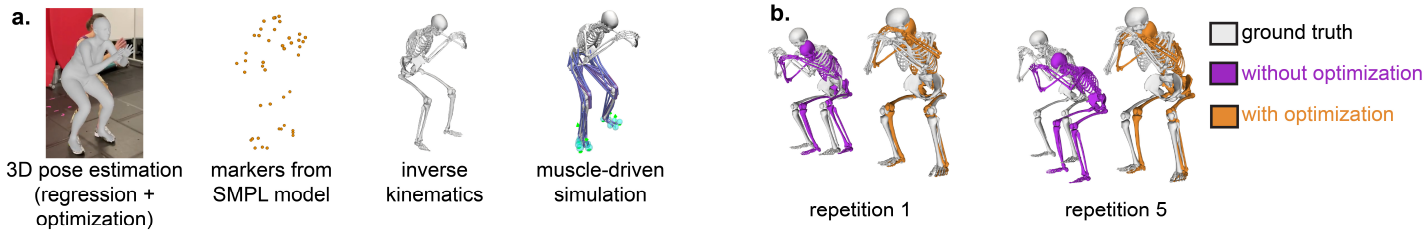
OPENCAP MONOCULAR: HUMAN MOVEMENT DYNAMICS FROM A SINGLE SMARTPHONE VIDEO

Scott D. Uhlrich[1,2*], Shardul Sapkota[1], Antoine Falisse[1], Scott L. Delp[1]
[1]Stanford University, Stanford, CA; [2]University of Utah, Salt Lake City, UT
*Corresponding author's email: suhlrich@stanford.edu

Introduction: The ability to easily measure the kinematics and kinetics of human movement could improve the prevention and treatment of neurological and musculoskeletal diseases. The time and expertise required to measure movement with marker-based motion capture has limited its adoption in the clinic and in large-scale research studies. We recently developed OpenCap, open-source software that estimates kinematics and kinetics using two smartphone videos [1]. OpenCap and other markerless motion capture technologies [2] lower the time and cost barriers to movement analysis by orders of magnitude [1]; however, they require multiple cameras. If we can achieve similar accuracy from a single video, the >4 billion smartphone owners around the world would have access to 3D motion capture. The computer vision field has made advances in estimating the global pose of a human mesh from a single video [3], but these algorithms are not designed for or evaluated on their biomedical utility. This study aims to develop and evaluate a pipeline for estimating kinematics and kinetics of common human movements (walking and squatting) from a single video.

Methods: Here we introduce OpenCap Monocular, which combines 3D human pose estimation, biomechanical modeling, and dynamic musculoskeletal simulation to estimate kinematics and kinetics from monocular video (Figure 1a). We evaluate the pipeline on walking and squatting trials with synchronously recorded video (45° from front-facing), marker-based motion capture, and force plate data [1] for one individual (33-year-old female). First, we estimate foot contact probability and the global pose of an SMPL model [4] using WHAM [3], a state-of-the-art regression-based 3D pose estimation algorithm. Using the WHAM results as the initial guess, we perform a two-stage optimization to solve for camera parameters and global human pose. The first optimization solves for camera parameters and SMPL shape parameters (i.e., anthropometry) by minimizing reprojection error assuming a fixed camera and known subject height. After fixing camera and SMPL parameters, the second optimization solves for the global SMPL pose by minimizing reprojection error, foot position variance during contact, and joint velocity. We then extract anatomical landmarks from SMPL vertices, scale a musculoskeletal model [1], and run Inverse Kinematics using OpenSim 4.4 [5]. We generate a muscle-driven dynamic simulation that tracks the kinematics to estimate ground forces [1,6]. To evaluate accuracy, we compute mean absolute differences between OpenCap Monocular and marker-based motion capture and force plates and average across trials.

Figure 1: (a) The OpenCap Monocular pipeline for estimating kinematics and kinetics from video. **(b)** Optimizing pose after regression-based pose estimation reduces drift over five repetitions of squatting (translational error reduced from 96mm to 46mm).



Results & Discussion: For walking and squatting, OpenCap Monocular had mean absolute differences from marker-based motion capture and force plates of 6.6° and 37mm (pelvis translation) for kinematics, and 7.7%bodyweight for reaction forces (Table 1). The joint angle differences during walking (5.6°) are comparable to the two-camera OpenCap system (5.2°), an eight-camera markerless motion capture system (5°) [2], and a 17-sensor IMU suit (5°) [7]; however, translational differences (27mm) are greater than two-camera OpenCap (7mm). Ground reaction force accuracy during walking (medio-lateral: 1.3% bodyweight, anterior-posterior: 4.9%, vertical: 21.5%) is worse than two-camera OpenCap (1.5–11.1% bodyweight) [1] and IMUs (1.7–9.3% bodyweight) [8], likely due to larger errors in pelvis translation. The pose optimization reduces drift from the regression-based pose estimation model (Figure 1b).

Table 1: Mean absolute differences between OpenCap and marker-based motion capture. Kinematic differences are averaged across 24 rotational and three translational degrees of freedom, and ground forces are averaged across three directions.

	Kinematics (rotations, °)	Kinematics (translations, mm)	Ground forces (% bodyweight)
Opencap Monocular	6.6	37	7.7
OpenCap two-camera	4.9	7	4.2

Significance OpenCap Monocular estimates kinematics and kinetics in two activities for a single individual with similar but slightly worse accuracy than two-camera OpenCap; further validation across individuals and activities is needed. Biomechanical analysis from a single video will dramatically expand the frequency and ease with which we can quantify movement in the home, clinic, and field. Rigorous analysis in the context of biomechanical assessments is a necessary step in this expansion.

References: [1] Uhlrich et al. (2023), *Plos Comput Biol* 19(10). [2] Kanko et al. (2021), *J Biomech* 127. [3] Shin et al. (2024), *Proc CVPR*. [4] Loper et al, (2015), *Proc SIGGRAPH Asia*. [5] Seth et al. (2018) *Plos Comput Biol* 14(7). [6] Falisse et al. (2019) *J R Soc Interface* 16. [7] Al Borno et al. (2022), *J Neuroeng Rehabil* 19(11). [8] Karatsidis et al. (2019), *Med Eng Phys* 65.