Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 2, 2023

3D Human Pose Estimation: A Survey

Shan Jia *

Department: Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
*** Corresponding author Email:** nningfvcg15@gmail.com

Abstract: This comprehensive review article explores the latest research advancements in the realm of estimating 3D human pose. Traditional methods such as PSM, SVM are discussed. Besides, this review also talks about deep learning-based approaches, including direct approaches, 2D-to-3D lifting and volumetric model approach for single person, top-down approaches and bottom-up approaches for multi-person pose estimation. The analysis covers the strengths and challenges of various methods, encompassing issues such as model generalization, occlusion robustness, and computational efficiency. Current research issues are identified, and future directions are proposed. By summarizing and evaluating existing methods, this paper aims to provide valuable insights for researchers in both academia and industry, driving the evolution of 3D human pose estimation for better practical applications.

Keywords: 3d Human Pose Estimation; Deep Learning; Neural Networks.

1. Introduction

In the field of computer vision, human pose estimation (HPE) is a significant and highly regarded research task. This process entails precise determination of human body pose and keypoint coordinate using image or video data. This data offers crucial insights into geometry and movement, applicable across various domains including behavior recognition, interactive technology, as well as healthcare in both augmented and virtual reality contexts.

With the advancement of deep learning, notable advancements have been realized in the field of HPE, especially the successful implementation of Convolutional Neural Networks (CNNs) within the realm of computer vision. Traditional methods have excelled in achieving impressive performance in 2D HPE, but challenges persist when dealing with 3D HPE. Recovering accurate 3D pose information from 2D images is complex and challenging as the absence of depth data and the effects of viewpoint variations.

Presently, approaches for3D HPE can be broadly classified into two categories: traditional methodologies and those based on deep learning techniques. Traditional methods depend on handcrafted features and models, achieving certain results in some cases, but they perform poorly in the face of complexity, occlusion, incomplete information, and data scarcity. Conversely, deep learning-based methods, through end-to-end training and automatic feature learning, have overcome the limitations of traditional methods and achieved superior performance in many visual tasks.

However, despite the impressive performance of deep learning methods in 3D HPE, challenges and issues persist. First, deep learning methods heavily rely on a substantial volume of annotated 3D pose information, and obtaining high-quality 3D pose annotations remains a costly and time-consuming task. Second, due to the complexity and parameterization of deep learning models, they demand substantial computational resources, making it challenging, especially for real-time scenarios with stringent requirements.

Therefore, this review paper aims to investigate, summarize, and analyze the latest research advancements in the field of 3D HPE. Through comparing and evaluating traditional methods and deep learning-based approaches, we will pinpoint the current issues and challenges in the research, such as insufficient depth information, viewpoint variations, occlusions, and data scarcity. The goal of this paper is to clarify these problems, guide future research directions, and explore how to address these challenges to enhance the accuracy, robustness, and real-time capabilities of 3D HPE. Through the summary and review of existing methods, we hope to provide valuable insights for researchers in academia and industry, driving the development of this field for better application in real-world scenarios.

2. 3d Human Pose Estimation

3D HPE has broad applications, such as rehabilitation training and can provide skeleton information for other computer vision tasks, like behavior recognition.

Human pose can be represented in two main ways: (1) Skeletal representation: This method constructs human poses as a series of keypoints connected by lines, describing the relationships and connections between keypoints. Skeletal representation is commonly used in applications like motion analysis and human-computer interaction. (2) Parametric human models: An example is SMPL (Skinned Multi-Person Linear Model), which represents human pose and body shape in the form of a mesh. This approach is suitable for detailed human modeling and motion reconstruction. The objective of 3D HPE involves the prediction of three-dimensional pose attributes, encompassing either the coordinates (x, y, z) or Euler angles, pertaining to each joint. This prediction is derived from RGB images (or RGB images coupled with 2D keypoints). It uses an individual's image to determine the XYZ coordinates of the human keypoints.

The process of 3D HPE is roughly as follows:

(1) Keypoint extraction: Detect 2D keypoints from images or videos, representing body parts such as the head, shoulders, elbows, waist, knees, etc. (2)3D HPE: Based on the extracted 2D keypoints, algorithms are used to estimate the three-dimensional coordinates (x, y, z) for each keypoint, determining the pose of the human in 3D space.

The challenge in 3D HPE lies in recovering 3D human pose from 2D images, as there is more uncertainty in the estimation process, and it requires greater computational and spatial

resources compared to 2D images. Additionally, dataset selection is a crucial challenge, as algorithms must exhibit invariance to a multitude of factors, including texture, human skin color, background interference, occlusion, etc. The following review covers both conventional and recent deep learning methods employed in 3D HPE.

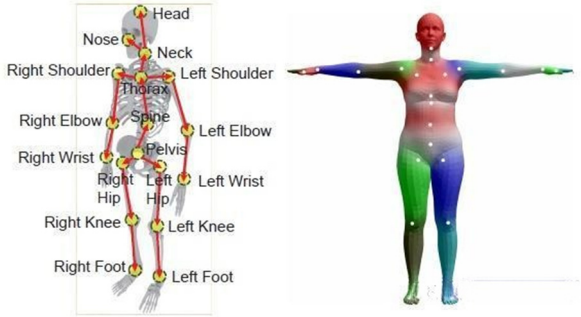

Figure 1. Skeletal form and Mesh form.

2.1. Classical Approaches

Pictorial Structure Models (PSM): Pictorial Structure Models (PSM) have been a prominent approach in 2D HPE. This model introduced a multi-view pictorial structure, extending the progress of 2D HPE. PSM is applicable to single-person HPE and extends to estimating the poses of multiple individuals. Essentially, it operates as a generative model for HPE. However, PSM suffers from lower accuracy due to certain conditional constraints among output factors.

Structured SVM: To address the conditional constraints of PSM, Structured Support Vector Machine (Structured SVM) methods were introduced. These methods are employed to get the transformation from segmentation features to keypoint positions, aiming to enhance the accuracy of HPE.

3D HPE with HOG Features: 3D HPE utilizing Histogram of Oriented Gradients (HOG) features. HOG features are employed to describe object shapes and have demonstrated stable application in 3D human pose analysis. By registering HOG features across the entire image, efficient measurement of HOG features is achieved. Subsequently, Principal Component Analysis (PCA) is employed on individual HOG blocks, thus assessing 3D human pose through linear regression based on HOG features.

In summary, these classical methods have achieved notable progress in the field of HPE. Among them, the HOG feature-based approaches performed well during the COCO 2016 keypoint challenge, surpassing earlier state-of-the-art methodologies. Satisfactory results were also obtained in the MPII MultiPerson benchmark tests.

2.2. Learning based Approaches

2.2.1. Single-person 3D HPE

Direct Estimation: This approach directly estimates 3D human pose from 2D images without intermediate steps for 2D HPE. For instance, Sun et al. utilize a bone-based representation, which offers some stability advantages by encoding long-range interactions between bones through a compositional loss. [1] Pavlakos and colleagues propose volumetric representations, which facilitate the conversion of the intricate 3D coordinate regression challenge into a tractable format within discrete space. Their approach involves forecasting the probability of individual joints within the volumetric context, utilizing convolutional networks. A challenge for this method is its potential reliance on the performance of 2D pose detectors. [2, 3]

2D to 3D Lifting: This method infers 3D human pose based on 2D HPE (2D HPE) and maps 2D pose to 3D pose. Generally, 2D-to-3D lifting methods exhibit good performance due to the strong performance of current 2D pose detectors. Martinez et al. use fully connected residual networks to regress 3D joint positions from 2D joint locations. However, this approach may be affected by the reconstruction ambiguities stemming from over-reliance on the 2D pose detector.[4] Other methods use 2D heatmaps rather than 2D pose as an intermediate representation for 3D HPE. Some approaches employ convolutional neural networks to predict the depth ordering of human joints, then regress 3D pose from 2D joints and the depth ordering matrix using a coarse-to-fine pose estimator. Additionally, the method by Li and Lee generates multiple 3D pose hypotheses and applies a sorting network to select the best 3D pose. The advantage of this approach is enhancing 3D HPE through intermediate steps of 2D HPE.[5]

Volumetric Models: This approach employs volumetric representations to model human pose in 3D HPE. It partitions 3D space into voxel grids, with each voxel storing information or probabilities of a certain joint. This can be seen as a 3D image classification problem where the goal is to determine the presence of each joint in every voxel. One advantage of this method is its ability to precisely model pose in 3D space, but it requires larger computational and memory resources. In the Volumetric models approach, some methods use Convolutional Neural Networks (CNNs) to predict the likelihood of each voxel, often employing 3D convolutions or variations. Such methods retain the inter-joint relationships in 3D space, leading to better capture of pose structure.

An example of this approach is the work by C. Qi et al., which employs CNNs to predict the joint presence probabilities for each voxel in the voxel grid, transforming the 3D HPE problem into a voxel-level classification task.[6] By processing the predicted results for these voxels, estimates for 3D human pose can be obtained. It's essential to note that while Volumetric models approach offers accurate 3D HPE, the processing of 3D space information incurs relatively higher computational and memory costs, especially when dealing with high-resolution data. Hence, when selecting a HPE method, it's crucial to weigh the pros and cons of different approaches based on application requirements and available computational resources.

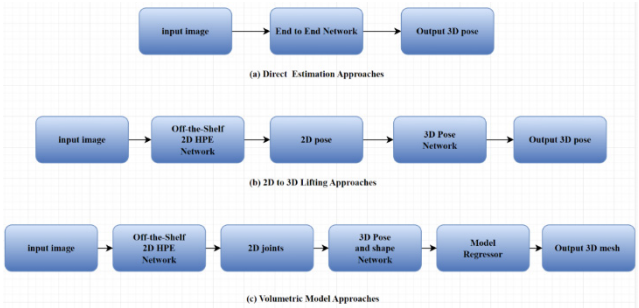

Figure 2. Single-person 3D HPE frameworks

2.2.2. Multi-person 3D HPE

Top-down approaches first perform human detection to locate individuals in the image. Zanfir et al. proposed an enhanced 3D multi-person human pose module by adding semantic segmentation information and scene constraints.[7] For video-based multi-person 3D HPE, the challenge of 3D temporal association needs to be addressed. The advantages of top-down methods include leveraging advanced human

detection and single-person HPE methods, but in scenarios featuring an extensive assembly of individuals, particularly within congested settings, the computational intricacy and inference duration could potentially escalate significantly. Additionally, top-down methods often first detect bounding boxes for each person, potentially leading to the neglect of global scene information, depth estimation inconsistency in cropped regions, and overlapping predicted poses. Li et al. addressed the issue of insufficient global context within top-down methods by introducing a hierarchical multi-person ordering scheme that utilizes body hierarchy semantics and global consistency to encode interaction information in a hierarchical manner, improving HPE accuracy, especially in scenes involving multi-person interactions.[8]

Bottom-up approaches is dissimilar to the top-down strategies, bottom-up methodologies exhibit a distinct approach by initiating with the generation of individual body joint positions and depth maps for all entities. Subsequently, the connection of body parts to each person is determined based on factors like root depth and relative part depths. One pivotal challenge encountered by bottom-up techniques pertains to the task of effectively grouping body joints corresponding to each individual. Addressing this, Zanfir et al. formalized this problem as a binary integer programming (BIP) challenge. The method's strengths reside in its linear computational nature and its minimal time complexity.[9]

However, when the goal centers around the reconstruction of 3D human meshes, bottom-up approaches adopt a less direct route. They necessitate supplementary model regression modules to rebuild human meshes rooted in the final 3D pose. In contrast, top-down methods offer a more straightforward route: once each individual is detected, the integration of 3D single-person human mesh recovery techniques seamlessly restores individual human meshes. In their study, Chen et al. ingeniously fused both top-down and bottom-up paradigms.[10] They initially employed a top-down network to gauge joint heatmaps within individual bounding boxes, subsequently applying a bottom-up network to consolidate these estimated heatmaps, effectively managing variations in scale.

Navigating occlusion presents another hurdle for bottom-up strategies. To mitigate this challenge, Mehta et al. introduced the concept of "Occlusion-Robust Pose-Maps (ORPM)" that infuses redundancy into position map formulations, thereby facilitating person association within heatmaps, particularly within occluded scenes.[11] Zhen et al., on the other hand, harnessed a depth-aware part association algorithm, factoring in occlusion and bone length constraints to assign joints to distinct individuals.[12]

In a different approach, Mehta et al. rapidly deduced intermediate 3D poses for visible body joints, regardless of their precision.[13] This was succeeded by inferring occluded joints based on acquired pose priors and overarching context, effectively accomplishing comprehensive 3D pose reconstruction. The final 3D pose is optimized by applying temporal coherence and adapting to motion skeleton models.

Comparative Analysis: Top-down techniques frequently depend on sophisticated human detection and single-person HPE methodologies, yielding commendable outcomes. Nevertheless, as the count of individuals escalates, notably within congested environments, the computational intricacy and inference duration could potentially escalate significantly. Additionally, since top-down methods primarily detect bounding boxes for each individual, they may neglect global information, leading to depth estimation inconsistency in cropped regions and overlapping predicted poses. Bottom-up methods offer the benefit of streamlined computation and time complexity. However, when the objective revolves around the reconstruction of 3D human meshes, they require additional model regression modules. Therefore, the choice of approach should be based on application requirements and computational resources.

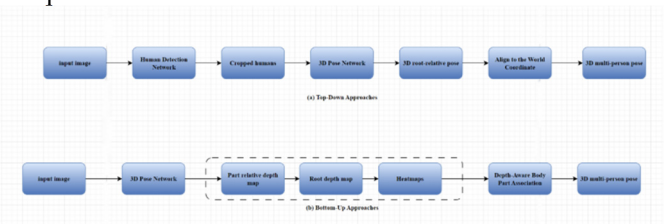

Figure 3. Schematic diagram of the multi-person 3D HPE framework

3. Conclusion

In recent years, significant progress has been made in 3D HPE. Many 3D HPE methods have benefited from the advancements in 2D HPE (2D HPE) due to the adoption of 2D-to-3D lifting strategies. Certain 2D HPE (HPE) techniques, namely OpenPose, AlphaPose, and HRNet, have gained widespread adoption as 2D pose detectors within the realm of 3D HPE methodologies. Furthermore, beyond their role in 3D HPE, specific approaches even exhibit the capability to restore 3D human meshes from images or videos. However, despite these advancements, several challenges still exist.

A notable challenge pertains to the model's capacity for generalization. The generation of precise 3D ground-truth pose annotations frequently hinges on motion capture systems, which can pose challenges in terms of implementation across a range of diverse real-world settings. Consequently, existing datasets are primarily collected in controlled scenarios. The most advanced techniques excel in these specific datasets. However, they encounter a decline in performance when confronted with in-the-wild data. The integration of synthetic datasets encompassing a spectrum of poses and intricate scenes, such as the SURREAL [14], or GTA-IM [15] datasets, derived from gaming engines, presents itself as a potential avenue. Nevertheless, deriving knowledge solely from synthetic data may not fully align with anticipated performance outcomes due to the divergence between the distributions of synthetic and authentic real-world data.

Similar to the challenges faced in 2D HPE (HPE), 3D HPE confronts the crucial hurdles of withstanding occlusions and maintaining computational efficiency. The existing 3D HPE methodologies encounter notable performance declines in scenarios marked by dense crowds, primarily due to pervasive mutual occlusions and potential limitations in image resolution for each individual. The computational demands of 3D HPE surpass those of its 2D counterpart. For instance, methods that transition from 2D to 3D pose rely on the 2D pose as an intermediary to deduce the 3D pose. Hence, the development of highly computationally efficient pipelines for 2D HPE stands as an imperative, all the while ensuring the retention of accurate HPE outcomes.

In conclusion, despite the significant progress in 3D HPE, challenges remain in aspects such as model generalization, occlusion robustness, and computational efficiency. Addressing these challenges will further drive the development of 3D HPE technology, enabling better

performance in broader scenarios and applications.

References

[1] Sun, X., Shang, J., Liang, S et al. (2017) Compositional human pose regression. In: 2017 IEEE International Conference on Computer Vision. Venice. pp. 2602-2611.

[2] Pavlakos, G., Zhou, X., Daniilidis, K. (2018) Ordinal Depth Supervision for 3D Human Pose Estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 7307-7316.

[3] Pavlakos, G., Zhou, X., Derpanis, K. G et al. (2017) Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu. pp. 7025-7034.

[4] Martinez, J., Hossain, R., Romero, J et al. (2017) A simple yet efective baseline for 3d human pose estimation. In: 2017 IEEE International Conference on Computer Vision. Venice. pp. 2640-2649.

[5] Li, C., Lee, G. H. (2019) Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach. pp. 9887-9895.

[6] Qi, C. R., Su, H., Nießner, M et al. (2016) Volumetric and Multi-view CNNs for Object Classification on 3D Data. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas. pp. 5648-5656.

[7] Zanfir, A., Marinoiu, E., Sminchisescu, C. (2018) Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes: The Importance of Multiple Scene Constraints. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 2148-2157.

[8] Wang, C., Li, J., Liu, W et al. (2020) Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In: 2020 European Conference on Computer Vision. Glasgow. pp. 242–259.

[9] Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A. I., & Sminchisescu, C. (2018) Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images. In: 32nd Conference on Neural Information Processing Systems. Montréal. pp. 8420–8429.

[10] Cheng, Y., Wang, B., Yang, B et al. (2021) Monocular 3D Multi-Person Pose Estimation by Integrating Top-Down and Bottom-Up Networks. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville. pp. 7649-7659.

[11] Mehta, D., Sotnychenko, O., Mueller, F et al. (2018) Single-Shot Multi-Person 3D Pose Estimation from Monocular RGB. In: 2018 International Conference on 3D Vision. Verona. pp. 120-130.

[12] Zhen, J., Fang, Q., Sun, J et al. (2020) SMAP: Single-Shot Multi-Person Absolute 3D Pose Estimation. In: European Conference on Computer Vision. Glasgow. pp. 550-566.

[13] Mehta, D., Sotnychenko, O., Mueller, F et al. (2020) XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera. ACM transactions on graphics, 39(4): 82-1.

[14] Varol, G., Romero, J., Martin, X et al. (2017) Learning from synthetic humans. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu. pp. 109-117.

[15] Cao, Z., Gao, H., Mangalam, K et al. (2020) Long-term human motion prediction with scene context. In: 16th European Conference on Computer Vision. Glasgow. pp. 387-404.