Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/109/20241407

A Review of 3D Human Pose Estimation

Lin Zhai

Northeastern University, CPS, Boston, Massachusetts, 002199, USA

franklin.zhai@outlook.com

Abstract. Visual perception and human body recognition are fundamental capabilities required for effective and safe interactions between artificial intelligence (AI), computer vision, and humans in real-world scenarios. Recent groundbreaking developments in AI and computer vision have resulted in major advancements in human body recognition technology. However, research in human body recognition is still in the early stages of the product lifecycle. Identifying the three-dimensional locations of the joints in the human body from pictures or videos is known as 3D posture estimation. Although it is widely used in areas like human motion analysis and robotics, it continues to be a difficult task due to challenges such as depth ambiguity and the scarcity of robust datasets. Over the past decade, numerous methods have been developed, many of which are based on deep learning, significantly improving the performance of existing benchmarks. A comprehensive literature review of this field is crucial for future development. However, in nowadays,more and more such research has mainly concentrated on traditional techniques, requirement for a comprehensive examination of tools based on deep learning. This paper delivers a thorough overview of current deep learning-based 3D pose estimation algorithms, outlining their advantages and limitations while providing a detailed understanding of the field. It also explores commonly used benchmark datasets and methods for analyzing human poses in unlabeled field images, providing a thorough comparative analysis. Finally, insights are provided to aid in the design of future models and algorithms.

Keywords: Visual perception, Body recognition, 3D human pose estimation, Benchmark datasets.

1. Introduction

The task of predicting the locations of human joints from a single image or a series of images is regularly defined as human pose estimation. In computer vision, 3D human pose estimation has become an important field of study, owing to its potential for numerous applications and its influence across various domains. With the advent of advanced deep learning methods and the availability of large-scale datasets, human pose estimation has made remarkable strides forward.

Previous research on 3D human pose estimation has primarily focused on traditional methods, such as sample-based approaches. A relevant survey mainly reviewed work conducted before 2015 [1]. The authors of the survey provided a comprehensive taxonomy of 3D pose estimation methods and discussed several deep learning-based approaches. They observed that the swift advancement of deep learning has naturally sped up progress in 3D human pose estimation. In a separate study, the researchers outlined the strengths and weaknesses of current deep learning-based 3D pose estimation methods and presented a range of algorithms in this area [2]. They suggested classifying input data into two categories—

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/109/20241407

skeleton and shape (contour)—from the standpoint of pose representation and carried out an extensive comparative analysis. This study provides insightful information for the development of future models and algorithms while highlighting recent developments in this techniques.

Recently, 3D as well as 2D domains, the rapid advancement of deep learning technologies has significantly accelerated the process of monocular human pose estimation.In a recent study, the authors offer a comprehensive and integrated approach to tackling this challenge, examining it from a 2D to 3D perspective [3]. Specifically, they provide an in-depth analysis of the intrinsic connections and methodological evolution between 2D and 3D pose estimation. Notably, the paper also examines solutions for challenging scenarios and summarizes the quantitative performance of benchmarks, evaluation metrics, and popular methods. This undoubtedly presents new challenges and opportunities.

Some researchers have introduced a novel approach to 3D human pose estimation in videos by designing and implementing a new model [4]. Building on cutting-edge methods, they frame the problem as 2D key-point detection, which enables more efficient 3D pose estimation. The model's performance was experimentally evaluated, demonstrating an 11% reduction in error compared to previous top results.

This paper conducts a literature review to revisit and summarize the strengths and weaknesses of various 3D human pose estimation algorithms and models. It explores the commonly utilized datasets in this field and discusses the techniques for analyzing human poses in real-world images.

The findings of this research offer valuable insights that build upon previous studies in 3D human pose estimation, underscoring the attractiveness and future potential of this area. Looking forward, numerous datasets and algorithms will continue to encounter challenges, fueled by the excitement surrounding ongoing technological innovations and various breakthroughs. This paper makes a meaningful contribution to the field of 3D human pose estimation and computer vision.

2. Case Analysis

2.1. Dilated temporal convolutions over 2D key-points serve as the basis

In 2019, a researcher proposed a simple yet effective method for 3D human pose estimation in videos, leveraging 2D key point trajectories and dilated temporal convolutions[4]. regards to computational complexity and parameter quantity the model outperforms earlier than RNN-based models, while maintaining the same level of accuracy. This breakthrough marks a substantial advancement in the field of computer vision.

Furthermore, the authors introduced a semi-supervised method that leverages unlabeled videos, which proves to be particularly effective in situations where labeled data is limited. Unlike previous semi-supervised methods, this approach requires only intrinsic camera parameters and does not necessitate real 2D annotations or multi-view images with external camera parameters. This provides a highly effective solution to the data scarcity issue in current computer vision model training and undoubtedly offers a broader practical development space for 3D human pose recognition. Overall, the proposed method surpasses previous state-of-the-art approaches in both supervised and semi-supervised settings. Notably, the supervised model demonstrates superior performance compared to other models, even those trained with additional labeled data, offering valuable insights for future computer vision model training and achieving significant progress.

2.2. Deep Learning

3D estimation of human pose algorithms based on deep learning that divide input data into two groups: skeleton and shape. [5]. For multi-person 3D pose estimation, the authors categorize methods into different approaches. The elaborate methods are separated the top-down and bottom-up further approaches. Specifically, this methods first recognize each individual, after that estimate their fracture separately, while bottom-up methods initially detect each of the joints. Afterward, allocate them to respective individuals. Contrastly, single-stage methods typically estimate both the root positions and joint displacements simultaneously [6]. The authors also present three types of human modeling

DOI: 10.54254/2755-2721/109/20241407

approaches: skeleton-based models, SMPL-based models, and surface-based models, and list various evaluation metrics such as MPJPE, PCP, and PCK [7-9]. This approach is generally faster but may be less accurate in complex scenarios with multiple overlapping people.

2.3. Based on image sequences

Among the various methods developed in the past, some approaches for Time-dependent information is not used in human pose estimate. In particular, 3D human pose estimation from monocular photos have been widely researched. A recent paper suggests using a part-aware observation system to improve monocular 3D human pose estimation, tackling the challenge of significant motion pattern inconsistencies across different parts of the human skeleton. This approach helps the model focus on specific body parts with varied movement patterns, resulting in a more precise posture prediction [10]. This method utilizes long-range part correlations to enhance 3D pose estimation further. Based on a part-aware dictionary, the part-aware attention module determines attention for the dictionary's input part features. On two popular public datasets, experimental findings demonstrate that Transformer-based models exhibit state-of-the-art 3D pose estimation performance due to the part-aware attention mechanism. Additionally, To forecast the 3D parameters of posture joints, an end-to-end network is frequently used. For example, one researcher proposed a volumetric representation of 3D poses and trained a CNN to anticipate each joint's voxel probability inside the 3D capacity, leading to more precise pose estimations [11]. In contrast, another paper presents a novel regression method for estimating human poses from static images, utilizing the soft-argmax operation. This operation is differentiable and can be seamlessly integrated into deep convolutional networks to indirectly learn part-based detection maps. By doing so, it significantly enhances the performance of the regression method, providing a more accurate estimation of human poses [12]. 2D pose estimation results are used for 3D human pose estimation to enhance natural generalization performance, which is also a common practice. Recent research has introduced a novel Transformer architecture, GraFormer, by embedding graph convolutional layers after multi-head attention blocks [13]. This architecture captures Worldwide information from every node in addition to the nodes' explicit proximity structures, enabling it to learn more robust features. Results indicate that GraFormer surpasses the performance of the state-of-the-art GraphSH on the Human3.6M dataset, showcasing its effectiveness in 3D tasks.

Estimating 3D human pose from image sequences is currently a highly effective approach, requiring the learning of temporal relationships using networks such as LSTMs. For example, a paper introduces the Recurrent 3D Pose Sequence Model (RPSM), which employs a multi-stage sequential refinement process to automatically learn structural constraints related to the image as well as temporal context. With this method, the model is able to recognize both temporal and spatial correlations in human posture sequences, leading to increasingly precise and consistent 3D pose predictions over time. [14]. As for significant depth ambiguity and severe self-occlusion in single-view images, some paper draws inspiration from the effectiveness of combining spatial dependencies and temporal consistency and introduces a local-to-global network architecture that learns multi-scale features based on graph representations [15]. Additionally, recent research on this technique from multi-view image sequences has gained attention. For example, scholars propose a natural form of supervision by exploiting the appearance constancy of a person across different frames. By utilizing this consistent visual information, the model can better track and refine the 3D pose estimation over time, enhancing accuracy and robustness, especially in challenging scenarios such as occlusion or motion blur. This approach helps the model maintain temporal coherence across frames [16]. By generating a texture map for each frame and assuming minimal variation in texture between frames, this method efficiently reconstructs the body model. The assumption of consistent texture across frames allows the system to capture the body's shape and structure more accurately, leading to a more reliable 3D reconstruction of the human pose across multiple frames. This approach capitalizes on the stability of appearance to improve pose estimation and model consistency over time.

DOI: 10.54254/2755-2721/109/20241407

2.4. Based on Transformer

Recently, By comprehensively assessing body joints in every skeletons, transformer-based techniques have been presented to estimate 3D human poses from 2D keypoint sequences. This approach leverages theTransformer's individual attention system detects long-term dependencies between joints and across frames, offering a comprehensive picture of the pose dynamics. By considering the entire sequence of keypoints, these models offer new opportunities for enhancing the accuracy and consistency of 3D pose estimation, particularly in capturing complex movements and interactions.

For example, a study proposes MixSTE (Mixed Spatiotemporal Encoder), which utilizes temporal Transformer blocks to model the temporal motion of each joint independently. In addition, it employs a spatial Transformer block to learn the spatial correlations between joints. This combination makes it possible for the model to accurately and robustly estimate 3D poses by capturing both the spatial connections between body components and the temporal dynamics of joint motions[17]. This model more effectively captures the general sequence coherence and the temporal movement trajectories of various body parts, leading to significant improvements in the efficiency and accuracy of 3D human pose estimation. By addressing both spatial and temporal relationships, the model enhances its ability to handle complex movements and ensures more consistent pose predictions across frames. Another approach introduced in a study addresses two seemingly unrelated yet conflicting issues in lifted 3D human pose estimation: processing long series inputs efficiently and being resilient to noise joint detection [18]. This method effectively integrates features from both the temporal and frequency domains, achieving a better speed-accuracy trade-off compared to its predecessors.

2.5. Arbitrary Video Inference

For inference and 3D human pose estimation on field and arbitrary videos, a study presents an approach using Detectron to infer 2D key-points and processing videos in advance to achieve optimal results [4]. However, this method encounters challenges such as the need to implement bounding box matching strategies for multi-person tracking. An alternative approach discussed in the paper addresses multi-person 3D pose estimation and tracking using multiple cameras with wide baselines [19]. This technique estimates 3D postures and re-identification (Re-ID) attributes for every person in the scene at the same time using a multi-branch network. By combining pose estimation with Re-ID, the approach not only tracks individuals across frames but also distinguishes between them, even in complex environments, thus enhancing both accuracy and robustness in multi-person scenarios. Unlike previous efforts that required noisy 2D pose estimation to establish cross-view correspondences, this approach estimates and trajectorys 3D poses directly from a voxel-based three-dimensional illustrations made from multi-view pictures. The three-dimensional space is first discretized into ordinary voxels, then feature vectors for each voxel are computedby summing together the heatmaps of the bodily joints that were displayed from every camera angle. The 3D pose is then estimated from the voxel representation by predicting whether each voxel contains a specific body joint. This method effectively leverages the multi-view information to create a detailed and accurate 3D pose representation, improving the precision of pose estimation and tracking in multi-person environments.

2.6. The popular datasets

Regarding the required datasets, the research presents commonly used 3D datasets such as Human3.6M, HumanEva, MPI-INF-3DHP, TotalCapture, and CMU Panoptic, among others [20]. The Human3.6M dataset, one of the largest motion capture datasets, contains 3.6 million human poses along with corresponding images. It offers precise 3D joint locations and synchronous higher-res videos, captured at 50 Hz by a dynamic capture system. This dataset is widely used for benchmarking 3D human pose estimation models due to its scale and accuracy.

Similarly, the HumanEva dataset comprises 7 calibrated video sequences synchronized with 3D body poses captured by a motion capture system. It features 4 subjects performing 6 common actions, such as walking, jogging, and gesturing. The dataset is split into training, validation, and test sets, making it a useful resource for evaluating pose estimation algorithms across different activities and conditions.

DOI: 10.54254/2755-2721/109/20241407

2.7. Discussion

3D human pose recognition technology undoubtedly has a significant impact across various fields, including healthcare, sports, entertainment, security surveillance, human-computer interaction, and education. It can enhance rehabilitation outcomes, optimize athletic performance, enrich gaming experiences, improve surveillance efficiency, and facilitate educational interactions. The role of data and models in this technology is irreplaceable; for instance, accurate annotations are fundamental to training effective models. Similarly, model robustness is crucial for handling pose data in diverse environments. Training models on diverse datasets can improve their performance in complex scenarios, such as pose recognition under occlusions or varying lighting conditions.

3. Conclusion

This paper reviews the current mainstream 3D pose estimation algorithms and explores commonly used benchmark datasets and methods for analyzing human poses in unlabeled field images. However, the study has limitations, as it cannot cover all human pose recognition methods comprehensively. Due to hardware constraints (e.g., GPU limitations) and missing packages in the code, full reproduction of the models and algorithms was not achievable. The future work will focus on further research in algorithm development and ensuring the reproducibility of code. What's more, human pose recognition still faces several challenges, including noise from camera views, varying lighting conditions, and diverse body shapes and clothing. Future research aims to address these issues and improve models for different environments and populations. In addition, future advancements in 3D human pose estimation will focus on several key areas. Real-time and on-device processing capabilities will be a significant development, enabling applications on mobile devices and embedded systems, thus broadening the technology's practical use. The integration of 3D pose estimation with other technologies, such as augmented reality (AR), virtual reality (VR), and robotics, promises to enhance user interactions and open up innovative applications. Moreover, as technology continues to advance, ethical considerations and privacy concerns will become more prominent, requiring the creation of clear guidelines and standards to ensure the responsible and ethical use of these technologies. Safeguards must be implemented to protect individuals' privacy, particularly when dealing with sensitive data such as human pose and motion capture, ensuring that these advancements are used in ways that respect personal rights and societal norms. The technology's scalability and efficiency will also be crucial, particularly in handling large-scale and complex datasets, which will facilitate widespread adoption and practical application across various fields. As these areas progress, the field of 3D human pose estimation is on the brink of significant evolution, with improvements in accuracy, applicability, and integration into emerging technologies.

References

[1] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, Ioannis A. Kakadiaris, 3D Human pose estimation: A review of the literature and analysis of covariates, Computer Vision and Image Understanding, Volume 152, 2016, Pages 1-20, ISSN 1077-3142, https://doi.org/10.1016/j.cviu.2016.09.002.

[2] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, Ling Shao, Deep 3D human pose estimation: A review, Computer Vision and Image Understanding, Volume 210, 2021, 103225, ISSN 1077-3142, https://doi.org/10.1016/j.cviu.2021.103225.

[3] Wu Liu, Qian Bao, Yu Sun, and Tao Mei. 2022. Recent Advances of Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective. ACM Comput. Surv. 55, 4, Article 80 (April 2023), 41 pages. https://doi.org/10.1145/3524497

[4] 3D human pose estimation in video with temporal convolutions and semi-supervised training. (n. d.). https://dariopavllo.github.io/VideoPose3D/

[5] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, Ling Shao, Deep 3D human pose estimation: A review, Computer Vision and Image Understanding, Volume 210, 2021, 103225, ISSN 1077-3142, https://doi.org/10.1016/j.cviu.2021.103225.

DOI: 10.54254/2755-2721/109/20241407

[6] Nie, X., Zhang, J., Yan, S., & Feng, J. (2019, August 24). Single-Stage Multi-Person pose machines. arXiv.org. https://arxiv.org/abs/1908.09220

[7] Cao, Z., Hidalgo, G., Simon, T., Wei, S., & Sheikh, Y. (2018, December 18). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv.org. https://arxiv.org/abs/1812.08008

[8] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2023). SMPL: a Skinned Multi-Person Linear model. In ACM eBooks (pp. 851–866). https://doi.org/10.1145/3596711.3596800

[9] Güler, R. A., Neverova, N., & Kokkinos, I. (2018). DensePOse: Dense human pose estimation in the wild. https://openaccess.thecvf.com/content_cvpr_2018/html/Guler_DensePose_Dense_Human_CVPR_2018_paper.html

[10] Y. Xue, J. Chen, X. Gu, H. Ma and H. Ma, "Boosting Monocular 3D Human Pose Estimation With Part Aware Attention, " in IEEE Transactions on Image Processing, vol. 31, pp. 4278-4291, 2022, doi: 10.1109/TIP.2022.3182269.

[11] Pavlakos, G., Zhou, X., Derpanis, K. G., & Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3D human pose. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7025-7034).

[12] Luvizon, D. C., Tabia, H., & Picard, D. (2019). Human pose regression by combining indirect part detection and contextual information. Computers & Graphics, 85, 15-22.

[13] Zhao, W., Tian, Y., Ye, Q., Jiao, J., & Wang, W. (2021, September 17). GRAFormer: Graph Convolution Transformer for 3D pose Estimation. arXiv.org. https://arxiv.org/abs/2109.08364

[14] Lee, K., Lee, I., & Lee, S. (2018). Propagating lstm: 3d pose estimation based on joint interdependency. In Proceedings of the European conference on computer vision (ECCV) (pp. 119-135)

[15] Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T. J., Yuan, J., & Thalmann, N. M. (2019). Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 2272-2281).

[16] Pavlakos, G., Kolotouros, N., & Daniilidis, K. (2019). Texturepose: Supervising human mesh estimation with texture consistency. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 803-812).

[17] Zhang, J., Tu, Z., Yang, J., Chen, Y., & Yuan, J. (2022). MIXSTE: SEq2SEQ Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in video. https://openaccess.thecvf.com/content/CVPR2022/html/Zhang_MixSTE_Seq2seq_Mixed_Spatio-Temporal_Encoder_for_3D_Human_Pose_Estimation_CVPR_2022_paper.html

[18] Zhao, Q., Zheng, C., Liu, M., Wang, P., & Chen, C. (2023). PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation. https://openaccess.thecvf.com/content/CVPR2023/html/Zhao_PoseFormerV2_Exploring_Frequency_Domain_for_Efficient_and_Robust_3D_Human_CVPR_2023_paper.html

[19] Y. Zhang, C. Wang, X. Wang, W. Liu and W. Zeng, "VoxelTrack: Multi-Person 3D Human Pose Estimation and Tracking in the Wild, " in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 2, pp. 2613-2626, 1 Feb. 2023, doi: 10.1109/TPAMI.2022.3163709.

[20] Zczcwh. (n.d.). DL-HPE/3D_dataset at main. zczcwh/DL-HPE. GitHub. https://github.com/zczcwh/DL-HPE/tree/main/3D_dataset