Real-time Full-Body Motion Capture from Video and IMUs

Charles Malleson, Marco Volino, Andrew Gilbert, Matthew Trumble, John Collomosse and Adrian Hilton

CVSSP, University of Surrey, Guildford, U.K.

Overview

A real-time full-body motion capture system is presented which uses input from a sparse set of inertial measurement units (IMUs) along with images from two or more standard video cameras and requires no optical markers or specialized infra-red cameras. A real-time optimization framework incorporates constraints from the IMUs, cameras and a prior pose model. The combination of video and IMU data allows the full 6-DOF motion to be recovered including axial rotation of limbs and drift-free global position. Tests on indoor and outdoor captured data show the effectiveness of the approach for tracking a wide range of human motion in real time.

Motivation

- Applications in entertainment (film, TV, games, VR, AR), life sciences
- Real-time full-3D kinematic motion capture with low encumbrance, flexible capture configurations
- Overcoming limitations of previous methods

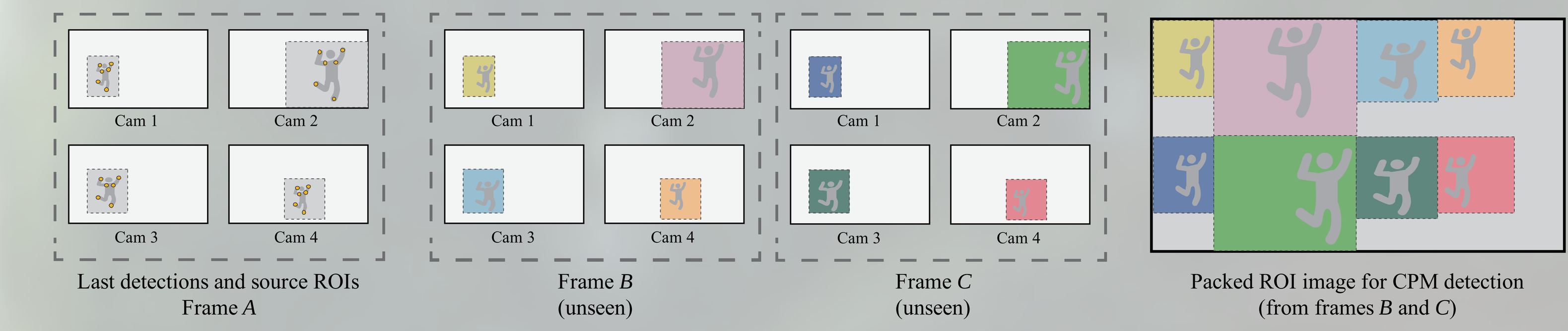
Features / Approach	Optical [4]	IMU [13]	Kinect	Andrews 2016 [6]	SIP [18]	CPM [19]	Vnect [12]	Trumble 2017[16]	Ours
Realtime, online (video rates)	✓	✓	✓	✓		✓	✓	✓	✓
Outputs full 6DOF motion (incl. axial rotation)	✓	✓	✓	✓	✓		✓		✓
Outputs unambiguous 3D global position	✓		✓	✓					✓
Kinematic skeleton for animation	✓	✓	✓	✓	✓		✓		✓
Dynamic lighting and background	✓	✓	✓	✓	✓	✓	✓		✓
Outdoor		✓			✓	✓	✓	✓	✓
Robust to heavy occlusion		✓		✓	✓				✓
Long range (> 5m)	✓	✓		✓	✓	✓	✓	✓	✓
Marker-less		✓	✓		✓	✓	✓	✓	✓
Subject fully unencumbered			✓			✓	✓		

Flexible input configuration

The method can trade accuracy against capture hardware and setup complexity (number of cameras and IMUs) as well as framerate. In the experiments, IMU input was from Xsens MTWs [13], while 2D joint detections are obtained from a state-of-the art convolutional pose machine (CPM) detector [19].

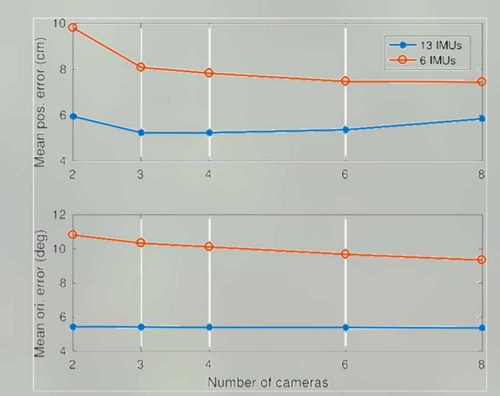
Frame packing for increased detection throughput

The 2D joint detection is a bottleneck (requiring > 150 ms per image on our system). In order to maintain video rate operation while detecting on multiple camera views, we pack regions of interest from several input images into a single image for detection and resolve the results to their originating frame and camera.



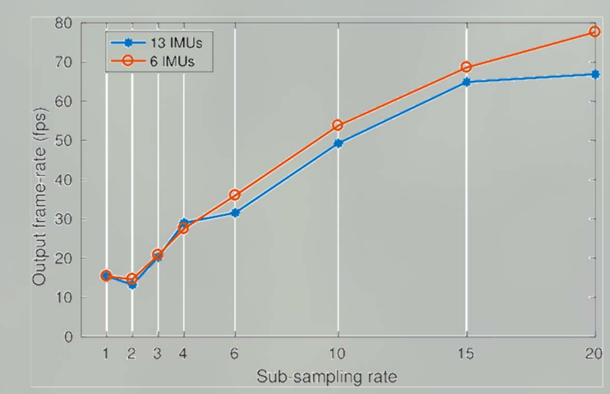
Number of cameras

Between 2 and 8 cameras were used (multiple cameras constrain depth)



Positional subsampling

To increase the output framerate, 2D detection may be performed on a subset of input frames



Comparison of configurations

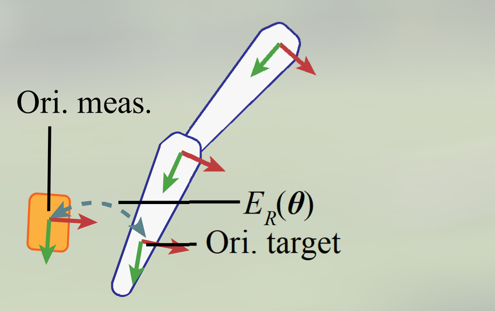
Quantitative evaluation on the *Total Capture* dataset [16] using A full set of 13 IMUs or a sparse set of 6 IMUs with HQ ('high quality', 8 cam, detection on all frames) and HS ('high speed', 4 cam, detection on 2/8 frames).

	S1 FS3	S2 FS1	S2 RM3	S3 FS1	S3 FS3	S4 FS3	S5 A3	S5 FS1	Mean
Pos. error (cm)									
Ours, 13 IMU, HQ	**7.4**	**5.3**	**3.9**	**6.7**	**6.7**	**6.4**	**6.4**	**7.0**	**6.2**
Trumble [16]	9.4	16.7	9.3	13.6	8.6	11.6	14.0	10.5	11.7
Ours, 13 IMU, HS	8.5	5.4	3.8	7.4	7.3	7.7	6.6	7.5	6.8
Ours, 6 IMU, HQ	9.8	7.1	6.6	10.0	10.7	9.2	9.0	10.0	9.1
Ours, 6 IMU, HS	14.3	9.4	10.8	19.4	17.1	13.9	13.3	16.5	14.3
Ori. error (deg)									
Ours, 13 IMU, HQ	11.2	5.1	5.0	8.3	9.3	8.0	7.6	8.2	7.8
Ours, 13 IMU, HS	11.2	5.1	5.0	8.3	9.3	8.0	7.6	8.2	7.8
Ours, 6 IMU, HQ	16.3	9.2	8.7	13.2	15.7	13.0	11.8	12.1	12.5
Ours, 6 IMU, HS	18.3	10.9	10.6	16.2	19.7	14.8	14.3	15.1	15.0

Cost function

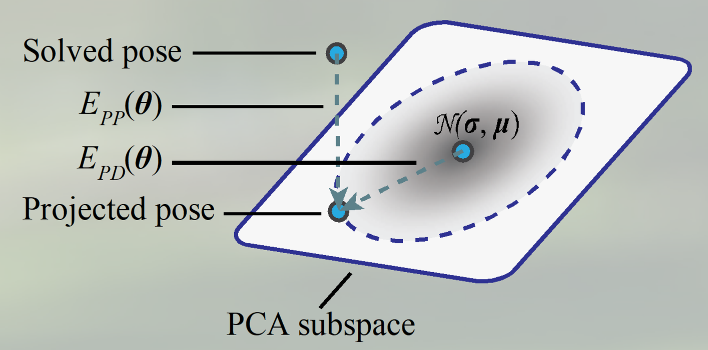
The solver combines input from IMUs (orientation and acceleration) and video (2D joint detections) with a pose prior to produce accurate and robust kinematic pose estimates for each frame.

$$E(\boldsymbol{\theta}) = \overbrace{E_R(\boldsymbol{\theta}) + E_P(\boldsymbol{\theta}) + E_A(\boldsymbol{\theta})}^{Data} + \overbrace{E_{PP}(\boldsymbol{\theta}) + E_{PD}(\boldsymbol{\theta})}^{Prior}$$



Orientation term

Minimizes the difference between the solved and measured orientation of the body segment

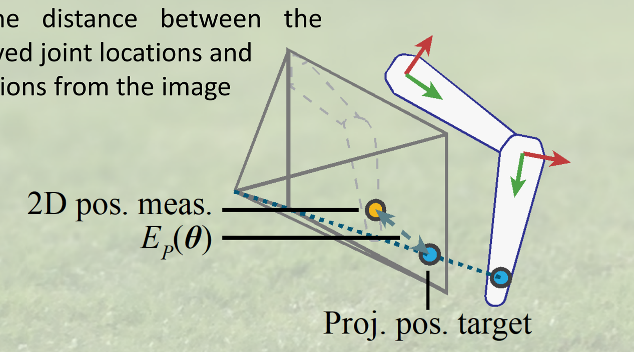


PCA prior projection term

Encourages the pose to lie close to a reduced dimensionality subspace of prior observed poses (soft joint type enforcement)

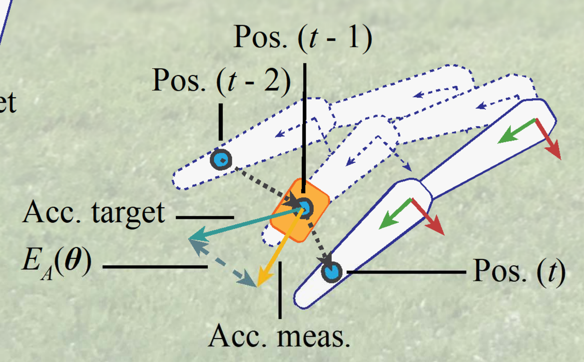
2D Position term

Minimizes the distance between the projected solved joint locations and the 2D detections from the image



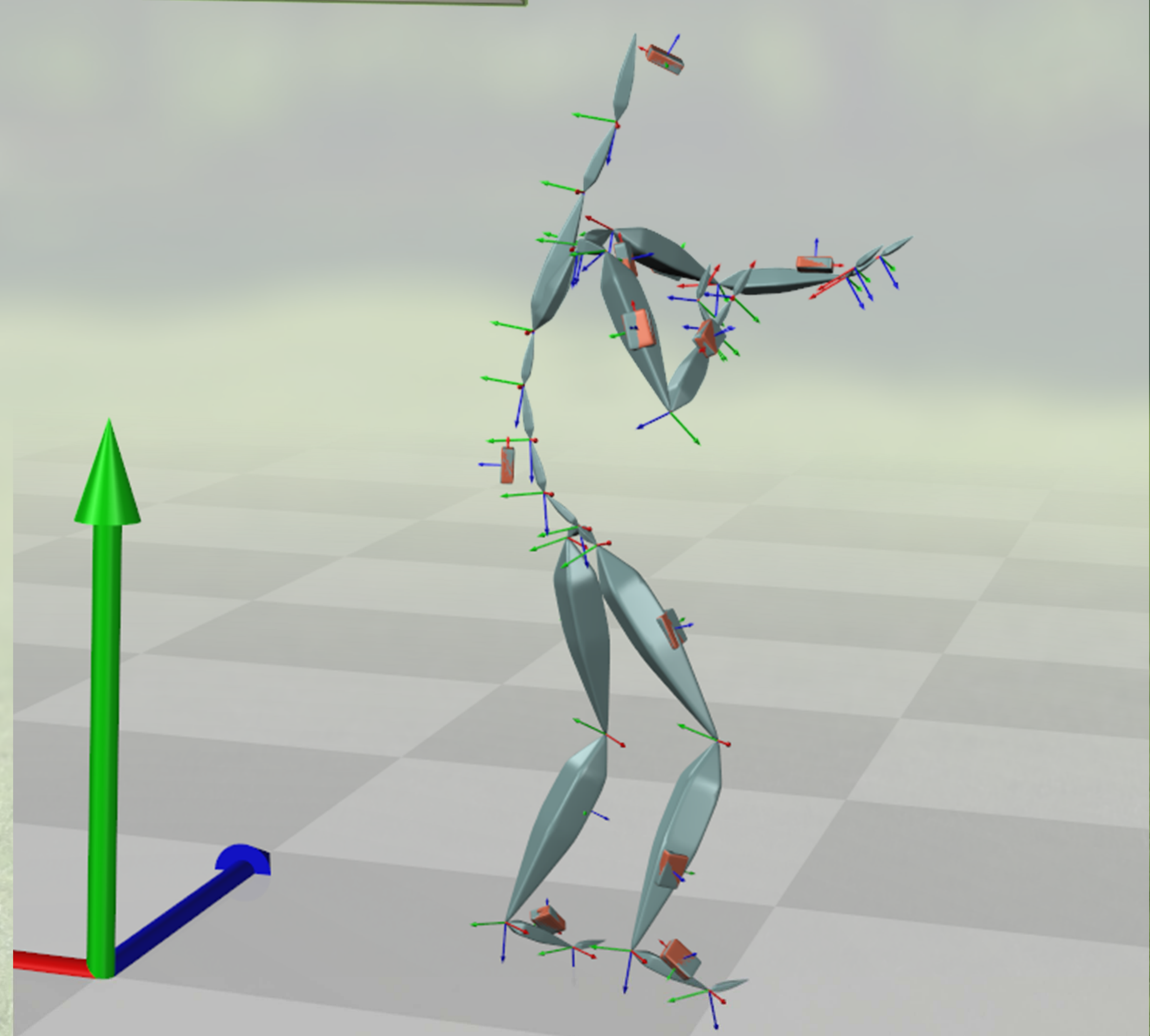
PCA prior deviation term

Discourages deviation beyond the prior observed range of motion (soft joint limit)



Acceleration term

Minimizes the difference between the solved and measured acceleration at the IMU site

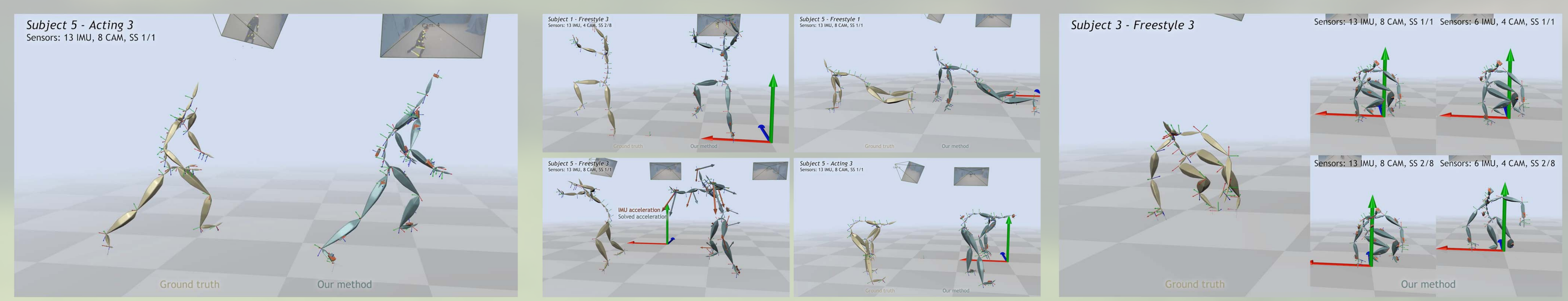


Solved skeletal pose showing IMU placement and 2D joint detections in one of the camera views

Results

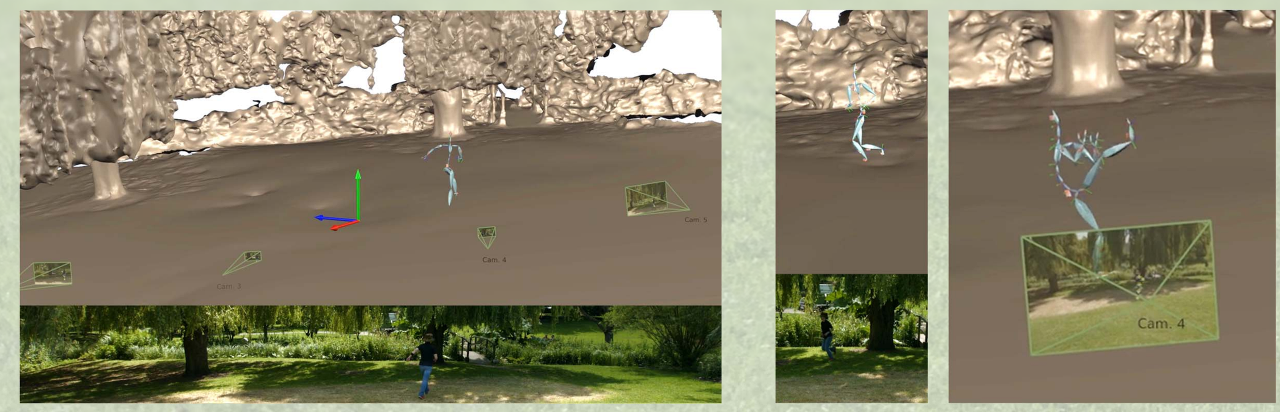
Indoor capture

The *Total Capture* indoor motion capture dataset [16] has 360 camera coverage and 13 IMUs. It also contains a marker based 'ground truth' reference motion for evaluation.

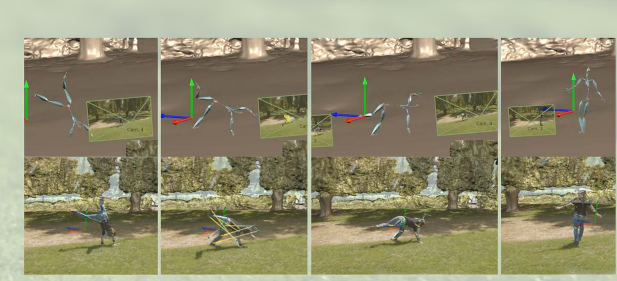


Outdoor capture

A an outdoor dataset *Outdoor 1* was captured with cameras in a 120 degree arc along with 12 IMUs.



Large, uncontrolled capture area



Acrobatic motion (note robustness to CPM mis-detections)

Prop interaction

References

[4] Vicon Blade. http://www.vicon.com

[6] S. Andrews, I. Huerta, T. Komura, L. Sigal. and K. Mitchell Real-time Physics-based Motion Capture with Sparse Sensors. CVMP2016

[13] D. Roetenberg, H. Luinge, and P. Slycke. Xsens MVN: Full 6DOF Human Motion Tracking Using Miniature Inertial Sensors. Technical report, pages 1–7, 2013

[16] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. BMVC 2017

[18] T. von Marcard, B. Rosenhahn, M. Black, and G. Pons-Moll. Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs. Eurographics 2017

[19] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. CVPR 2016

Acknowledgements: This work was supported by the Innovate UK Total Capture project (grant 102685) and in part by the EU H2020 Visual Media project (grant 687800). We wish to thank Anna Korzeniowska, EvrenImre, Joao Regateiro and Armin Mustafa for their help with data capture.

UNIVERSITY OF SURREY

CVSSP
Centre for Vision, Speech and Signal Processing

3DV 2017