Temporally-aware Human Pose Estimation on 3D videos: A study of the State of the Art

Joris LIMONIER[1][0000−0002−0393−2247], Frédéric PRECIOSO[2][0000−0001−8712−1443], and Lucile SASSATELLI[2][0000−0003−1232−1787]

[1] Université Côte d'Azur, Biot, France joris.limonier@etu.univ-cotedazur.fr
[2] {frederic.precioso, lucile.sassatelli}@univ-cotedazur.fr

Abstract. The abstract should briefly summarize the contents of the paper in 15–250 words.

Keywords: Computer Vision · Human Pose Estimation · 3D Human Pose Estimation · Human Pose Estimation in Videos

1 Introduction

The Deep Learning revolution, coupled with increasing computing power and the improved use of GPU opened new opportunities in the field of Computer Vision. New architectures arose and new techniques suggested numbers of parameters that hadn't been seen before, some of them reaching hundreds of millions of parameters [4]: up to 94.9M for Faster R-CNN [5], 127.3M for Cascade R-CNN [6], 51.0M for FCOS [7], 210.1M for CenterNet [8], 135.2M for Cascade Mask R-CNN [6], 138.2M for Hybrid Task Cascade [9] and 63.4M for Mask R-CNN [5]. Such complex networks manage to segment images, detect objects images or identify the pose of a person. Our interest goes to the latter task. The task of Human Pose Estimation (HPE) aims at detecting joints of a human being in a frame. This could be considered a solved problem in the 2D case when the person is clearly visible. Some other cases are more challenging, one of which is when some body parts are hidden (occlusions) in a 2D image. Another challenging case is finding spatial coordinates in a 3D image, this will be our focus throughout this study.
Furthermore, one can consider HPE applied to images but also HPE applied to videos. We want to focus on videos, as well as the interest of considering the temporal dimension rather than a frame-per-frame joints detection. Doing so brings its set of challenges and its complexity to the problem, which makes it even more interesting.
We want to study the current state of the art for 3D HPE on videos while considering the temporal dimension. In order to do so, we will list and examine related work, then we will gather existing datasets and metrics while analysing their strengths and weaknesses. Subsequently, we will evaluate and compare existing methods on common datasets. Finally, we will give our conclusions and propose study pathways for the future of this study.

2 Related Work

3 Datasets and Metrics

3.1 Datasets

3.2 Metrics

4 Evaluation and Comparison

5 Conclusion and Perspectives

5.1 Conclusion

5.2 Perspectives

References

1. C. Zheng et al., "Deep Learning-Based Human Pose Estimation: A Survey," arXiv:2012.13392 [cs], Jan. 2021, Accessed: Jan. 12, 2022. [Online]. Available: http://arxiv.org/abs/2012.13392
2. H.-Y. Wu, L. Nguyen, Y. Tabei, and L. Sassatelli, "Evaluation of deep pose detectors for automatic analysis of film style," in EUROGRAPHICS Workshop on Intelligent Cinematography and Editing, Reims, France, 2022, p. 9.
3. W. Li, H. Liu, H. Tang, P. Wang, and L. V. Gool, "MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation," in 2022 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, Jun. 2022.
4. Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X, Liu W. Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence. 2020 Apr 1;43(10):3349-64.
5. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
6. Cai, Z. and Vasconcelos, N., 2019. Cascade R-CNN: high quality object detection and instance segmentation. IEEE transactions on pattern analysis and machine intelligence, 43(5), pp.1483-1498.
7. Tian, Z., Shen, C., Chen, H. and He, T., 2019. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9627-9636).
8. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q. and Tian, Q., 2019. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6569-6578).
9. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W. and Loy, C.C., 2019. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4974-4983).