Monocular Video-to-Kinematics Pipeline for Injury Reconstruction

Wenxuan Li, Svein Kleiven, Xiaogai Li

I. INTRODUCTION

Injuries in traffic, sports, and forensic contexts remain a major public health concern, and finite element human body models (HBMs) have become important numerical tools for injury analysis. However, their use is often limited by the difficulty of obtaining 3D kinematic boundary conditions, such as initial position, orientation and velocity, from monocular real-world video. In many real-world injury cases, ordinary footage from closed-circuit television (CCTV), dashcams or mobile devices is the only record of the event. Converting such footage into 3D kinematic boundary conditions is therefore an important step toward subsequent HBM-based reconstruction.

Recovering 3D kinematics from monocular video is inherently under-constrained, because depth, out-of-plane motion, and metric scale are not directly observable from a single view. This is particularly challenging for ordinary CCTV footage, where camera parameters, scale references, image quality, motion blur, and occlusion are often uncontrolled. Prior video-informed reconstruction work has demonstrated the value of estimating impact boundary conditions from monocular footage for HBM-based injury reconstruction [1], but front-end extraction remains labor-intensive and scenario-dependent. Recent computer-vision foundation models now enable reliable video segmentation [2] and temporally consistent monocular depth estimation [3]. This study introduces a monocular video-to-kinematics pipeline that combines segmentation-based tracking, scene calibration, monocular depth estimation and temporal calibration to estimate video-derived 3D kinematic inputs in cases where instrumented data are unavailable.

II. METHODS

The proposed workflow (Fig. 1) consists of a spatial reconstruction branch and an independent temporal calibration module. The spatial branch extracts 2D target motion from monocular video using segmentation-based tracking. Segment Anything Model 2 (SAM 2) [2] is used to propagate a dense target mask throughout the sequence. Region-specific constraints are then applied to isolate the anatomical region of interest, combining appearance cues (e.g., color, intensity, texture) with geometric cues (e.g., compactness, area, centroid continuity). A frame-wise 2D centroid or representative trajectory point is then computed from the refined mask.

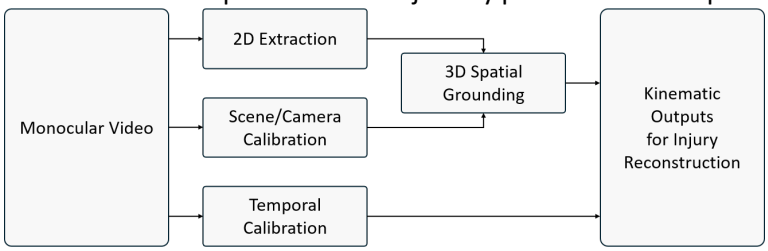

Fig. 1. Overview of the proposed monocular video-to-kinematics pipeline

For 3D spatial calibration, camera parameters are estimated from visible scene geometry using vanishing points derived from approximately orthogonal structural directions [4]. The recovered camera model is then used to project 2D observations into a 3D coordinate system aligned with the scene geometry.

Depth is estimated for each frame using Video Depth Anything (VDA) [3] and fused with the calibrated projection model. Relative depth from VDA is converted to metric depth using a global scale factor recovered from a known scene reference (e.g., a structural element with known size). The metric depth is sampled at the 2D target location and combined with the back-projected camera ray from the calibrated intrinsics to obtain a 3D position in camera coordinates, which is then transformed to the world frame via the ground-plane-aligned extrinsics. The resulting 3D trajectory is post-processed by outlier rejection based on inter-frame displacement, short-gap interpolation, and temporal smoothing. Component-wise velocity histories are derived by numerical differentiation. These outputs serve as video-derived kinematic inputs for subsequent multibody or HBM-based reconstructions.

The temporal calibration module recovers frame rate independently: from visible timestamps when available, and from external timing references otherwise.

W Li (e-mail: wenxuanl@kth.se) is a doctoral student, S Kleiven is a professor, X Li is an associate professor, all at the Division of Neuronic Engineering, Department of Biomedical Engineering and Health Systems, KTH, Sweden.

III. INITIAL FINDINGS

The pipeline was applied to a CCTV recording of a suspected infant abusive-shaking incident. A frame rate of 55 Hz was recovered from the visible timestamp, enabling time-scaled trajectory and velocity estimation. The head trajectory was extracted by restricting the propagated full-body mask to the upper body, applying a skin-color similarity cue from initial-frame head pixels, and filtering candidates by compactness and frame-to-frame centroid continuity. In the absence of manufacturer-provided camera intrinsics, a camera model was recovered from visible structural edges using vanishing-point-based scene calibration, and the floor-tile side length provided the global scale reference; metric depth was then sampled at the head centroid in each frame. After depth alignment, outliers were rejected using a median absolute deviation (MAD) threshold on inter-frame displacement, gaps of ≤ 3 frames were filled by linear interpolation, and the trajectory was smoothed with a Savitzky-Golay filter. Across the analysed sequence, 3% of centroid samples were rejected as outliers and 2% required short-gap interpolation. The pipeline produced a continuous 3D head trajectory and corresponding velocity histories (Fig. 2).

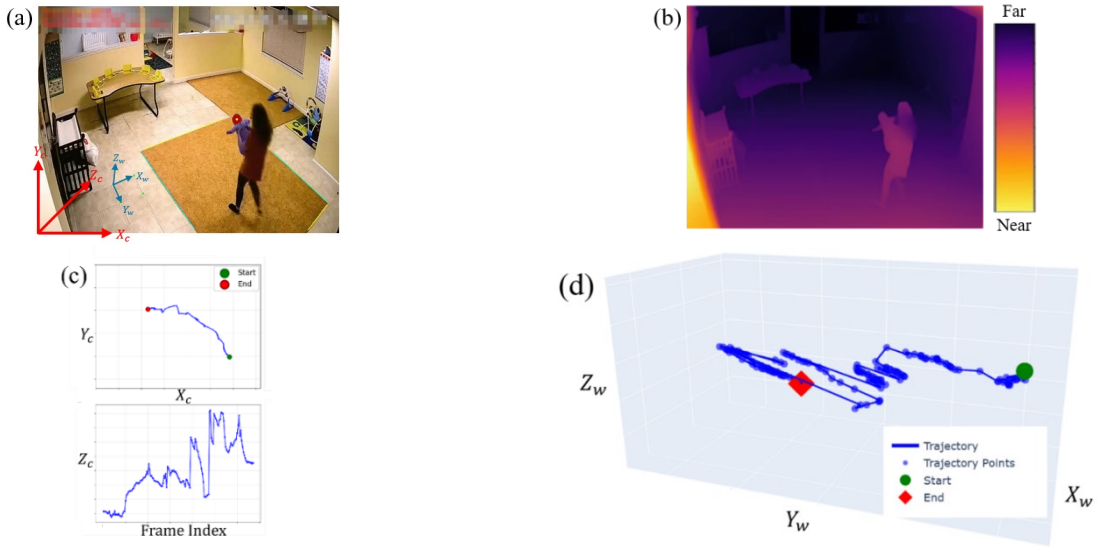

Fig. 2. Preliminary results of the proposed pipeline: (a) 2D target extraction and scene calibration; (b) relative depth map estimated from the monocular video; (c) recovered 2D trajectory on the camera image plane; and (d) post-processed 3D trajectory in the world coordinate system, after outlier rejection, short-gap interpolation, and temporal smoothing. In (a), the red coordinate system indicates the camera coordinate system, the dark blue coordinate system indicates the world coordinate system, and the yellow and light blue lines denote reference lines used for calibration.

IV. DISCUSSION

The proposed pipeline was demonstrated on a CCTV recording of a suspected infant abusive-shaking incident, producing a continuous 3D head trajectory suitable as kinematic input for injury reconstruction. Accuracy remains to be quantified against controlled experiments, and length scaling depends on visible scene geometry and manual calibration. SAM 2 propagation can also drift on long sequences under heavy occlusion or appearance change; although this was not encountered here, mask-quality monitoring and detection-based re-anchoring could stabilise tracking on longer sequences.

Future work includes controlled validation, sensitivity analyses with respect to camera viewpoint, image quality, post-processing choices, and extension to other body segments and impact scenarios. Although demonstrated on indoor CCTV, the segmentation, calibration, depth estimation and temporal calibration are not scene-specific and may extend to other ordinary-video settings, including dashcam and sports footage. Once validated, the pipeline could lower the barrier to HBM-based injury reconstruction by converting ordinary monocular video from qualitative visual evidence into video-derived 3D kinematic inputs.

V. REFERENCES

[1] Yuan Q *et al.*, Brain Multiphys, 2023.
[2] Ravi N *et al.*, arXiv, 2024.
[3] Chen S *et al.*, CVPR, 2025.
[4] Caprile B *et al.*, Int. J. Comput. Vis., 1990.