Master Research Proposal Nr:

Topic: Investigation of Robustness of Single-View 3D Pose Estimators under Challenging Conditions

Untersuchung der Robustheit von Single-View-3D-Pose-Estimators unter schwierigen Bedingungen

3D human pose estimation is a widely used aspect of computer vision with applications ranging from guided movement exercises in the home environment [1] to supportive systems for athletes and coaches [2]. In recent years, pose estimation incorporating depth information, e.g. projecting humans in three-dimensional space, has progressed extensively. Technologies in 3D human pose estimation and neural radiance fields have been significantly advanced, such as musculoskeletal dynamics in biomechanics [3], photorealistic head avatars [4], and intuitive physics-based motion prediction models [5], with broad applications across these fields. Some presented methods include uplifting from 2D to 3D space using a fully connected residual network to regress 3D joint locations from 2D joint locations, while another method introduced a volumetric representation for direct prediction of 3D poses from 2D images [6]. There have been significant improvements in deep learning methods, resulting in more stable and accurate models that only require a single camera view [7]. However, these models are still not robust enough to handle all real-life scenarios [8]. The robustness of existing estimators under challenging conditions, such as difficult camera angles, varying lighting conditions especially in outdoor environments, and the presence of occlusions such as overlapping body parts or objects blocking parts of the person being estimated, remains a significant concern in both 2D and 3D approaches [6]. Numerous advances, such as dual-stream spatiotemporal transformer models [9], modulated graph neural networks (GCNs) [10], or combinational networks of transformers and GCNs [12] were introduced to enhance the performance and robustness of single-view 3D pose estimations.

This thesis proposes to investigate the robustness of single-view 3D pose estimators by evaluating their performance under different conditions and testing novel approaches to improve the stability of pose predictions in a temporally coherent context. Prediction pipeline design choices, such as heatmap-based methods like generating Gaussian heatmaps centered at the ground truth joint locations for joint localization and regression-based approaches for estimating 3D coordinates and angles, along with the generalization of data augmentation techniques for addressing overfitting and specific loss functions are thoroughly tested.

Consequently, this work consists of the following parts

- Capturing data that includes a range of human actions, such as jumping jacks and tying shoelaces, while taking into account various conditions like different lighting, occlusion by objects (such as furniture), and occlusion by other parts of the human body in the frame, to create a comprehensive and robust dataset for fine-tuning and evaluation.

- Capture keypoints with OMC (optical motion capture) setup and compare global joint positions to test model generalization. If necessary, fine-tune with recorded data to evaluate model performance, considering the impact on generalizability due to reduced domain gap.

- Evaluate the accuracy [12], aiming to achieve Mean Per Joint Position Error (MPJPE) of at least 20mm [6] on the captured robust data, for various state-of-the-art single-view 3D pose estimation approaches, including graph-based models, detection-based methods such as keypoint mask R-CNN, spatio-temporal transformers, convolutional models, and others. Furthermore, testing the effectiveness of various pose estimation methods such as uplifting from 2D to 3D space, and direct spatiotemporal joint prediction.

- Introduce an approach to enhance the stability and robustness of single-view 3D pose estimators by incorporating techniques such as adding additive temporal noise or 2D pose augmentation with added jitter [13].

Advisors: Jonas Müller, M. Sc., Alexander Weiß, M. Sc.,
Prof. Dr. Bjoern Eskofier, Prof. Dr. Anne Koelewijn
Student: Abhinav Jaivishnu Choudhary
Start – End: 15.07.2024 – 15.01.2025

References

[1] Hellsten T, Karlsson J, Shamsuzzaman M, Pulkkis G. The Potential of Computer Vision-Based Marker-Less Human Motion Analysis for Rehabilitation. *Rehabilitation Process and Outcome*. 2021;10. doi:10.1177/11795727211022330.

[2] Wang, Jianbo, et al. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. *Proceedings of the 27th ACM international conference on multimedia*. 2019.

[3] Uhlrich SD, Silder A, Beaupre GS, Shull PB, Delp SL. Subject-specific toe-in or toe-out gait modifications reduce the larger knee adduction moment peak more than a non-personalized approach. *Journal of Biomechanics*. 2018;66:103-110. doi:10.1016/j.jbiomech.2017.11.003.

[4] Qian S, Kirschstein T, Schoneveld L, Davoli D, Giebenhain S, Nießner M. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. *Technical University of Munich*. 2023. arXiv:2312.02069.

[5] Tripathi S, Trivedi R, Tang S, Black MJ. 3D Human Pose Estimation via Intuitive Physics. *CVPR*. 2023.

[6] Zheng, Ce, et al. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys* 56.1 (2023): 1-37.

[7] El Kaid, Amal, and Karim Baïna. A Systematic Review of Recent Deep Learning Approaches for 3D Human Pose Estimation. *Journal of Imaging* 9.12 (2023): 275.

[8] Yu, Cheng, Bo Yang, Bo Wang, and Robby T. Tan. Occlusion-Aware Networks for 3D Human Pose Estimation in Video. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2020.

[9] Zhu, Wentao, et al. Motionbert: A unified perspective on learning human motion representations. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[10] Zou, Zhiming, and Wei Tang. Modulated graph convolutional network for 3d human pose estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[11] Mehraban, Soroush, Vida Adeli, and Babak Taati. MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024.

[12] Gozlan, Yoni, Antoine Falisse, Scott Uhlrich, Anthony Gatti, Michael Black, and Akshay Chaudhari. OpenCapBench: A Benchmark to Bridge Pose Estimation and Biomechanics. *arXiv preprint arXiv:2406.09788*. 2024.

[13] Hoang, Trung-Hieu, Mona Zehni, Huy Phan, Duc Minh Vo, and Minh N. Do. Improving the Robustness of 3D Human Pose Estimation: A Benchmark and Learning from Noisy Input. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2024.