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

3DV 2017

Motivation

Motivation

Realtime, unconstrained motion capture

- Numerous **applications** in entertainment (film, TV, games, VR, AR) and life sciences
- Existing approaches typically place many **restrictions** on the capture setting or offer limited accuracy
- Goal: real-time, full-3D kinematic motion capture with low encumbrance, **flexible** capture configurations

Traditional IR marker-based approach

Our approach

Motivation

Overcoming limitations of previous methods

- Our method: **high fidelity**, full skeletal solve in **realtime**, with **modest hardware requirements**, low encumbrance and **flexible** capture environments

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			✓			✓	✓		

Hybrid video and IMU solution

Realtime, unconstrained motion capture

- Combining complementary input modalities, multiple-view **video** and **IMU**s
  - Full **6DOF kinematic skeleton** solve suitable for character **animation** (axial rotation recovered from IMU input)
  - Drift-free **global 3D position** without depth ambiguity (multiple-view video)
  - Indoor or **outdoor, uncontrolled conditions**, e.g. moving background, changing illumination, heavy occlusion (no silhouettes, visual hulls or appearance consistency)
  - Minimal incumbrance (**no markers**, only a few IMUs)
  - **Flexible** hardware configuration (number of cameras and IMUs)
  - **Realtime, online** operation at video rates (efficient per-frame pose optimization rather than batch processing)

Approach

Approach

Data sources

- Inertial measurements
    - Xsens MTw **IMUs**, worn on body
        - Orientation
        - Acceleration
- 2D keypoint detections
    - Standard **video** input (no optical markers or IR cameras)
    - State-of-the art convolutional pose machine (**CPM**) detector [19]
        - Labelled keypoint (joint) position estimates
        - Detection confidences

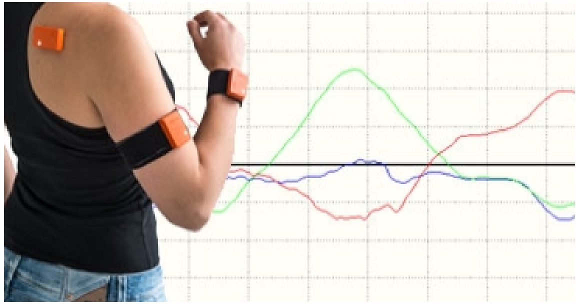


Image: www.xsens.com

Image: [19]

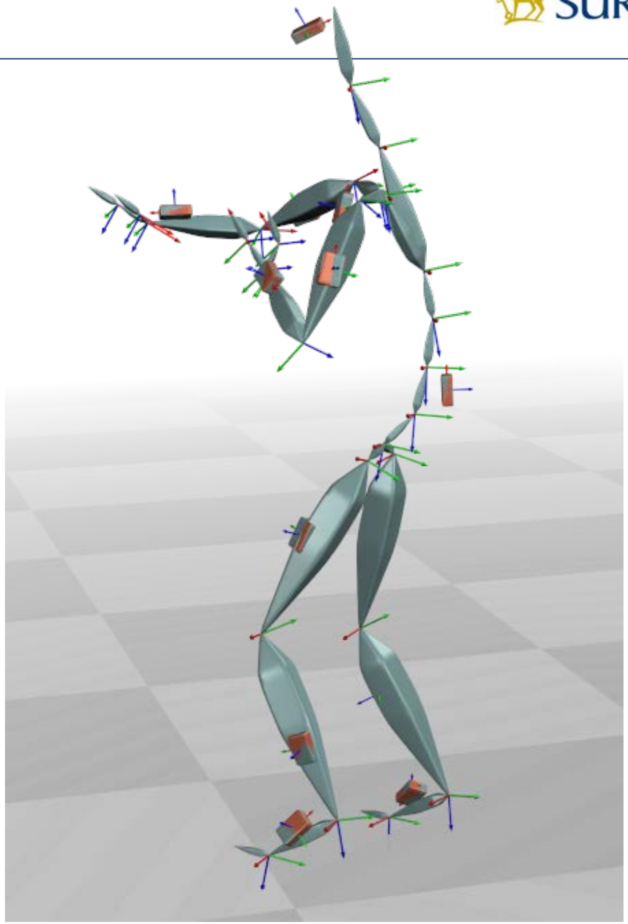
Overview

Hybrid kinematic solver using video and IMU input

- **Kinematic skeleton**, parameterised by a 66D pose vector $\boldsymbol{\theta}$ containing:
    - Root translation (3D)
    - Root orientation (3D)
    - Joint rotations (3 x 20 non-root bones)
- Bone positions and orientations determined from parameter vector by forward kinematics:

$$\mathbf{T}_b^g(\boldsymbol{\theta}) = \prod_{b' \in \mathcal{P}(b)} \left[\begin{array}{c|c} \underbrace{\mathbf{R}_{b'}}_{\text{Joint rotation}} & \underbrace{\mathbf{t}_{\mathbf{b}'}}_{\text{Bone offset}} \\ \hline 0 & 1 \end{array} \right]$$

- Minimization of a **cost function** yields the optimal parameter vector for each frame



Cost function

Overview

- Cost function to optimize pose parameter vector $\boldsymbol{\theta}$ based on sum of terms
- Optimized using **non-linear least squares** [5], initializing each frame with the previous 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 (IMUs)

Position (images)

Acceleration (IMUs)

PCA projection

PCA deviation

Cost function

Orientation terms

- Minimize **relative orientation** between body segment and measured IMU orientation

IMU global reference frame

Raw IMU ori.

IMU bone calibration

$$\mathbf{R}^{g}_{b_i} = \mathbf{R}_{ig} \cdot \mathbf{R}_{i} \cdot (\mathbf{R}_{ib})^{-1}$$

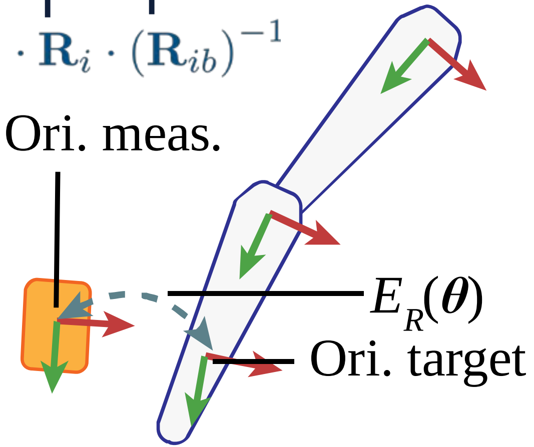
For all IMUs

$$E_R(\boldsymbol{\theta}) = \sum_{i \in [1, n_i]} \rho_R \left(\lambda_R \left\| \boldsymbol{\psi}\left((\hat{\mathbf{R}}^{g}_{b_i})^{-1} \mathbf{R}^{g}_{b_i} \right) \right\|_2^2 \right)$$

Loss function

Term weighting

Angle operator



$$\hat{\mathbf{R}}^{g}_{b} = \prod_{b' \in \mathcal{P}(b_i)} \mathbf{R}_{b'}.$$

Cost function

Position terms

- Minimize the distance between the **projected solved keypoint** locations and the 2D keypoint **detections**

For all cameras

For all keypoints

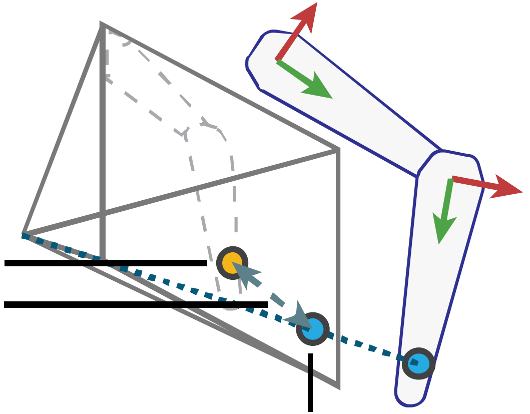
$$E_P(\boldsymbol{\theta}) = \sum_{c\in[1,n_c]} \sum_{p\in[1,n_p]} \rho_P\left(\lambda_P c_p^c \|\hat{\mathbf{t}}_p^c - \mathbf{t}_p^c\|_2^2\right)$$

Robust Cauchy loss function

$$\rho(x) = \log(1+x).$$

Term weighting

Detection confidence



$$\hat{\mathbf{t}}_p^g = \boldsymbol{\tau}_t\left(\boldsymbol{\tau}_T(\mathbf{t}_{pb}) \cdot \mathbf{T}_{b_p}^g\right)$$

$$\hat{\mathbf{t}}_p^c = \mathrm{dh}(\mathbf{P}_c \hat{\mathbf{t}}_p^g)$$

Cost function

Acceleration terms

- Minimize the difference between the **solved** and **measured acceleration** at each IMU site

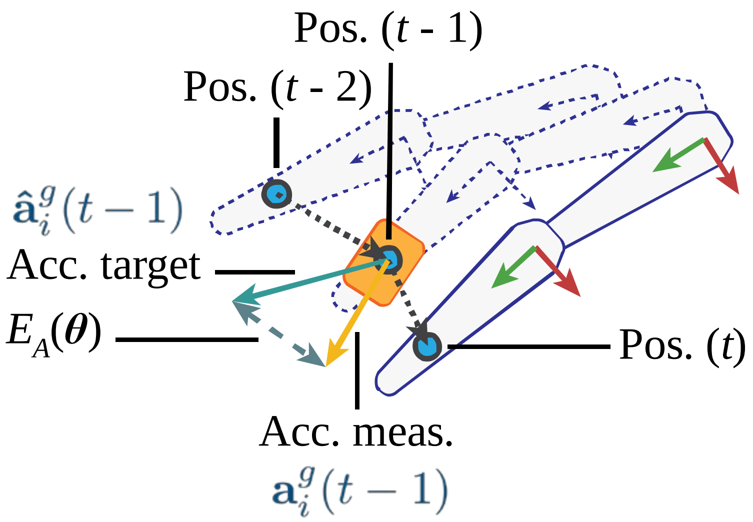
$$\hat{\mathbf{a}}_i^g(t-1) = \left(\hat{\mathbf{t}}_i^g(t) - 2\hat{\mathbf{t}}_i^g(t-1) + \hat{\mathbf{t}}_i^g(t-2)\right)/(\Delta t)^2$$

For all IMUs

$$E_A(\boldsymbol{\theta}) = \sum_{i\in[1,n_i]} \rho_A\left(\lambda_A \left\|\hat{\mathbf{a}}_i^g - \mathbf{a}_i^g\right\|_2^2\right)$$

Loss function

Term weighting



IMU orientation

IMU global ref. frame

Raw IMU acc.

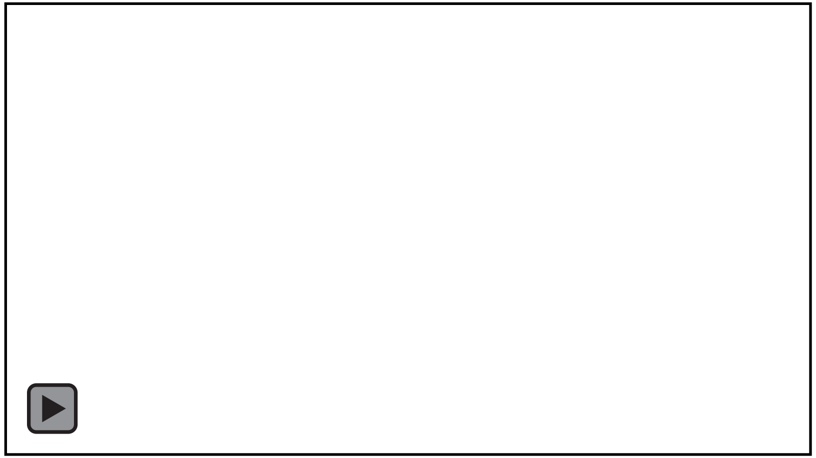
Gravity

$$\mathbf{a}_i^g(t-1) = \mathbf{R}_{ig} \cdot \mathbf{R}_i(t-1) \cdot \mathbf{a}_i(t-1) - \mathbf{a}_g$$

Cost function

Pose prior terms

- The skeletal pose is not fully constrained by position and orientation data alone
- Prior terms are needed to encourage **plausible poses** (e.g. of the spine)
- **PCA** model from prior pose database
    - DOF excluding root joint – invariance to position and heading
    - *k*-means clustering to avoid over-representation of common poses
    - 95% of the variance, dimensionality from 60 to 23



Visualization of pose principal components

Cost function

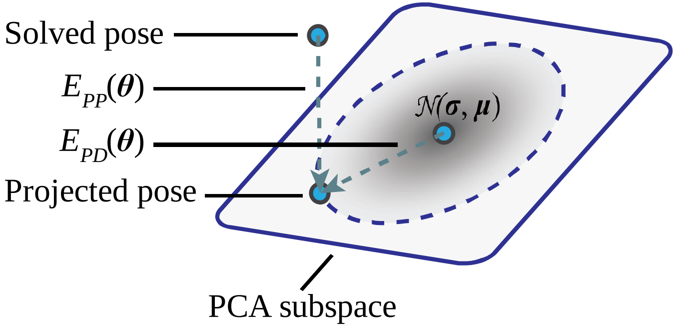
Pose prior terms

- PCA **projection** prior - encourages the pose to lie close to a subspace of prior observed poses (*soft* **dimensionality reduction**)

$$E_{PP}(\boldsymbol{\theta}) = \rho_{PP}\left(\lambda_{PP}\left\|(\bar{\boldsymbol{\theta}}-\boldsymbol{\mu})-\mathbf{M}\mathbf{M}^T(\bar{\boldsymbol{\theta}}-\boldsymbol{\mu})\right\|_2^2\right)$$

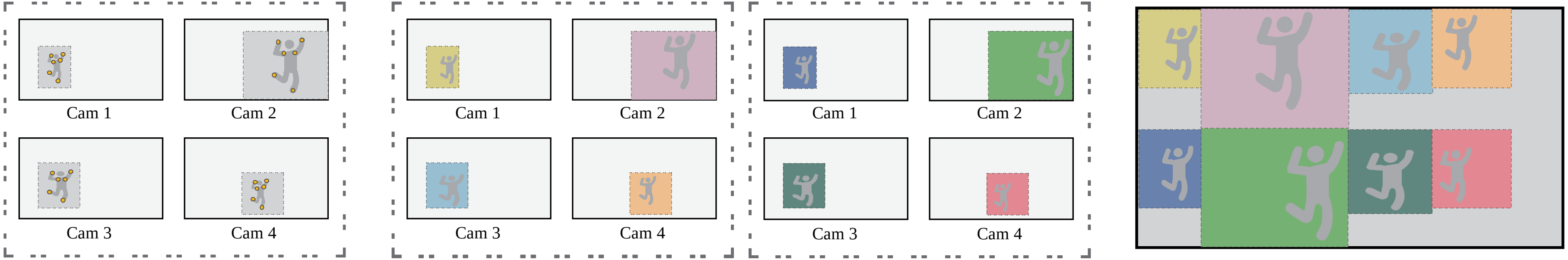
- PCA **deviation** prior - discourages deviation beyond the prior observed range of motion (*soft* **joint limit**)

$$E_{PD}(\boldsymbol{\theta}) = \rho_{PD}\left(\lambda_{PD}\left\|\operatorname{diag}(\boldsymbol{\sigma})^{-1}\mathbf{M}^T(\bar{\boldsymbol{\theta}}-\boldsymbol{\mu})\right\|_2^2\right)$$



Increasing 2D detection throughput

- The **CPM** keypoint detection [19] is a **bottleneck** (requiring > 150 ms per image)
- Aim to achieve video rate operation while detecting on **multiple** camera views
- CPM detector – detect **multiple people** in a **single image**
- Solution: **pack regions of interest** from several input images into a single image for detection, then **resolve** to originating frame and camera
- **8x increase** in throughput



Last detections and source ROIs Frame *A* | Frame *B* (unseen) | Frame *C* (unseen) | Packed ROI image for CPM detection (from frames *B* and *C*)

Results

Results

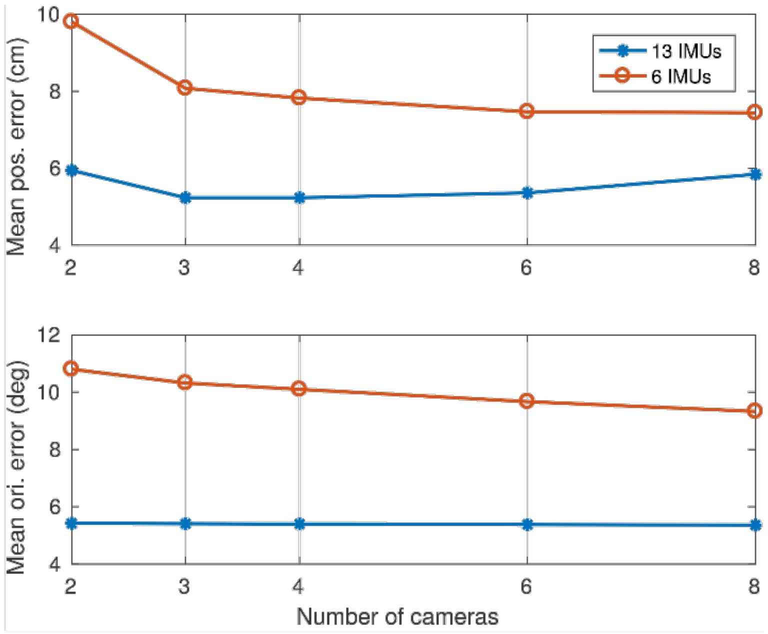
Overview

- **Quantitative** evaluation on **indoor** data (*Total Capture* dataset [16])
    - Number of cameras
    - Subsampling of 2D detections
    - Number of IMUs
        - 13 IMUs – head, upper/lower back, upper/lower limbs and feet
        - 6 IMUs – head, lower back, lower limbs (sparse)
    - Ablation study
- **Qualitative** evaluation on **outdoor** data, captured in **uncontrolled** conditions

Input configuration

Number of cameras

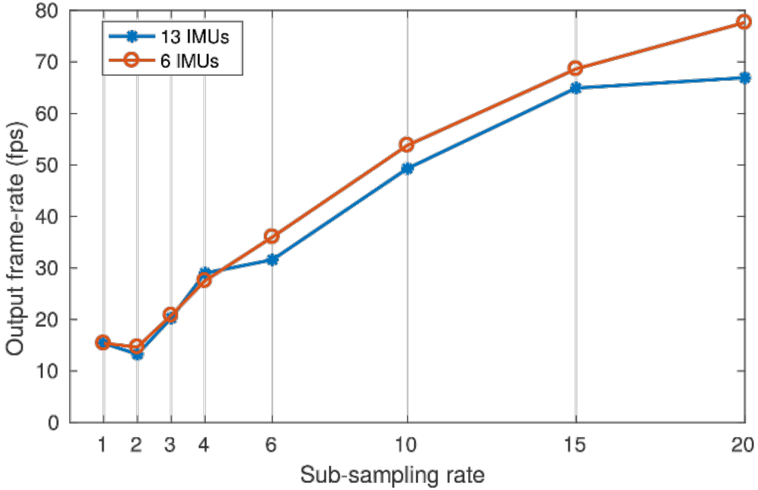
- Can use as **few as 2 cameras**
- Limited benefit in using more than 3-4 cameras
- In principle, a single camera could be used, but having multiple views **avoids depth ambiguity**
- No requirement for foreground segmentation or visual hulls, thus more **freedom in capture environment** and camera layout

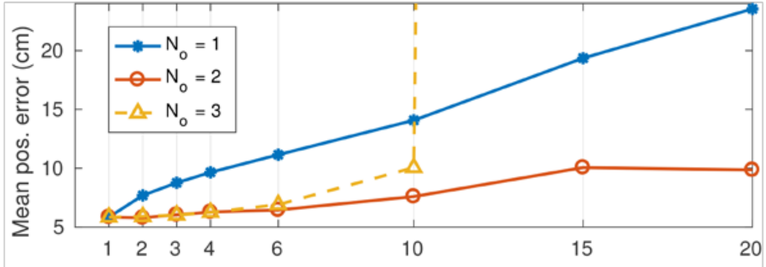


Input configuration

2D detection subsampling

- **Increase** output **frame-rate** by performing **expensive CPM detection** on a **subset of input frames**
- High quality (**HQ**) setting detect on all frames (1/1), 8 cameras
- Hight speed (**HS**) – detect on 2/8 frames, 4 cameras
- Best to detect **2 consecutive frames** rather than 1 frame and shorter interval (bottom right-hand figure)





(N_o is number of consecutive frames detected)

Input configuration

Number of IMUs and quality/speed trade-off

	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

Input configuration

Omitting terms from the cost function

- Orientation term important for removing jitter in position as well as disambiguating axial orientation
- Acceleration term has relatively small impact
- Position term important to lock down global 3D position (avoids run-away drift from double integration of noisy acceleration)
- PCA projection and deviation prior terms important for constraining pose

	13 IMUs		6 IMUs	
Terms Omitted	Pos.	Ori.	Pos.	Ori.
IMU (E_R, E_A)	1.97	4.82	1.27	2.38
Ori. (E_R)	2.63	6.27	1.54	2.89
Acc. (E_A)	1.11	0.99	1.01	0.97
Pos. (E_P)	188.58	1.00	194.82	1.05
Prior (E_{PP}, E_{PD})	1.50	4.68	1.42	4.33
Prior Proj. (E_{PP})	2.26	6.29	1.63	6.46
Prior Dev. (E_{PD})	1.16	2.86	1.46	3.24

Position and angle error with terms omitted (relative to full cost function below)

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

Conclusion

Conclusions and future work

- Hybrid motion capture approach
    - Full 6DOF kinematic solve
    - Drift-free 3D global translation
    - Unconstrained capture environment
    - Flexible, sparse input configurations
    - Real-time, online (suitable for on-set pre-vis, interactive applications)
- Future work
    - Improve real-time performance by using multiple GPUs for CPM detection
    - Extending to work with multiple people

References

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

[5] S. Agarwal, K. Mierle, and Others. Ceres solver. http://ceres-solver.org

[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, Evren Imre, Joao Regateiro and Armin Mustafa for their help with data capture.

Questions?