Temporally-aware Human Pose Estimation on 3D videos: A review 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

HPE encompasses several subtasks, not all of which we are interested in. In this section, we describe the tasks we want to tackle, then we describe existing approaches.

2.1 Task Description

As mentioned in section 1, HPE can be applied in 2D (planar coordinates), as well as in 3D (spatial coordinates), our focus is on the 3D case. This means that we want to predict the 3D location of body joints. Several industries find applications of these techniques, namely the movie & animation industry and the sport industry, only to name a few.
When talking about images that have been taken from one viewpoint only, we talk about *monocular* images. Such images result in a projection from the 3D space we live in to the 2D space that is the image. As such, each point in the image is the projection of one of an infinite number of points (namely, the whole straight line starting at the camera and going infinitely in the direction of the projected point). For this reason, the 3D HPE monocular problem is ill-posed.
Although ill-posed, inferences could be made in the the 3D HPE monocular problem thanks to the temporal awareness. This is why we want to work with videos and take advantage of the consecutiveness of their frames. We want the algorithm to understand that the frames are in a sequence, not simply making a frame-per-frame prediction. Predicting one frame at a time is the simplest approach, but completely diregards the advantage given by a video over an equivalent number of independent frames.

2.2 Approaches

Several methods can be used to solve the 3D, temporally-aware video HPE problem, we will detail them in this section.

3D HPE from monocular RGB videos This is the simplest case where one uses a unique conventional camera to capture videos. As mentioned previously, this is however an ill-posed problem as the captured 2D images come from a projection of the 3D space onto a 2D screen. Reverting the operation seems impossible in the general, mathematically-posed problem but using some commonsensical tricks could help achieve better results in some cases. For instance, the length of a human arm lies within a given interval for all human beings, which a smart algorithm could use to infer a set of plausible positions to lift such an arm from 2D to 3D.
We consider the case of single-person videos for now. In this setting, we distinguish three types of tasks: skeleton-only, the kinematic model and Human Mesh Recovery (HMR). We will detail each of these techniques in their own paragraph.
The Skeleton-only method aims at predicting the spatial (*i.e.* 3D) coordinates

of the joints of a person. Estimating a 3D skeleton can be done in two ways. The first way is called direct estimation. It consists of training a neural network to predict the 3D joints directly from the image. The other way is to use an existing 2D joint predictor, which are very performant as we mentioned before, then try to lift the joints from 2D to 3D with a custom neural network.
The kinematic model is another task which consists in estimating the segments between joints. This task allows to pass custom rules to the network and leverage prior knowledge such as restricting joint rotation angles or fixing bone-length ratios.
HMR is the final task in this list. It aims at superimposing a 3D, volumetric body model correctly in space. The body is seen as a unique, connected (in the topological sense) structure where the limbs are able to evolve within their respective ranges of motion.

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).