# The Role of Depth for Human Motion Assessment with a Single RGB-D Camera: Preliminary Findings

B. Lagomarsino✉, V. Falzarano, G. Marchesi, T. Falchi Delitala, F. Odone, M. Casadio, M. Moro

*Abstract*— **Telerehabilitation is an emerging remote therapy method, particularly for people with neurological disorders. In this context, single-camera markerless motion capture technology offers a cost-effective, accessible and easy-to-use solution for motion assessment. Differently from multi-camera systems or those requiring physical markers, single-camera approaches are suitable for home environments, which makes them ideal for telerehabilitation. However, most studies have focused on movements parallel to the camera's plane, leaving a gap in understanding their accuracy and precision for more complex, multi-directional movements. This study aims to fill this gap by analyzing hip movements in different directions of ten unimpaired participants using a single RGB-D camera. Comparing markerless approaches with a marker-based system, we found that single-camera methods perform well in the frontal plane, while the error increases in movements involving a sagittal component. This study highlights the strengths and limitations of camera-based motion analysis, contributing to the advancement of telerehabilitation.**

## I. INTRODUCTION

TELEREHABILITATION offers remote rehabilitation services via telecommunication networks and the internet, allowing access to therapy from home or remote locations. It has demonstrated its potential in treating different neurological disorders as stroke, Parkinson's disease and multiple sclerosis [1]. One of the primary challenges is ensuring reliable assessment and feedback of individuals' movements and performance, which is fundamental to monitor progress and guide recovery [2]. The less invasive solution involves markerless motion capture technology that allows movements to be tracked from video data, without physical markers or any other wearable sensors [3], [4]. Recent advancements in video analysis techniques, driven by deep learning and computer vision, hold significant promise for telerehabilitation. These technologies enable users to assess their performance over time and receive online feedback to correct wrong executions without the constant monitoring of a therapist. Recent 3D pose estimation algorithms that can estimate depth information from a single RGB image (*e.g.*, MediaPipePose [5]) and cameras with built-in depth sensors have the potential to simplify motor task analysis by eliminating the need for multiple camera systems. Although some studies have evaluated the feasibility of using a single camera for human motion characterization, they have exclusively focused on movements performed in the plane parallel to that of camera acquisition, neglecting the depth information [6], [7]. Therefore, comprehensive research on the accuracy of video analysis tools analyzing out-of-plane movements with respect to camera position appears necessary. Given the current state of the art, this study aims to characterize motion using a single RGB-D camera by employing two methods of depth estimation: : (i) applying a monocular 3D pose estimation algorithm to the RGB video , and (ii) using the depth information from the sensor itself. The objective is to evaluate how these approaches affect the estimation of quantitative angle measures compared to those extracted with a gold-standard stereophotogrammetric system.

This work was funded by the European Union—NextGenerationEU and by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.5, project "RAISE—Robotics and AI for Socio-economic Empowerment" (ECS00000035). G. M., V. F., T. F. D., M. M., F. 0. and M. C. are part of RAISE Innovation Ecosystem.

B. L. and M.C. are with DIBRIS, University of Genova, Italy, ( ✉ corresponding author, e-mail: beatrice.lagomarsino@edu.unige.it).

G. M., V. F. and T.F. D. are with Movendo Technology s.r.l., Genoa, Italy.

M. M. and F. O. are with DIBRIS, University of Genova, Italy and with the Machine Learning Genoa (MaLGa) Center.

M. C. is with DIBRIS, University of Genova, Italy.

## II. MATERIAL AND METHODS

### A. Experimental Setup and Protocol

Ten unimpaired volunteers performed hip movements in different planes: abduction (A) in the frontal plane, flexion (F) and extension (E), in the sagittal plane, and flexion-abduction (FA), extension-abduction (EA) in both planes (Figure 1). We asked the participants to perform these movements facing an RGB-D camera (*i.e.*, with their frontal plane in front of the device), the Intel RealSense D435 acquiring at 30Hz. This camera captures both RGB video and depth information using an integrated built-in depth sensor, which uses infrared and stereo cameras for markerless motion tracking. The setup included a calibrated motion capture system (*i.e.*, the Vicon, acquiring at 100 Hz) that we considered the gold-standard. To accurately capture hip movements, eight markers were placed on each leg according to the Davis Protocol for lower limb kinematics. The markers were positioned on the anterior superior iliac spine, posterior superior iliac spine, lateral thigh, lateral epicondyle of the femur, lateral shank, lateral malleolus, heel, and the head of the fifth metatarsal. Participants performed 10 repetitions of each movement with both left and right legs at distances of 2 and 3 meters from the camera. The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the local ethical committee (Comitato etico per la ricerca di Ateneo CERA- Università degli Studi di Genova, protocol n. 2023/93, 14/12/2023).

### B. Data Analysis

For the gold standard marker-based approach (*Marker*), we considered *(x, y, z)* coordinates of the passive markers positioned on pelvis and knees. For the markerless approaches, we used MediaPipe to estimate the *(x, y)* coordinates of the corresponding keypoints (*i.e.*, hips and knees). For the z-coordinates, we employed two approaches: *Markerless1* leveraged the depth information estimated by MediaPipe, and *Markerless2* extracted the depth information from the sensor integrated in the RealSense. Specifically, for *Markerless1* we considered single RGB images as input for MediaPipe and we obtained *(x, y, z)* keypoints' positions directly as output. For *Markerless2* we extracted the 3D coordinates from the depth sensor's point cloud corresponding to the pixel position of the keypoints of interest. We computed the hip angle in the 3D space for each approach as the angle between the leg body segment (*i.e.*, the vector connecting hip and knee landmarks) during the task and the same during the initial position (*i.e.*, participants standing still and facing the camera, Figure 1). After obtaining angular data from each frame of all systems, for each movement we extracted the peak values representing the maximum angular displacement. Then, mean absolute errors (MAEs) were calculated by computing the difference between the maximum angular displacement obtained from both *Markerless* approaches and that from *Marker* approach (*MAE#* representing the error between *Marker* and *Markerless#*, # referring to 1 and 2). The obtained values were finally averaged for repeated movements. We aimed to evaluate if the performance of the markerless approaches was comparable to a marker-based method for assessing movement in different directions. Additionally, we examined potential differences by considering two distances from the camera and movements performed with the left and right legs. We conducted a repeated measures ANOVA (rANOVA) on angle values, considering four factors: approaches (*Marker*, *Markerless*), body side, movement direction, and distance from the camera. A Bonferroni-Holm correction for multiple comparisons was applied to the post hoc analysis performed following the significant factors.

## III. Results and Discussion

As expected, the angles did not significantly differ for the body side and the distance from the camera. Therefore, we averaged the data across both distances and legs. A significant difference was found among the measurement approaches depending on the movement direction (approaches*direction interaction p=0.034). Figure 1 shows the mean angle values for each direction of movement for both *Marker* and *Markerless* approaches. The angles computed with the three approaches were not significantly different for the movements performed purely in the frontal plane (*i.e.*, parallel to the acquisition plane, A in Figure 1). This aligns with existing literature, indicating that the *Markerless* systems provide results comparable to the gold standard for these movements. Additionally, the errors relative to the gold standard were similar for both camera-based approaches.

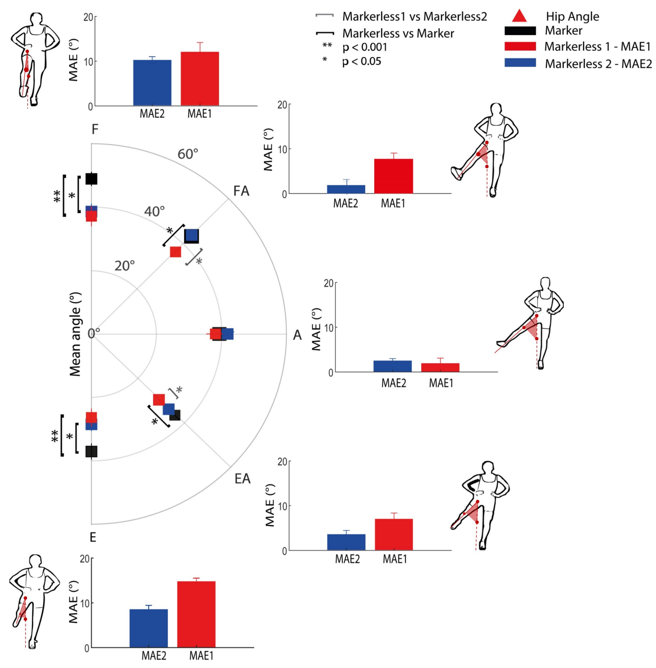


Fig. 1. The polar plot reports mean angles and standard deviations for each method and direction. Bar graphs illustrate MAE and standard deviations between each camera-based method and the gold standard.

However, in directions where depth information is crucial, the *Markerless* approaches underestimated the angle in all movements with a sagittal component (F, E, FA, and EA, Figure 1). This difference was significantly higher for the *Markerless1* compared to *Markerless2* in the combined planes (FA and EA, Figure 1), resulting in a significant error with respect to the gold standard. In contrast, for movements performed purely in the sagittal plane, perpendicular to the camera acquisition plane, both camera-based methods showed comparable significant differences from the gold standard (F and E, Figure 1). However, it's important to emphasize that both Markerless methods reported significantly higher error in detecting sagittal movements, indicating a notable limitation in accurately capturing angles for movements in the sagittal plane compared to other directions.

## IV. Conclusion

This study presents a preliminary evaluation of two distinct approaches for estimating depth information in the quantitative analysis of human motion using a single RGB-D camera. Both approaches demonstrated comparable performance in the frontal plane, with particularly low errors compared to the gold-standard. However, in other planes where depth information is crucial, such as the combined and sagittal planes, they reported more limitations, especially the method based on MediaPipe Pose, which showed higher error values. These accuracy variations highlight the need to select systems based on specific movement tracking requirements. Future steps include improving the accuracy of these systems exploring ways to integrate the two approaches, also in presence of occlusions, extending analysis to multiple joints to understand complex movements, and evaluating the applicability of these methods in clinical applications.

# References


[1] W. W. Lam, Y. M. Tang, and K. N. Fong, "A systematic review of the applications of markerless motion capture (mmc) technology for clinical measurement in rehabilitation," 2023.

[2] H. Yan, B. Hu, G. Chen, and E. Zhengyuan, "Real-time continuous human rehabilitation action recognition using openpose and fcn," 2020.

[3] A. Avogaro, F. Cunico, B. Rosenhahn, and F. Setti, "Markerless human pose estimation for biomedical applications: a survey," *Frontiers in Computer Science*, vol. 5, 7 2023.

[4] M. Moro, G. Marchesi, F. Hesse, F. Odone, and M. Casadio, "Markerless vs. marker-based gait analysis: A proof of concept study," *Sensors*, vol. 22, 3 2022.

[5] V. Bazarevsky and I. Grishchenko, "Google ai blog: On-device, real-time body pose tracking with mediapipe blazepose," 2020.

[6] T. B. de Gusmão Lafayette *et al.*, "Validation of angle estimation based on body tracking data from rgb-d and rgb cameras for biomechanical assessment," *Sensors*, vol. 23, 2023.

[7] B. Scott, M. Seyres, F. Philp, E. K. Chadwick, and D. Blana, "Healthcare applications of single camera markerless motion capture: a scoping review," 2022.