Optimizing Monocular Camera Perspectives for Accurate 3D Human Pose Estimation

Muhammad Shahid [1], Rob Argent[2], Georgiana Ifrim[1], and Brian Caulfield[1]
[1]Insight SFI Centre for Data Analytics, University College Dublin, Ireland
[2]School of Pharmacy and Biomolecular Sciences RCSI University of Medicine and Health Sciences, Dublin, Ireland
[1] Email: muhammad.shahid1@ucd.ie

Summary

Marker-based motion capture systems (MoCap) are widely recognized for their high precision in measuring biomechanical parameters. However, their accessibility is limited due to high cost. This has spurred interest in markerless systems, which leverage deep learning advancements to provide comparable accuracy with greater accessibility across domains. Yet, the reliability of multicamera MoCap systems depends on precise 3D human pose estimation (HPE), requiring accurate camera calibration, synchronization and placement. In this study, we evaluate monocular camera-based 3D HPE and the effect of camera perspective. Specifically, we assess PoseFormerV2's performance in predicting 3D HPE from single-camera inputs across eight perspectives and compare its accuracy to multi-view-based 3D keypoints. The optimally placed single-camera setup achieved a Euclidean error of 63 mm, compared to 51 mm in multi-view HPE, while an ill-posed camera resulted in a higher error of 94 mm.

Introduction

Motion capture (MoCap) systems are vital in fields like rehabilitation, sports analysis, animation, and human-computer interaction. Traditional marker-based systems, while highly precise but costly , restricting their use to well-funded labs. These limitations have fueled interest in markerless MoCap systems. A recent advancement in computer vision have enhanced 2D human pose estimation (HPE) models, enabling markerless motion capture with multi-camera setups [3]. However, these systems face challenges like synchronization, calibration. Recent progress in computer vision now allows 3D HPE [1] from monocular video data by leveraging temporal relationships, offering a single-camera solution. This approach eliminates calibration needs, making it accessible to non-technical users on pervasive technology.

Accurate 3D human pose estimation (HPE) from monocular video streams can be challenging due to self-occlusion, making camera placement critical. This study investigates the optimal camera perspective for 3D keypoint detection by analyzing how perspectives and movements impact accuracy. Using PoseFormerV2 [1], a state-of-the-art deep learning model, we aim to enhance single-camera 3D HPE, enabling motion capture (MoCap) to be accessible through devices like smartphones for resource-constrained applications without requiring technical expertise.

Methods

The adapted pipeline summarized as follows. First, 2D human poses are estimated for each camera video stream using the HRNet model. As we used RRIS40 [2] dataset which has eight synchronized cameras. Subsequently, the 3D poses are reconstructed using two different approaches. The first approach employs a multiview epipolar geometry method, while the second approach utilizes a pre-trained PoseformerV2 to infer 3D poses independently for each camera stream based on the previously computed 2D poses. The RRIS40 dataset lacks predefined temporal segmentation by movement type. We manually annotated videos into four categories: static, squatting, stepping, and jumping, while maintaining consistent camera perspectives.

Table 1: Comparison of Monocular and Multi-Camera 3D HPE: Procrustes-Aligned MPJPE Errors (mm)

Camera	Static	Squatting	Stepping	Jumping	Average
FL-Cam	**35**	75	75	70	64
RS-Cam	75	115	91	93	94
LS-Cam	75	111	86	84	89
BL-Cam	70	97	63	76	76
F-Cam	**35**	**70**	82	73	65
BR-Cam	48	88	68	78	70
FR-Cam	41	95	75	78	72
B-Cam	42	82	**59**	**71**	**63**
Multi-View	**33**	**69**	**46**	**56**	**51**

FL: Front Left, **RS:** Right Side, **LS:** Left Side, **BL:**Front Left, **F:**Front, **BR:** Back Right, **FR:**Front Right, **B:** Back

Results and Discussion

To assess the impact of camera perspective on 3D HPE accuracy, we calculated the Procrustes-Aligned Mean Per Joint Position Error (MPJPE) in millimeters (mm), as shown in Table 1. The Back (B-Cam), Front Left (FL-Cam), and Front (F-Cam) cameras achieved the lowest competitive average MPJPE (63, 64, and 65 mm, respectively), indicating superior accuracy for single-camera setups, while the Left Side Camera (LS-Cam) performed the worst (94 mm). This superior performance can be attributed to minimal occlusion and the robustness of the deep learning model. As expected, the multi-camera approach achieved the lowest MPJPE (51 mm), demonstrating the advantages of multi-view triangulation.

Conclusions

This study demonstrated that a single monocular camera, when optimally positioned, can achieve 3D HPE accuracy comparable to multi-camera systems, emphasizing the importance of camera perspective. The results highlight the potential of monocular motion capture for real-world applications, particularly in clinical settings, due to its practicality and convenience.

Acknowledgments

This study has received funding from Science Foundation Ireland [12/RC/2289 P2] at the Insight SFI Research Centre for Data Analytics and the European Union's H2020 Marie Sklodowska-Curie Cofund programme, NeuroInsight [Grant ID: 101034252].

References

[1] Zhao Qitao et al. *IEEE/CVF CVPR*, (2023).
[2] Jatesiktat Prayook et al. (2024). *IEEE JBHI*.
[3] Scott et al. (2023). *PLoS Comp. Bio.*, **19.10**: e1011462.