摘要

人体形态感知旨在从输入图像中推断细粒度的人体姿态与形状信息，为后续视觉分析任务提供结构化的人体表示，是计算机视觉领域的重要研究问题、支撑智能服务应用的基础技术。大量文献表明，人体形态感知不仅在人机交互、数字娱乐以及体育运动等诸多领域有着广阔的应用前景，而且在计算机视觉领域中也能为可控人体渲染、动作识别和行人表征学习等众多与人有关的基础分析任务提供有效的指导信息，具有重要的应用价值与研究意义。

受益于深度学习与大规模神经网络的流行，人体形态感知领域有了长足发展。然而，在复杂场景下现有方法仍难以高效且鲁棒地估计人体形态，不能满足实际应用的需求。复杂场景具有密集遮挡严重、感知目标众多以及场景类型多变等特点，给现有人体形态感知方法带来了不同层次的挑战。首先，复杂场景包含的密集重叠人体与遮挡给多人感知处理带来了难题，导致方法出现漏检与准确率下降。其次，在感知众多目标时，现有模型不能兼顾感知精度与效率的要求，难以高效感知人体形态。最后，复杂多变的场景对现有方法的泛化能力提出了要求，而传统人体形态感知方法受限于训练数据，难以泛化到感知不同输入场景以及新类型的人体结构。为此本论文以复杂场景人体形态感知任务为研究重点，从遮挡鲁棒性、感知高效性以及跨场景泛化性三个方面出发设计全新且高效的方法，并展示了人体形态感知在典型场景下的验证。论文的主要贡献包含以下几个方面：

针对复杂场景中密集遮挡引起的多人感知漏检问题，本文提出了全新的鲁棒多人处理框架，用于准确分离密集重叠行人。本文首先提出基于直接姿态推理的多人姿态估计框架，来避免人体姿态与关键点在多人密集场景下的漏检与不准确定位。不同于之前定位单个关键点来组合成人体姿态的方法，所提出的框架将人体姿态视为一个基本的估计单元进行推理，以此避免人体姿态的漏检。更进一步本文提出了适用于多种人体形态感知任务的上下文实例解耦框架。该框架能够从特征层高效地分离不同人体，同时保留充足的上下文信息对每个人进行人体形态感知，具有对检测误差鲁棒、推理效率高、充分利用图像上下文信息等优点。上述多人处理框架在多个复杂密集场景人体形态感知任务中展现出性能优势，在人体姿态估计、前景分割以及部件分割任务上分别比之前最佳方法提升了 5.6%、11.1% 和 3.5% 的准确率。

针对现有模型难以兼顾感知精度与效率的难题，本文提出了新颖的高效预测模型，用于准确且快速地感知人体姿态与形状。本文首先提出了空间感知回归模型来定位人体形态中的关键点信息。该模型将输入图像中的空间位置信息引入到回归过程中来降

低直接回归的难度，从而避免了已有模型存在的准确率低下、计算存储高与量化误差等弊端。引入的空间位置先验也能辅助模型感知不同实例的同类型关键点，扩大了回归模型的适用范围。其次，本文还提出顺序全局旋转估计模型来估计人体形态中的另一类人体网格信息，以此缓解已有相对旋转估计模型存在的误差累积问题。所提出的方法直接估计每个关节点的全局旋转矩阵，无需通过人体运动学链连乘相对旋转矩阵来得到最后的全局旋转，因而避免了误差累积问题。上述形态预测模型能高效感知人体形态，在多个人体关键点定位与网格重建数据集上取得领先性能，如在维持相似计算量的情况下提升了 17.5% 的关键点定位准确率与降低了 5.0mm 的网格重建误差，取得了准确率与效率的平衡。

针对传统感知方法难以处理场景类型多变的难题，本文提出了新颖的跨场景可泛化人体形态感知方法，用于准确感知新类型场景下的人体形态。不同于已有方法从训练数据中学习人体结构先验，本文引入文本来编码人体结构信息，同时利用在海量数据上预训练好的大语言模型来理解文本并进行人体形态感知。该方法以图像和文本指令作为输入并利用大语言模型优异的推理能力与文本指令中的关键点类型、位置以及关系信息来输出目标人体关键点坐标。实验结果表明，本文所提出的多模态定位大模型在多个人体姿态估计任务上取得了优异的性能。同时得益于引入的文本先验与大语言模型，定位大模型也展示出优异的泛化能力，在跨场景数据测试和新类型人体关键点定位中均取得优异性能，领先当前方法 11.0% 与 24.1% 的准确率。

基于上述三方面工作，本文形成了较为完备且高效的人体形态感知方法，能有效处理复杂场景下的感知任务。围绕这些方法，本文还展示了人体形态感知在典型场景下的验证，分别构建了自动驾驶下的人体感知数据集与复杂场景人体视觉分析系统。上述方法、数据集与系统的研究有效提升了复杂场景人体形态感知的鲁棒性、高效性以及泛化性，有望推进人体形态感知相关方法的实际落地应用，对实现视觉智能提供有效的支持。

关键词：人体形态感知、复杂场景、鲁棒性、高效性、泛化性

Human Pose and Shape Perception in Complex Scenes

Dongkai Wang (Computer Applied Technology)
Directed by: Prof. Shiliang Zhang

ABSTRACT

Human Pose and Shape Perception (HPSP) aims to infer the fine-grained human structure information from monocular RGB images, *e.g.*, human pose and shape, which can provide structured representation of human body for subsequent visual analysis tasks. HPSP is one of the key problems in computer vision and a basic technology supporting intelligent service applications. A large amount of literature shows that HPSP not only has broad application prospects in many fields such as human-computer interaction, digital entertainment, and sports, but also can be used for controllable human body rendering, behavior recognition, and pedestrian reconstruction in many tasks in the field of computer vision. It provides effective guidance information for many basic tasks related to people, and has important application value and scientific research significance.

Benefited by the development of deep learning and large-scale neural networks, the field of HPSP has made great progress. However, existing methods are still difficult to estimate human body shape efficiently and robustly in complex scenes, unable to meet the needs of practical applications. HPSP in complex scenes contains three major challenges, including severe occlusion, numerous targets and diverse scenes. Firstly, complex scenes contain highly overlapping human bodies and occlusions, making it hard to separate overlapped persons and resulting in missed or wrong detection of human pose and shape. Secondly, when processing numerous targets, existing models cannot balance perception accuracy and efficiency, making it difficult to efficiently estimate corresponding human pose and shape. Finally, diverse scenes bring challenges to the generalization capability of existing methods. Traditional HPSP method is limited by the training data and is difficult to generalize to detect input images from different scenes and novel types of human body structures. This thesis focuses on HPSP in complex scenes and designs several novel and efficient methods from three aspects: occlusion robustness, perception efficiency and cross scene generalization, and validates the performance of HPSP in typical scenes. The main contributions of this thesis include the following aspects:

In terms of missed and wrong detection of HPSP methods in complex scenes, this thesis

proposes several novel and robust multi-person processing frameworks to separate overlapped instances. This thesis firstly proposes a Direct Pose-Level Inference (DPLI) framework to avoid missed and inaccurate localization of human poses and keypoints in complex scenes. Unlike previous methods that infer human pose from individual keypoint, the proposed method considers the human pose as an inference objective, thus avoid the missed detection of keypoints. Furthermore, this thesis proposes a more general Contextual Instance Decoupling (CID) framework to separate different persons in crowded scenes. The proposed CID efficiently separates overlapped persons in feature map, while retaining sufficient contextual information to infer dense body pose and shape for each instance, therefore achieving superior performance in various instance-level human analysis tasks. For example, the proposed method outperforms previous methods by 5.6%, 11.1%, and 3.5% on the human pose estimation dataset CrowdPose, the human foreground segmentation dataset OCHuman, and the human part segmentation dataset CIHP, respectively.

In terms of low efficient HPSP models, this thesis proposes several efficient HPSP models to estimate human pose and shape accurately and efficiently. This thesis firstly proposes a Spatial-Aware Regression (SAR) model for human keypoint localization, to address the low accuracy, high computational storage and quantization error of previous perception model. SAR introduces the spatial location prior in the input image into the regression process to ease the difficulty of direct regression. The introduced spatial location prior can also benefit model to locate multi-person keypoints, making SAR a more general keypoint localization model. For rotation matrix estimation, this thesis proposes a Sequentially Global Rotation Estimation (SGRE) to directly estimate the global rotation of each joint to avoid error accumulation of previous relative rotation estimation model and pursue better accuracy. The proposed SGRE alleviates error accumulation and produces more accurate 3D human mesh. The above model can effectively and efficiently estimate human pose and shape, achieving superior performance on various human pose and shape perception benchmarks. For example, it improves the localization accuracy of regression model on COCO Keypoint dataset by 17.5% and reduce the reconstruction error on 3DPW dataset by 5.0 mm, while maintaining the same computational cost.

In terms of limited generalization capability of traditional HPSP methods, this thesis proposes a novel text-guided HPSP method to handle diverse scenes. Different from existing methods that learn human body structure priors from training data, the proposed method introduces text to encode human body keypoint information, and utilizes the large language models

that pretrained on large amount of data to understand textual input and estimate corresponding human pose and shape. The proposed method takes both image and text instruction as input and utilizes the powerful reasoning capability of large language models to understand the keypoint type, location and relationship to other keypoints to output the desired human keypoint coordinates. Experimental results show that the proposed method obtains superior performance on both 2D/3D human pose estimation tasks. Benefited by the introduced textual description and large language model, our method shows superior generalization capability in cross dataset validation and novel type keypoint localization, outperforming previous methods by 11.0% and 24.1%, respectively.

Relying on the efforts in the above three directions, this thesis proposes several unified and effective algorithms for HPSP in complex scenes. Besides on above methods, this thesis also verifies the effectiveness of HPSP in typical scenes, constructing a human perception dataset for autonomous driving and a human visual analysis system for complex scenes, respectively. The research on the above methods, datasets and systems has effectively improved the accuracy, robustness and generalization capability of HPSP method in complex scenes, which is expected to promote the practical application of HPSP and provide support for visual intelligence.

KEY WORDS: Human Pose and Shape Perception, Complex Scenes, Robustness, Efficiency, Generalization