3D pose estimation in a multi-camera scene: Main approaches and What is the Future?

Ana Filipa Rodrigues Nogueira[1,2]
ana.f.rodrigues@inesctec.pt

Hélder P. Oliveira[1,3]
helder.f.oliveira@inesctec.pt

Luís F. Teixeira[1,2]
luisft@fe.up.pt

[1] INESC TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência
R. Dr. Roberto Frias, Porto, Portugal

[2] FEUP - Faculdade de Engenharia da Universidade do Porto
R. Dr. Roberto Frias, Porto, Portugal

[3] FCUP - Faculdade de Ciências da Universidade do Porto
Rua do Campo Alegre 1021 1055, Porto, Portugal

Abstract

3D pose estimation is a crucial task for numerous real-world applications. However, several obstacles, like the lack of 3D annotated datasets, occlusions or variations in human appearance, have been limiting the efficiency and reliance on the existing 3D pose estimation models and restraining their employment in real-world settings. Multi-view machine learning solutions have been increasingly explored to overcome those obstacles due to the greater representation power of deep neural networks and the increasing availability of environments with several cameras.

This paper focuses on deep learning methods for 3D pose estimation in a multi-camera setting. After comparing the various methodologies, it was possible to conclude that the best method depends on the intended application. Thus, future research should focus on finding a solution that allows a fast inference of a highly accurate 3D pose while keeping a low computational complexity. To this end, techniques like active learning, selection of views and multi-modal approaches should be further explored.

Keywords: 3D pose; Multi-view; Deep Learning (DL).

1 Introduction

The goal of 3D pose estimation is to recover the body configuration of every person in a scenario. Solving this task is critical for several significant real-world applications such as human-computer interaction, surveillance systems, and action recognition where detecting and identifying abnormal behaviour might be crucial to minimise repercussions, for example, in an assisted living situation. Also, it can be used in gaming, animations, rehabilitation or sports assessments, among many others [1, 4, 8].

Nonetheless, obstacles such as occlusions of people, incorrect pose detection, poor camera calibration, scarcity of 3D labelled data, depth ambiguity, resemblances or variances in human appearance, and random camera viewpoints have made it challenging to develop an efficient and accurate model [4, 8]. The higher capability of deep learning-based models to capture the most relevant features has led to better results than traditional approaches in settings with either one or multiple cameras [1]. Nevertheless, multi-camera solutions, because they can capture more 3D geometry information about the human body, allow the reconstruction of more reliable 3D poses [7].

Hence, this paper will compare the existing DL approaches for multi-view 3D pose estimation. Depending on the number of people in the scene, the existing approaches can be divided into single-person or multi-person [1, 8]. Therefore, Section 2 addresses single-person approaches and Section 3 multi-person approaches. Then, Section 4 points to new directions of research and Section 5 presents the overall conclusions.

2 Single-person approaches

Single-person approaches can be divided into two distinct categories: one-stage methods and two-stage methods, which are described in length in the following subsections.

2.1 One-stage methods

In one-stage methods, the 3D pose is directly reconstructed from the input image (see Figure 1).

This type of approach avoids the accumulation of errors due to incomplete or erroneous 2D pose estimations by directly predicting the 3D pose. Nonetheless, the scarcity of 3D labelled datasets severely limits the results. To cope with this problem, Rhodin et al. [6] proposed a semi-supervised method that learns a geometry-aware representation of the human body from unlabeled images by predicting an image from one point of view based on an image from another. Then, uses some supervision to learn the mapping from the 3D geometry representation to the 3D pose, thereby, reducing the required amount of training data.

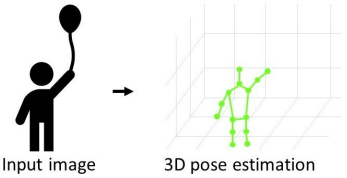

Figure 1: One-stage methods process

2.2 Two-stage methods

Two-stage methods regress the 3D pose by merging the predicted 2D joint positions (see Figure 2).

Figure 2: Two-stage methods process

An example of a two-stage approach is AdaFuse [11], an adaptive fusion method that attributes weights to each camera view to favour those with better feature quality, allowing to combat the occlusion problem. However, the quality of the predicted 3D poses is highly affected by the inaccuracies of the 2D pose estimations. The 2D-to-3D lifting human pose is one of the biggest issues of two-stage methods. PoseFormerV2 [12] tries to surpass that problem for transformer-based approaches by converting the input joint sequences into low-frequency coefficients which are sufficient to express the entire visual identity and also, filter out the high-frequency noise. Moreover, it fuses temporal and frequency features to have complementary information to obtain better 2D joint estimation. Even though this methodology has better results than some preceding implementations based on transformers, compared with AdaFuse, the results of PoseFormerV2 still fall short. This shows the continuing need to search for the best combination of model architecture and 2D-to-3D lifting technique to get more accurate results.

3 Multi-person approaches

Multi-person approaches can be divided into bottom-up and top-down methods, which will be discussed in further detail in the following subsections.

3.1 Bottom-up methods

Bottom-up approaches start by determining all body joint locations in the scene and then grouping them for each individual (see Figure 3).

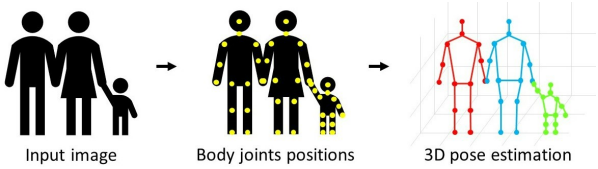

Figure 3: Bottom-up methods process

The key advantages of this type of approach are the fast inference time and the fact that the number of persons as little influence on the computational complexity [2]. For example, Light3DPose [2] uses a Volumetric Network to aggregate the joints and predict the pose in the 3D space, making the model more robust to occlusions and suitable for crowded scenes. Nonetheless, volumetric computations are computationally costly, so, it uses a lightweight 2D backbone to predict and then, project the 2D feature maps into a lower-dimensional feature space to promote fast inference.

3.2 Top-down methods

In top-down methods, a bounding box is defined around each person, then a single-person pose estimator is executed and the final 3D pose is obtained (see Figure 4).

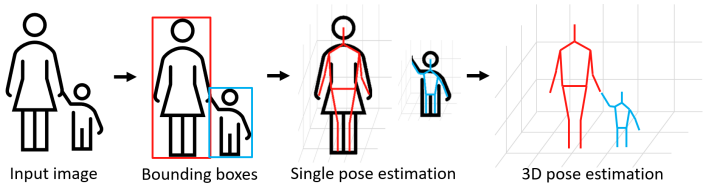

Figure 4: Top-down methods process

A popular top-down approach that has been serving as the basis for other methods, like [5, 10], is VoxelPose [7] which operates directly in the 3D space, to avoid mistakes due to incorrect 2D pose estimations. VoxelTrack [10] extends VoxelPose to also, track the 3D poses throughout time. Besides, VoxelTrack uses a more robust 2D backbone network and adds extra 3D heatmap supervision of all joints to improve the 3D pose estimation results. Nonetheless, both methods suffer a huge performance drop when the number of camera views decreases. TesseTrack [5], on the other hand, is less affected by this, due to the incorporation of temporal information via the 4D spatio-temporal Convolutional Neural Network (CNN) which makes the model better at estimating the joint positions, more robust to occlusions and able to deal with appearance ambiguities in a single frame. Moreover, according to Papers With Code website[1], TesseTrack holds the best results in 3 benchmark datasets: Campus, Shelf and CMU Panoptic. Nevertheless, although these methods are highly accurate, the computational cost tends to grow substantially with the number of people in the scene, making them inadequate for crowded scenarios [1, 8, 10].

4 Challenges and Opportunities

The scarcity of labelled datasets affects the performance of all types of models [1, 4, 8]. Despite some works, such as [6], propose semi-supervised or weakly-supervised models to address this challenge, they still require some labelled data. Hence, other studies have concentrated on using active learning to enhance the data labelling process [3]. Feng et al. [3] demonstrated that using active learning, whether alone or combined with self-training, effectively enhances the data annotation process and, consequently, 3D pose estimation. As only very few studies have explored this cost-effective technique for this task, probably the full potential benefits are yet to be unfolded.

Other methods like the generation of synthetic data or data augmentation techniques such as multi-image mixing, information dropping, and image corruption by adding noise or blur can, also, help overcome the limited number of available datasets and consequently, improve the accuracy of the models by providing them with more training data.

Regarding approaches that rely on proper 2D pose estimations to accurately predict the 3D pose, an interesting technique to investigate to boost the accuracy while minimising the energy usage is to solely take into consideration the subset of cameras with better view perspectives concerning occlusions and energy states to predict the 2D pose [9]. This will allow for greater camera battery conservation and better 3D pose estimations, both of which are important to be able to implement an edge-assisted 3D pose estimation system in the real-world [9].

Finally, multi-modal approaches can combine multi-view images with signals from other sensors, such as radio signals, IMUs, LiDAR, or Wi-Fi. Thus, multi-modal learning can provide a more robust and accurate solution by overcoming obstacles like occlusions, which is crucial for real-world applications. However, further research is still necessary [4].

[1] https://paperswithcode.com

5 Conclusion

In conclusion, the lack of 3D labelled datasets strongly hampers the potentialities of one-stage approaches, whereas two-stage methods rely heavily on correct 2D pose estimations. On the other hand, for multi-person settings, top-down methods can generally produce better results, but at the expense of higher computational complexity and longer inference times when compared to bottom-up approaches [1, 8]. So, depending on the intended application, the best model type may differ because of the trade-off between complexity and performance. Therefore, a model must balance high accuracy, quick inference, and low computational cost to be suitable for a wide range of real-world applications. A multi-modal approach combined with active learning strategies might be the future for an efficient and effective solution.

6 Acknowledgments

This work has received funding from the European Union's Horizon Europe research and innovation programme under the Grant Agreement 101094831-CONVERGE project and by National Funds through the Portuguese funding agency, FCT-Foundation for Science and Technology Portugal, a PhD Grant Number 2023.02851.BD.

References

[1] M. Ben Gamra and M. A. Akhloufi. "A review of deep learning techniques for 2D and 3D human pose estimation". In: *Image and Vision Computing* 114 (2021), p. 104282. DOI: 10.1016/j.imavis.2021.104282.

[2] A. Elmi, D. Mazzini, and P. Tortella. "Light3DPose: Real-time Multi-Person 3D Pose Estimation from Multiple Views". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jan. 2021, pp. 2755–2762. DOI: 10.1109/ICPR48806.2021.9412652.

[3] Q. Feng et al. "Rethinking the Data Annotation Process for Multi-View 3D Pose Estimation With Active Learning and Self-Training". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2023, pp. 5695–5704.

[4] G. Lan et al. "Vision-Based Human Pose Estimation via Deep Learning: A Survey". In: *IEEE Transactions on Human-Machine Systems* 53.1 (2023), pp. 253–268. DOI: 10.1109/THMS.2022.3219242.

[5] N. D. Reddy et al. "TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 15190–15200.

[6] H. Rhodin, M. Salzmann, and P. Fua. *Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation*. 2018. DOI: 10.48550/ARXIV.1804.01110.

[7] H. Tu, C. Wang, and W. Zeng. *VoxelPose: Towards Multi-Camera 3D Human Pose Estimation in Wild Environment*. 2020. DOI: 10.48550/ARXIV.2004.06239.

[8] D. Zhang et al. "Deep Learning Methods for 3D Human Pose Estimation under Different Supervision Paradigms: A Survey". In: *Electronics* (2021). DOI: 10.3390/electronics10182267.

[9] L. Zhang and J. Xu. *E^3Pose: Energy-Efficient Edge-assisted Multi-camera System for Multi-human 3D Pose Estimation*. 2023. DOI: 10.48550/ARXIV.2301.09015.

[10] Y. Zhang et al. "VoxelTrack: Multi-Person 3D Human Pose Estimation and Tracking in the Wild". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.2 (2023), pp. 2613–2626. DOI: 10.1109/TPAMI.2022.3163709.

[11] Z. Zhang et al. "AdaFuse: Adaptive Multiview Fusion for Accurate Human Pose Estimation in the Wild". In: *International Journal of Computer Vision* 129.3 (Mar. 2021), pp. 703–718. DOI: 10.1007/s11263-020-01398-9.

[12] Q. Zhao et al. "PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 8877–8886.