This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2026.3733600

Upper-Limb Pose Estimation from Sparse Body-Worn Visual-Inertial Sensors

Ferdinando Pompanin, Enrico Martini*, Andrea Calanca, Franco Fummi, and Nicola Bombieri

***Abstract*—Estimating the upper-body pose is essential for a wide range of applications, from rehabilitation therapy to industrial injury prevention. Current solutions face several limitations, such as heavy setup requirements, and often suffer from occlusion and drift issues. We propose a novel, sparse sensor fusion framework that estimates upper-limb kinematics using two shoulder-mounted inertial measurement units (IMUs) and a chest-mounted camera with an integrated IMU, without visual markers. The proposed pipeline integrates an orientation estimation module with a quadratic programming inverse kinematics (QPIK) solver. This formulation enables the system to enforce biomechanical constraints and utilize low-frequency visual input to correct inertial drift and recover joint positions that are unobservable by sparse IMUs alone. We evaluate the framework on the Total Capture dataset, where egocentric targets are simulated from motion-capture keypoints masked by a virtual chest-camera field of view, allowing the fusion layer to be assessed under controlled visual rate and accuracy. In this setting, our method achieves lower tracking error than adapted EKF-based visual-inertial baselines while using fewer wearable sensors. Compared with an IMU-only data-driven method, the proposed fusion pipeline achieves lower error with sufficiently frequent and accurate visual updates, demonstrating the benefit of exploiting intermittent egocentric constraints. The code is publicly available at https://github.com/PARCO-LAB/COMETH.**

***Index Terms*—Accelerometers, Biomechanics, Kalman filters, Motion capture, Motion analysis, Multimodal sensors, Multisensor systems, Sensor fusion.**

I. Introduction

Human pose estimation (HPE) aims at estimating the positions and configurations of human body limbs. This task presents several variations depending on the application context and the available sensing modalities. In particular, upper-body pose estimation enables a wide range of motion analysis applications, including rehabilitation support and injury prevention, especially in industrial environments [1].

State-of-the-art approaches [2], [3], [4] have achieved remarkable accuracy. However, each sensing technology exhibits intrinsic limitations. Fixed camera-based systems, although widely adopted, are prone to occlusions and self-occlusions and constrain the subject to a predefined capture volume [5]. Wearable egocentric cameras, commonly used in *Ego-HPE* systems, alleviate these constraints but introduce new challenges, such as severe self-occlusions and perspective distortions caused by wide-angle lenses used to maximize the field of view [6]. Inertial measurement unit (IMU)-based systems represent an alternative solution that is portable and inherently occlusion-free [7]. Zhang et al. [8] propose DynaIP, a deep learning framework that exploits spatial correlations by dividing the human body into local regions and incorporating pseudo-velocity regression. Zhu et al. [9] propose ProgIP to reconstruct full-body motion using IMUs mounted on the head and wrists, inferring leg motion from learned kinematic correlations. Deep Inertial Poser (DIP) [10] reconstructs full-body pose in real time from six body-worn IMUs using a recurrent neural network. The model is trained primarily on synthetic inertial data generated from motion-capture sequences and exploits temporal context through a sliding-window inference scheme. However, approaches based solely on IMUs suffer from integration drift, which is difficult to compensate for using magnetometers due to environmental magnetic disturbances. Von Marcard et al. [11] address this issue by performing global offline optimization over the entire motion sequence, at the cost of preventing real-time deployment. Li et al. [12] propose a system based on a chest-mounted camera and IMUs placed on the sternum, upper arm, and forearm, using ArUco markers for visual tracking. Their estimation framework relies on an extended Kalman filter (EKF) and a modified Denavit–Hartenberg (DH) representation for the upper body [13].

In this work, we propose a novel sparse sensor fusion framework for upper-body pose estimation that achieves accurate tracking using two shoulder-mounted IMUs and a chest-mounted camera with an integrated IMU. The novelty is not the use of complementary filtering or inverse kinematics in isolation, but their coupling through a QPIK fusion layer that activates inertial and visual tasks at their native rates. High-rate IMU updates constrain the model through gravity-vector residuals, whereas intermittent egocentric keypoints re-anchor the kinematic chain and recover elbow/wrist positions that are not observable from shoulder IMUs alone. QPIK also returns virtual IMU attitudes, which reset the filter baseline after each constrained update.

The contributions of this work are twofold. First, we introduce a markerless upper-limb visual-inertial pipeline compatible with a minimal wearable setup and integration in upper-limb exoskeletons. Second, we compare the proposed method on the *Total Capture* dataset [14] against two adapted EKF

This work was supported by the Veneto Region under the PR Veneto FESR 2021–2027 Programme, Action 1.1.1 Sub-action A, “Rafforzare la ricerca e l’innovazione tra imprese e organismi di ricerca” (DGR No. 729/2024), through the SUPREME project (CUP D19J24000680007; ID 24729_001692). The work was also supported in part by the high-performance computing infrastructure developed within the CONVECS project, funded under the PR Veneto FESR 2021–2027 Programme, Priority 1, Specific Objective 1.1, Action 1.1.2, and by the SHIELD project, “Sustainable, Human-centric, Innovative, Efficient Logistics through Digital Advancements” (Project ID 24729_002199).

*Corresponding author: Enrico Martini. The authors are with the Department of Engineering for Innovation Medicine, University of Verona, Verona, Italy (e-mail: name.surname@univr.it).

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2026.3733600

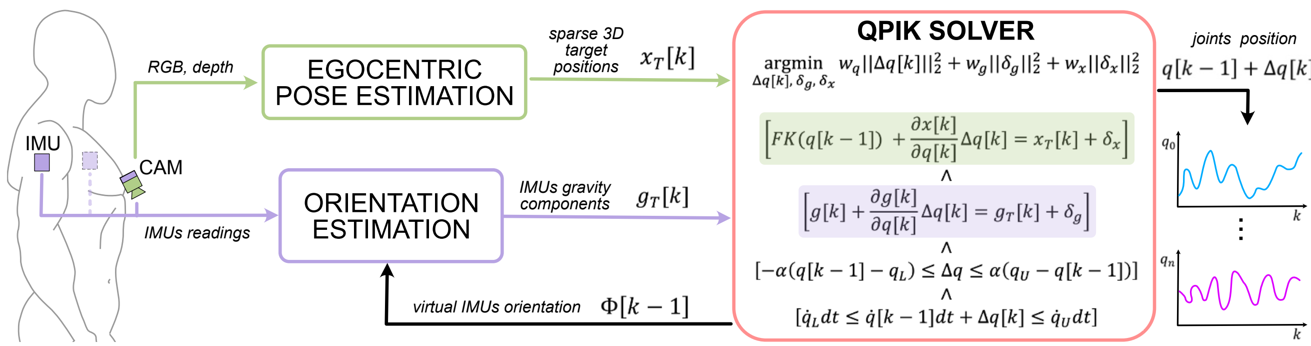

Fig. 1: Overview of the proposed multimodal sensor fusion pipeline. IMU readings are converted into gravity-vector targets, while the egocentric HPE module provides sparse 3D keypoints when they are visible. The quadratic programming inverse kinematics (QPIK) solver jointly enforces inertial, visual, and biomechanical constraints and returns joint angles and virtual IMU orientations. The latter are fed back to the orientation estimation module to reset the next prediction baseline.

baselines and DIP retrained for the same three-IMU layout. While DIP performs best in the IMU-only condition, our formulation achieves lower MPJPE with sufficiently frequent and accurate egocentric updates.

II. Method

The proposed algorithm fuses orientation data from a sparse set of IMUs with positional data from a body-mounted egocentric camera. The algorithm outputs the joint angles of the full upper-body kinematic chain. Figure 1 shows an overview of the proposed pipeline, whose individual blocks are described in detail in the following sections.

A. Orientation Estimation

We use a complementary filter [15] to estimate IMU gravity components by fusing gyroscope and accelerometer data. An IMU integrates an accelerometer, which provides a reliable estimate of the gravity direction but is sensitive to external vibrations, and a gyroscope, which measures angular velocity. We use this filter to provide gravity vector targets to QPIK. We compute the yaw ψ_G^-, pitch θ_G^-, and roll ϕ_G^- through integration of gyroscope readings, using the Euler angle rate transformation. We also calculate ϕ_A^- and θ_A^- as the projections of the accelerometer measurement a:

$$\phi_A^- = \operatorname{atan2}(a_y, a_z) \quad \theta_A^- = \operatorname{atan2}(-a_x, \sqrt{a_y^2 + a_z^2}) \tag{1}$$

The ϕ, θ, and ψ angles are obtained with:

$$\phi = \alpha\phi_G^- + (1-\alpha)\phi_A^- \ , \quad \theta = \alpha\theta_G^- + (1-\alpha)\theta_A^- \quad \psi = \psi_G^- \tag{2}$$

where α is the weighting coefficient. We adapt α using $e_a = |\|\mathbf{a}\| - 1|$: $\alpha = 1$ for $e_a > a_U$, rejecting accelerometer corrections; $\alpha = 0$ for $e_a < a_L$, resetting roll and pitch from gravity; and $\alpha = 0.95$ otherwise. Yaw remains gyroscope-integrated.

Once the orientation is estimated, for each IMU we calculate the expected accelerometer readings:

$$\tilde{g}_{\mathrm{IMU}_i} = g \begin{bmatrix} -\sin\theta \\ \cos\theta\sin\phi \\ \cos\theta\cos\phi \end{bmatrix} \tag{3}$$

where g is the gravity magnitude. The IMU gravity components are then fed into the solver.

B. Egocentric Pose Estimation

We define the positional target $x_T \in \mathbb{R}^{N\times 3}$ as the collection of 3D keypoints belonging to the camera wearer, where each row $\begin{bmatrix} x_i & y_i & z_i \end{bmatrix}$ represents the 3D position in the camera frame of a specific virtual marker $i \in [1, N]$. Depending on the vision sensor modality, the Ego-HPE node employs different strategies to estimate joint positions. When using an RGB-D camera, the node first extracts 2D keypoints from the RGB image and then projects them into 3D space using the pinhole camera model. With a fisheye RGB camera, the node directly infers 3D keypoints from the image data. Regardless of the method, the output is a set of 3D keypoints $x_T[k]$ for each time frame k.

Due to the body-worn camera's specific placement and field of view, certain joints (such as the shoulders) are permanently outside the frame. Others, such as elbows and wrists, are only intermittently visible, depending on the subject's motion. Consequently, the target position matrix $x_T[k]$ is inherently sparse.

C. Quadratic Programming Inverse Kinematics Solver

In biomechanics, kinematic models represent the human body as a series of rigid links, constraining joint positions to fixed distances to accurately simulate skeletal structure and dynamics [16]. Resolving the configuration of these articulated systems relies on inverse kinematics, which maps a desired 3D target position back to joint-space angular positions [13]. While redundant systems often employ gradient-based optimization, QPIK addresses the problem via quadratic programming, a subset of convex optimization in which a quadratic function is minimized under linear constraints [17]. QPIK is essential for applications requiring the simultaneous satisfaction of task objectives and strict constraints, such as joint limits, stability margins, and self-collision avoidance [18].

The core of the proposed methodology lies in the QPIK formulation. At each time step k, given the IMUs' gravity vectors $g_T[k]$ and sparse 3D target positions $x_T[k]$ (if available), the algorithm computes the optimal joint angular displacement $\Delta q = q[k] - q[k-1]$. The optimization problem is formulated as:

$$\underset{\Delta q, \delta_x, \delta_g}{\arg\min} \; w_q ||\Delta q||_2^2 + w_g ||\delta_g||_2^2 + w_x ||\delta_x||_2^2 \tag{4}$$

$$\text{s.t.} \begin{cases} x[k] + \frac{\partial x}{\partial q}[k]\Delta q = x_T[k] + \delta_x \\ g[k] + \frac{\partial g}{\partial q}[k]\Delta q = g_T[k] + \delta_g \\ q[k-1] - q_L \leq \Delta q \leq q_U - q[k-1] \\ \dot{q}_L \leq \dot{q}[k-1] + \frac{\Delta q}{dt} \leq \dot{q}_U \end{cases} \tag{5}$$

where $q \in \mathbb{R}^{49}$ represents the joint angles, $J_x = \frac{\partial x}{\partial q}$ and $J_g = \frac{\partial g}{\partial q}$ are the analytic Jacobians associated with the position and gravity tasks, respectively, and $\delta_x \in \mathbb{R}^{3N}, \delta_g \in \mathbb{R}^{3|\text{IMU}_s|}$ are slack variables allowing for relaxation of the task constraints. Equation (5) details the constraints required to ensure a realistic human configuration. The equality constraints enforce the differential kinematics: the current model state (position x and gravity g), when updated by the Jacobian-mapped displacement Δq, must match the targets (x_T, g_T) within a minimal error margin defined by the slack variables. The inequality constraints enforce anatomical feasibility. First, the joint positions are bounded by the range of motion (ROM) limits $q_L, q_U \in \mathbb{R}^{49}$, derived from normative biomechanical data [19]. Second, velocity constraints are imposed to ensure dynamic feasibility and to adhere to normative velocity ranges, as described in [20]. The three scalar weighting coefficients w_q, w_g, w_x normalize the different physical units and magnitudes inherent to the problem variables, ensuring numerical stability.

At initialization and after a QPIK update supported by a valid positional observation, we overwrite the complementary-filter state $\Phi = [\phi, \theta, \psi]$ with the Euler angles extracted from the corresponding virtual-IMU orientations produced by the biomechanically constrained model. The subsequent gyroscope prediction is therefore integrated from this corrected orientation. This model-based feedback also corrects the yaw component, which cannot be observed from gravity alone, without explicitly estimating or injecting a separate drift term.

III. Experiments

A. Experimental Setup

a) Dataset: We evaluate the proposed method on the *Total Capture* dataset [14], a widely used benchmark for multimodal human motion analysis. It provides synchronized 60 Hz body IMUs and marker-based motion capture for five subjects performing complex activities. The 3D positions extracted from both the compared methods and our approach are evaluated using mean per joint position error (MPJPE). Since TotalCapture does not include a real chest-mounted camera stream, we simulated the egocentric visual input by retaining only the motion-capture keypoints that would fall inside a virtual chest-mounted camera (HFOV$= 102°$, VFOV$= 57°$), tilted $75°$ from the thorax's surface normal. To test sensitivity to Ego-HPE uncertainty, we perturbed visible 3D targets with Gaussian noise $\mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma \in \{0.00, 0.02, 0.05\}m$.

Recent egocentric datasets, such as Ego-Exo4D [21] and Nymeria [22], provide richer real-world egocentric recordings, but they use viewpoints different from those of the proposed chest-mounted upper-limb setup. They are therefore promising for future validation, but require extrinsic remapping, upper-limb visibility filtering, and retraining or reannotation before enabling a direct comparison.

b) Compared Methods: Due to the lack of directly comparable methods in the literature for our specific sparse sensing configuration, we benchmarked our approach against the most closely related existing solution, derived from the work of Li et al. [12]. Their solution uses five IMUs and a chest-mounted camera with an 8-DoF upper-limb model. We extended it with two independent arm chains. The implementation described in [12], hereafter referred to as *Li2022*, adopts a uniformly accelerated motion model as the state transition model. To linearize the Jacobians of accelerometers and gyroscopes during the Kalman filter update phase, we computed numerical Jacobians. We also consider a second method, denoted *EKF (ZeroVel)*, similar to *Li2022* except that it uses an identity matrix as the state transition model. Both EKF baselines use shoulder and forearm IMUs plus thorax and wrist pose observations. We calibrated the models through grid search on a subset of the TotalCapture sequences. In addition to the two fusion pipelines, we implemented DIP [10] as a data-driven comparison method. We followed the AMASS split as in [10], and changed only the neural network input from six sensor blocks to three. At inference time, real IMU orientations were estimated with a magnetometer-free Madgwick filter [23].

c) Implementation details: For the accelerometer and gyroscope readings, we selected only the IMUs attached to the shoulders and the chest from the TotalCapture dataset. In the correction step of the orientation estimator (Equation 2), we set $\alpha = 0.95$, $a_U = 0.2$, and $a_L = 10^{-3}$. In the QPIK formulation (Equation (4)), we set the three scalar weighting coefficients $w_q = 10, w_g = 10^2, w_x = 10^3$, assigning the highest magnitude to w_x to ensure the solver prioritizes precise visual constraints over inertial data whenever available.

Parameters were selected via grid search over calibration sequences and then held fixed across subjects, actions, visual rates, and noise levels. The QPIK Jacobians are obtained from Nimble Physics [24] at the current linearization point and the QP is solved through CVXPY [25] using OSQP [26], warm-started from the previous solution. On an NVIDIA Jetson Orin Nano Super, the QPIK step averaged 9.74 ms per IMU-only update and 31.84 ms when visual updates were active.

B. Experimental Results

Table I compares the MPJPE of the two EKF baselines, DIP, and our method. Without egocentric updates, DIP achieves the lowest average error (193.8 mm), compared with 302.5 mm for our IMU-only configuration, 412.8 mm for EKF (ZeroVel), and 462.1 mm for Li2022. This result highlights the benefit of DIP's learned motion prior when pose estimation relies solely on sparse inertial measurements. With noise-free positional targets, our method surpasses DIP already at 1 Hz, reducing the average MPJPE from 193.8 mm to 163.8 mm. The error

TABLE I: Comparison on the Total Capture dataset [14] (MPJPE in mm); simulating positional targets at different rates and perturbations. Starred DIP values (*) were obtained without visual input and are repeated for reference.

Positional Target Noise			$\sigma = 0.00m$				$\sigma = 0.02m$				$\sigma = 0.05m$			
Action	**Method**	**0 Hz**	**1 Hz**	**6 Hz**	**12 Hz**	**30 Hz**	**1 Hz**	**6 Hz**	**12 Hz**	**30 Hz**	**1 Hz**	**6 Hz**	**12 Hz**	**30 Hz**
Acting	Li2022 [12]	454.0	398.5	370.5	359.4	326.2	407.1	399.0	375.6	354.5	443.3	426.6	417.7	407.7
	EKF (ZeroVel)	453.7	288.5	218.5	219.0	217.0	320.7	288.7	290.7	278.6	367.1	373.6	365.4	362.5
	DIP [10]	**188.8**	188.8*	188.8*	188.8*	188.8*	188.8*	188.8*	188.8*	188.8*	**188.8***	188.8*	188.8*	188.8*
	Ours	329.9	**173.4**	**121.3**	**110.0**	**102.9**	**187.5**	**140.3**	**130.8**	**126.4**	210.7	**184.3**	**179.6**	**181.4**
ROM	Li2022 [12]	456.3	431.9	429.9	414.6	388.4	441.9	433.5	419.5	404.3	453.3	453.8	444.1	432.3
	EKF (ZeroVel)	296.5	311.4	293.7	268.1	280.9	337.3	326.2	314.8	319.7	386.8	398.6	377.1	408.0
	DIP [10]	**218.0**	218.0*	218.0*	218.0*	218.0*	218.0*	218.0*	218.0*	218.0*	218.0*	218.0*	218.0*	218.0*
	Ours	262.6	**147.3**	**113.7**	**110.0**	**100.8**	**153.2**	**135.0**	**134.3**	**130.8**	**204.4**	**184.4**	**190.3**	**197.3**
Walking	Li2022 [12]	475.9	428.1	406.5	392.8	360.2	439.4	416.3	408.8	383.0	456.0	447.4	438.2	422.6
	EKF (ZeroVel)	488.2	313.5	251.4	248.5	244.5	349.1	315.6	297.5	292.3	387.9	374.5	371.2	364.1
	DIP [10]	**174.7**	174.7*	174.7*	174.7*	174.7*	**174.7***	174.7*	174.7*	174.7*	**174.7***	174.7*	174.7*	174.7*
	Ours	315.1	**170.7**	**98.7**	**89.1**	**76.6**	180.3	**114.5**	**106.8**	**96.4**	208.1	**155.8**	**152.6**	**151.4**
Average	Li2022 [12]	462.1	419.5	402.3	388.9	358.3	429.5	416.3	401.3	380.6	450.9	442.6	433.3	420.9
	EKF (ZeroVel)	412.8	304.5	254.5	245.2	247.5	335.7	310.2	301.0	296.9	380.6	382.2	371.2	378.2
	DIP [10]	**193.8**	193.8*	193.8*	193.8*	193.8*	193.8*	193.8*	193.8*	193.8*	**193.8***	193.8*	193.8*	193.8*
	Ours	302.5	**163.8**	**111.2**	**103.0**	**93.4**	**173.7**	**129.9**	**124.0**	**117.9**	207.7	**174.8**	**174.2**	**176.7**

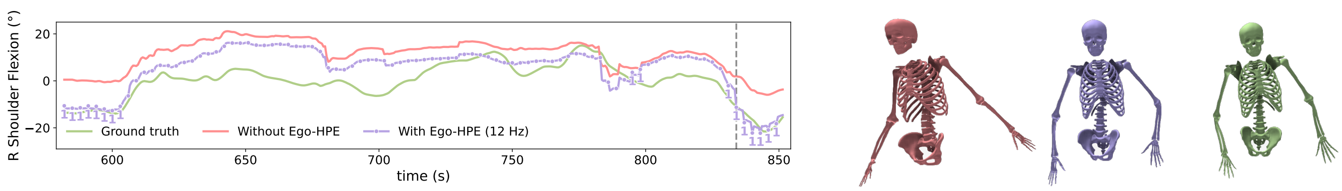

Fig. 2: Qualitative evaluation of right-shoulder flexion over a challenging motion sequence, comparing our method with 12 Hz egocentric updates (purple), its IMU-only configuration (red), and the ground truth (green). The right panel shows the corresponding 3D poses at the instant marked by the gray dot in the temporal plot. Numbers below the purple curve indicate the visible simulated keypoints supplied to the fusion algorithm. The visual-inertial estimate remains closer to the ground truth by limiting the drift observed in the IMU-only trajectory.

then decreases to 111.2, 103.0, and 93.4 mm at 6, 12, and 30 Hz, respectively. The advantage is particularly evident for dynamic motions: at 30 Hz, our method obtains 76.6 mm for Walking and 100.8 mm for ROM, compared with 174.7 mm and 218.0 mm for DIP, respectively. These results show that even intermittent positional constraints compensate for inertial drift more effectively than the learned inertial prior alone. At $\sigma = 0.02$ m, our method remains below DIP at every update rate, reducing the average MPJPE from 173.7 mm at 1 Hz to 117.9 mm at 30 Hz. Under the larger perturbation $\sigma = 0.05$ m, DIP is better at 1 Hz (193.8 mm versus 207.7 mm), but our method regains the advantage from 6 Hz onward, reaching 176.7 mm at 30 Hz. Thus, even with highly noisy positional measurements, our fusion method can outperform the IMU-only configuration when visual updates are sufficiently frequent.

Figure 2 shows the impact of the sensor fusion on the right shoulder flexion over a challenging motion sequence (`s1_acting1`). The trajectory generated without Ego-HPE highlights the inherent limitations of sparse inertial tracking, as the estimate suffers from progressive integration drift. This unbounded accumulation of error confirms that, in the absence of distal IMUs or magnetometers, the sparse three-IMU configuration cannot maintain a reliable heading reference over prolonged periods. In contrast, the 12 Hz visual-update condition substantially reduces drift. The QPIK formulation successfully leverages the sparse, intermittent visual inputs to anchor the global orientation of the limbs. When a keypoint is available, the solver enforces the geometric constraint, "pulling" the drifting inertial estimate back towards the true pose. As a result, the trajectory stays closer to the ground truth over the sequence.

IV. Conclusion

We presented a sparse visual-inertial framework for upper-body pose estimation using two shoulder-mounted IMUs and a chest-mounted camera with an integrated IMU. QPIK fuses high-rate IMU gravity targets with intermittent egocentric keypoints, subject to joint and velocity limits. The formulation is not tied to a dataset or to a specific vision module: other upper-limb wearable settings can be handled by changing the sensor extrinsics, visibility mask, and kinematic model, while additional cameras, hand trackers, or IMUs can be added as extra QPIK tasks. On Total Capture with simulated egocentric keypoints, the method reduced MPJPE relative to the adapted EKF baselines. DIP performed best in the IMU-only condition, whereas our method achieved lower error when visual updates were sufficiently frequent and accurate.

Future work will validate the pipeline on real chest-mounted egocentric recordings and recent egocentric datasets after the required viewpoint adaptation.

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2026.3733600

References

[1] L. Geretti, S. Centomo, M. Boldo, E. Martini, N. Bombieri, D. Quaglia, and T. Villa, "Collision prediction using plan learning in mixed human-robot work cells," *Robotics and Autonomous Systems*, p. 105292, 2025.

[2] A. Zhu, J. Li, and C. Lu, "Pseudo view representation learning for monocular rgb-d human pose and shape estimation," *IEEE Signal Processing Letters*, vol. 29, pp. 712–716, 2021.

[3] J. Liu, J. Rojas, Y. Li, Z. Liang, Y. Guan, N. Xi, and H. Zhu, "A graph attention spatio-temporal convolutional network for 3d human pose estimation in video," in *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 3374–3380, IEEE, 2021.

[4] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM computing surveys*, vol. 56, no. 1, pp. 1–37, 2023.

[5] S. Cao, C. Cao, and W. Zhao, "St-cfnet: Spatio-temporal cross-feature fusion networks for 3d human pose estimation," *IEEE Signal Processing Letters*, 2025.

[6] S. Chi, P.-H. Huang, E. Sachdeva, H. Ma, K. Ramani, and K. Lee, "Estimating ego-body pose from doubly sparse egocentric video data," *Advances in Neural Information Processing Systems*, vol. 37, pp. 55178–55203, 2024.

[7] T. L. Baldi, F. Farina, A. Garulli, A. Giannitrapani, and D. Prattichizzo, "Upper body pose estimation using wearable inertial sensors and multiplicative kalman filter," *IEEE Sensors Journal*, vol. 20, no. 1, pp. 492–500, 2019.

[8] Y. Zhang, S. Xia, L. Chu, J. Yang, Q. Wu, and L. Pei, "Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1889–1899, 2024.

[9] Z. Zhu, Y. Zhao, Y. Hu, G. Wang, H. Qiu, B. Zheng, C. Yan, and F. Xu, "Progressive inertial poser: Progressive real-time kinematic chain estimation for 3d full-body pose from three imu sensors," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[10] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, "Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time," *ACM Transactions On Graphics (TOG)*, vol. 37, no. 6, pp. 1–15, 2018.

[11] T. Von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, "Sparse inertial poser: Automatic 3d human pose estimation from sparse imus," in *Computer graphics forum*, vol. 36, pp. 349–360, Wiley Online Library, 2017.

[12] T. Li, X. Wu, H. Dong, and H. Yu, "Estimation of Upper Limb Kinematics with a Magnetometer-Free Egocentric Visual-Inertial System," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 1668–1674, 2022.

[13] B. Siciliano, O. Khatib, and T. Kröger, *Springer handbook of robotics*, vol. 200. Springer, 2008.

[14] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, "Total capture: 3d human pose estimation fusing video and inertial sensors," in *Proceedings of 28th British Machine Vision Conference*, pp. 1–13, 2017.

[15] P. Narkhede, S. Poddar, R. Walambe, G. Ghinea, and K. Kotecha, "Cascaded complementary filter architecture for sensor fusion in attitude estimation," *Sensors*, vol. 21, no. 6, p. 1937, 2021.

[16] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman, and D. G. Thelen, "Opensim: open-source software to create and analyze dynamic simulations of movement," *IEEE transactions on biomedical engineering*, vol. 54, no. 11, pp. 1940–1950, 2007.

[17] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[18] S. S. Mirrazavi Salehian, N. Figueroa, and A. Billard, "A unified framework for coordinated multi-arm motion planning," *Int. J. Rob. Res.*, vol. 37, no. 10, pp. 1205–1232, 2018.

[19] E. L. Zwerus, N. W. Willigenburg, V. A. Scholtes, M. P. Somford, D. Eygendaal, and M. P. van den Bekerom, "Normative values and affecting factors for the elbow range of motion," *Shoulder & elbow*, vol. 11, no. 3, pp. 215–224, 2019.

[20] E. Martini, H. J. Choi, N. Figueroa, and N. Bombieri, "Cometh: Convex optimization for multiview estimation and tracking of humans," *Expert Systems with Applications*, vol. 314, p. 131728, 2026.

[21] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, *et al.*, "Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

[22] L. Ma, Y. Ye, F. Hong, V. Guzov, Y. Jiang, R. Postyeni, L. Pesqueira, A. Gamino, V. Baiyya, H. J. Kim, *et al.*, "Nymeria: A massive collection of multimodal egocentric daily motion in the wild," in *European Conference on Computer Vision*, pp. 445–465, Springer, 2024.

[23] S. O. Madgwick *et al.*, "An efficient orientation filter for inertial and inertial/magnetic sensor arrays," 2010.

[24] K. Werling, D. Omens, J. Lee, I. Exarchos, and K. Liu, "Fast and feature-complete differentiable physics engine for articulated rigid bodies with contact constraints," in *Robotics: Science and Systems XVII*, RSS2021, Robotics: Science and Systems Foundation, July 2021.

[25] S. Diamond and S. Boyd, "Cvxpy: A python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[26] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd, "OSQP: an operator splitting solver for quadratic programs," *Mathematical Programming Computation*, vol. 12, no. 4, pp. 637–672, 2020.