置热图。从图像到关节坐标的直接映射是一个非常困难的问题，因为它是一个高度非线性的问题，而小区域表示提供了具有更强鲁棒性的密集像素信息。与原始图像尺寸相比，小区域表示的检测结果限制了最终关节坐标的精度。

(4) 单阶段 V.S. 多阶段

基于深度学习的一阶段方法旨在通过使用端到端网络将输入图像映射到人体姿势，而多阶段方法通常在多个阶段中预测人体姿势，并伴有中间监督。例如，一些多人姿势估计方法首先检测人的位置，然后为每个检测到的人估计人的姿势。其他 3D 人姿势估计方法则首先在 2D 平面中预测关节位置，然后将其扩展到 3D 空间。单阶段方法的训练比多阶段方法更容易，但中间约束更少。

2.2 人体模型

人体模型是人体姿态估计的关键组成部分。人体是一个柔性的、复杂的非刚性物体，具有运动结构、体形、表面纹理、各部位或各关节的位置等特性。

一个成熟的人体模型不一定要包含所有的人体属性，而应该满足特定任务的要求来建立和描述人体姿势。基于不同层次的表示和应用场景，如图 2 所示，人体姿态估计中有三种常用的人体模型：基于骨架的模型、基于轮廓的模型和基于体积的模型。

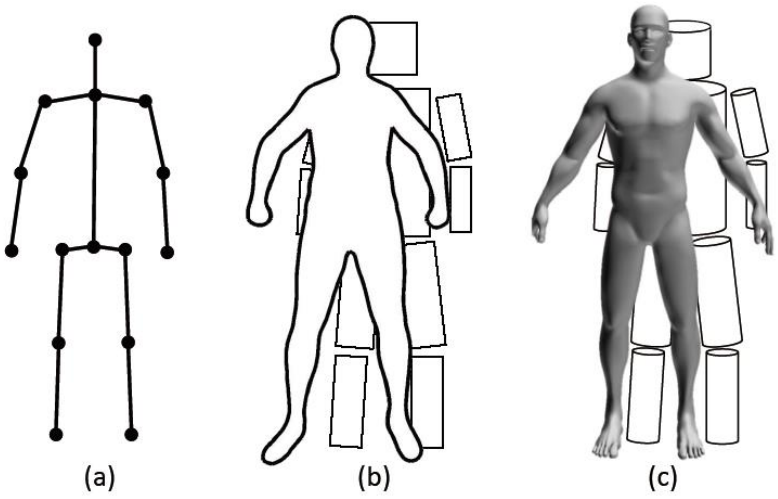

Fig. 2. Commonly used human body models. (a) skeleton-based model; (b) contour-based models; (c) volume-based models.

(1) 基于骨架的模型：基于骨架的模型，也称为棍状模型或运动学模型，表示一组关节（通常在 10 到 30 之间）的位置以及人体骨架结构之后相应的肢体方向。基于骨架的模型也可以描述为一个图，其中顶点指示骨骼结构中关节的约束和边缘编码约束或关节的先验连接。这种人体拓扑结构非常简单和灵活，广泛用于 2D 和三维人体姿态估计。它具有表示简单、灵活的优点，但也存在着缺乏纹理信息，即没有人体的宽度和轮廓信息等缺点。

(2) 基于轮廓的模型：基于轮廓的模型广泛应用于早期的人体姿态估计方法中，它包含了人体肢体和躯干的大致宽度和轮廓信息。人体各部分用人体轮廓的矩形或边界近似表示。广泛使用的基于轮廓的模型包括纸板模型和活动形状模型。

(3) 基于体积的模型：三维人体形状和姿势通常由具有几何形状或网格的基于体积的模型来表示。早期用于建模身体部位的几何形状包括圆柱、圆锥等。基于体积的现代模型以网格形式表示，通常通过 3D 扫描捕获。广泛使用的基于体积的模型包括人的形状完成和动画，蒙皮多人线性模型和统一的变形模型。

三、三维人体姿态估计

Table 5. Comparison of 3D single person pose estimation methods. Here E. stands for Extra data and T. indicates Temporal info. The last column is the Mean Per Joint Position Error (MPJPE) in millimeter on Human3.6M dataset under protocol #1. The results with ∗ were reported from 6 actions in testing set, while others from all 17 actions. The results with † were reported with 2D joint ground truth. The methods with # report joint rotation as well.

Methods	Backbone	E.	T.	Highlights	MPJPE (mm)
Model-free					
(Li and Chan, 2014)	shallow CNNs	✗	✗	A multi-task network to predict of body part detection with sliding windows and 3D pose estimation jointly	132.2*
(Li et al., 2015b)	shallow CNNs	✗	✗	Compute matching score of image-pose pairs	120.2*
(Tekin et al., 2016)	auto-encoder+ shallow CNNs	✗	✗	Employ an auto-encoder to learn a high-dimensional representation of 3D pose; use a shallow CNNs network to learn the high-dimensional pose representation	116.8*
(Tekin et al., 2017)	Hourglass	✓	✗	Predict 2D heatmaps for joints first; then use a trainable fusion architecture to combine 2D heatmaps and extracted features; 2D module is pre-trained with MPII	69.7
(Chen and Ramanan, 2017)	CPM	✓	✗	Estimate 2D poses from images first; then estimate depth of them by matching to a library of 3D poses; 2D module is pre-trained with MPII	82.7 /57.5†
(Moreno-Noguer, 2017)	CPM	✓	✗	Use Euclidean Distance Matrices (EDMs) to encoding pairwise distances of 2D and 3D body joints; train a network to learn 2D-to-3D EDM regression; jointly trained with other 3D (Humaneva-I) dataset	87.3
(Pavlakos et al., 2017)	Hourglass	✓	✗	Volumetric representation for 3D human pose; a coarse-to-fine prediction scheme; 2D module is pre-trained with MPII	71.9
(Zhou et al., 2017)	Hourglass	✓	✗	A proposed loss induced from a geometric constraint for 2D data; bone-length constraints; jointly trained with 2D (MPII) dataset	64.9
(Martinez et al., 2017)	Hourglass	✓	✗	Directly map predicted 2D poses to 3D poses with two linear layers; 2D module is pre-trained with MPII; process in real-time	62.9 /45.5†
(Sun et al., 2017)#	ResNet	✓	✗	A bone-based representation involving body structure information to enhance robustness; bone-length constraints; jointly trained with 2D (MPII) dataset	48.3
(Yang et al., 2018)	Hourglass	✓	✗	Adversarial learning for domain adaptation of 2D/3D datasets; adopted generator from (Zhou et al., 2017); multi-source discriminator with image, pairwise geometric structure and joint location; jointly trained with 2D (MPII) dataset	58.6
Pavlakos et al. (2018a)	Hourglass	✓	✗	Volumetric representation for 3D human pose; additional ordinal depths annotations for human joints; jointly trained with 2D (MPII) and 3D (Humaneva-I) datasets	56.2
(Sun et al., 2018)	Mask R-CNN	✓	✗	Volumetric representation for 3D human pose; integral operation unifies the heat map representation and joint regression; jointly trained with 2D (MPII) dataset	40.6
(Li and Lee, 2019)	Hourglass	✓	✗	Multiple hypotheses of 3D poses are generated from 2D poses; the best one is chosen by 2D reprojections; 2D module is pre-trained with MPII	52.7
Model-based					
(Bogo et al., 2016)#	DeepCut	✗	✗	SMPL model; fit SMPL model to 2D joints by minimizing the distance between 2D joints and projected 3D model joints	82.3
(Zhou et al., 2016)#	ResNet	✗	✗	kinematic model; embedded a kinematic object model into network for general articulated object pose estimation; orientation and rotational constrains	107.3
(Mehta et al., 2017c)#	ResNet	✓	✓	A real-time pipeline with temporal smooth filter and model-based kinematic skeleton fitting; 2D module is pre-trained with MPII and LSP; process in real-time; provide body height	80.5
(Tan et al., 2017)	shallow CNNs	✗	✗	SMPL model; first train a decoder to predict a 2D body silhouette from parameters of SMPL; then train a encoder-decoder network with images and corresponding silhouettes; the trained encoder can predict parameters of SMPL from images	-
(Mehta et al., 2017a)	Resnet	✓	✗	Kinematic model; transfer learning from features learned for 2D pose estimation; 2D pose prediction as auxiliary task; predict relative joint locations following the kinematic tree body model; jointly trained with 2D (MPII and LSP) datasets	74.1
(Nie et al., 2017)	RMPE + LSTM	✓	✗	Kinematic model; joint depth estimation from global 2D pose with skeleton-LSTM and local body parts with patch-LSTM; 2D module is pre-trained with MPII	79.5
(Kanazawa et al., 2018)#	ResNet	✓	✗	SMPL model; adversarial learning for domain adaptation of 2D images and 3D human body model; propose a framework to learn parameters of SMPL; jointly trained with 2D (LSP, MPII and COCO) datasets; process in real-time	88.0
(Pavlakos et al., 2018b)#	Hourglass	✓	✗	SMPL model; first predict 2D heatmaps of joint and human silhouette; second generate parameters of SMPL; 2D module is trained with MPII and LSP	75.9
(Omran et al., 2018)#	RefineNet	✗	✗	SMPL model; first predict 2D body parts segmentation from the RGB image; second take this segmentation to predict the parameters of SMPL	59.9
(Varol et al., 2018)	Hourglass	✓	✗	SMPL model; first predict 2D pose and 2D body parts segmentation; second predict 3D pose; finally predict volumetric shape to fit SMPL model; 2D modules are trained with MPII and SURREAL	49.0
(Arnab et al., 2019)#	ResNet	✓	✓	SMPL model; 2D keypoints, SMPL and camera parameters estimation; off-line bundle adjustment with temporal constraints; 2D module is trained with COCO	77.8 /63.3†
(Tome et al., 2017)	CPM	✓	✗	Pre-trained probabilistic 3D pose model; 3D lifting and projection by probabilistic model within the CPM-like network; 2D module is pre-trained with MPII; process in real-time	88.4
(Rhodin et al., 2018a)	Hourglass	✗	✗	A latent variable body model learned from multi-view images; an encoder-decoder to predict a novel view image from a given one; the pre-trained encoder with additional shallow layers to predict 3D poses from images	-

三维人体姿态估计是从图像或其他输入源中预测人体关节在三维空间中的位置。尽管商业产品，如带有深度传感器的 Kinect、带有光学传感器的 VICON 和带有多个摄像头的 The Captury 已被用于 3D 身体姿势估计，所有这些系统都在非常有限的环境中工作，或者需要

人体上的特殊标记。单目摄像机作为应用最为广泛的传感器，对三维人体姿态估计具有重要意义。深度神经网络具有从单目图像估计密集深度和稀疏深度点（关节）的能力。此外，基于单目输入的三维人体姿态估计的进展可以进一步改善约束环境下的多视点三维人体姿态估计。因此，本节重点介绍基于深度学习的方法，这些方法从单目 RGB 图像和视频中估计 3D 人体姿势，包括 3D 单人姿势估计和 3D 多人姿势估计。

3.1. 三维单人姿态估计

与二维人体姿态估计相比，3D-人体姿态估计更具挑战性，因为它需要预测人体关节的深度信息。另外，3D-人体姿态估计的训练数据也不像 2D-人体姿态估计那样容易获得。现有的数据集大多是在有限的可推广的受限环境下获得的。对于单人姿势估计，通常提供图像中的人的边界框，因此不需要结合人检测过程。在本节中，我们将三维单人姿势估计方法分为无模型和基于模型两类。

3.1.1. 无模型方法

无模型方法不使用人体模型作为预测目标或中间线索。它们大致可分为两类：1）直接将图像映射到三维姿态；2）根据二维姿态估计方法得到的中间预测的二维姿态估计深度。

直接从图像特征估计三维姿态的方法通常包含很少的约束。Li 和 Chan 采用浅层网络直接回归三维关节坐标，并使用滑动窗口进行身体部位检测的同步任务。Pavlakos 等人用人体关节的额外顺序深度作为约束来训练网络，通过这些约束，2D 人体数据集也可以输入顺序深度注释。Li 等人设计了一种嵌入子网络学习潜在姿势结构信息来指导三维关节坐标映射。该子网络可以为输入图像姿势对分配匹配分数，并具有最大的边际代价函数。Tekin 等人预先训练了一个无监督的自动编码器来学习 3D 姿势的高维潜在姿势表示，以添加关于人体的隐式约束，然后使用浅层网络来学习高维姿势表示。Sun 等人提出了一种结构感知回归方法。他们设计了一种基于骨骼的表示方法，它包含了身体结构信息，比仅仅使用关节位置更稳定。

3.1.2. 基于模型的方法

基于模型的方法通常采用一个参数化的人体模型或模板来从图像中估计人体的姿势和形状。本文不包括早期的几何模型。最近的模型是通过对不同人群的多次扫描或不同身体模型的组合来估计的。这些模型通常由单独的身体姿势和形状组件进行参数化。一些工作采用了 SMPL 的身体模型，并试图从图像中估计 3D 参数。例如，Bogo 等人将 SMPL 模型拟合到估计的 2D 关节，并提出了一种基于优化的方法从 2D 关节恢复 SMPL 参数。Tan 等人通过首先训练解码器以使用合成数据从 SMPL 参数预测轮廓，然后使用训练的解码器学习图像编码器来推断 SMPL 参数。训练后的编码器可以根据输入图像预测 SMPL 参数。

3.2 三维多人姿态估计

单目三维多人姿态估计是在三维单人姿态估计等深度学习方法的基础上发展起来的。这一研究领域比较新，提出的方法也不多。

Mehta 等人提出了一种自下而上的方法，通过使用 2D 姿势和部分相似性字段来推断人物实例。提出了一种遮挡鲁棒姿态映射(ORPM)算法，该算法可以在不受人数影响的情况下提供多种类型的遮挡信息。Rogez 等人提出了一个局部分类回归网络(LCR-Net)，经过三个阶段的处理。首先，采用更快的 R-CNN 来检测人的位置。第二，每个姿势候选被分配一个分类器评分的最近的姿势。最后的姿态分别用一个回归器进行细化。Zanfir 等人提出了一个具有前馈和反馈阶段的框架，用于 3D multi 人体姿态估计和形状估计。前馈过程包括身体部位的语义分割和基于 DMHS 的 3D 姿势估计。然后，反馈过程细化 SMPL 的姿势和形状参数。Mehta 等人通过三个阶段实时估计多个姿势。首先，SelecSLS 网络为可见的身体关节推断 2D 姿势和中间 3D 姿势编码。然后基于每个检测到的人，重建完整的三维姿态，包括遮挡的关节。最后，对时间稳定性和运动骨架拟合作出了改进。

发布链接：https://mp.weixin.qq.com/s/8DRMEzGFepS--WMtRtNQ1A