Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors

UNIVERSITY OF SURREY
Centre for Vision Speech and Signal Processing (CVSSP)

BMVC 2017 LONDON

M. Trumble, A. Gilbert, C. Malleson, A. Hilton and J. Collomosse
{matthew.trumble | a.gilbert | charles.malleson | a.hilton | j.collomosse}@surrey.ac.uk

Motivation

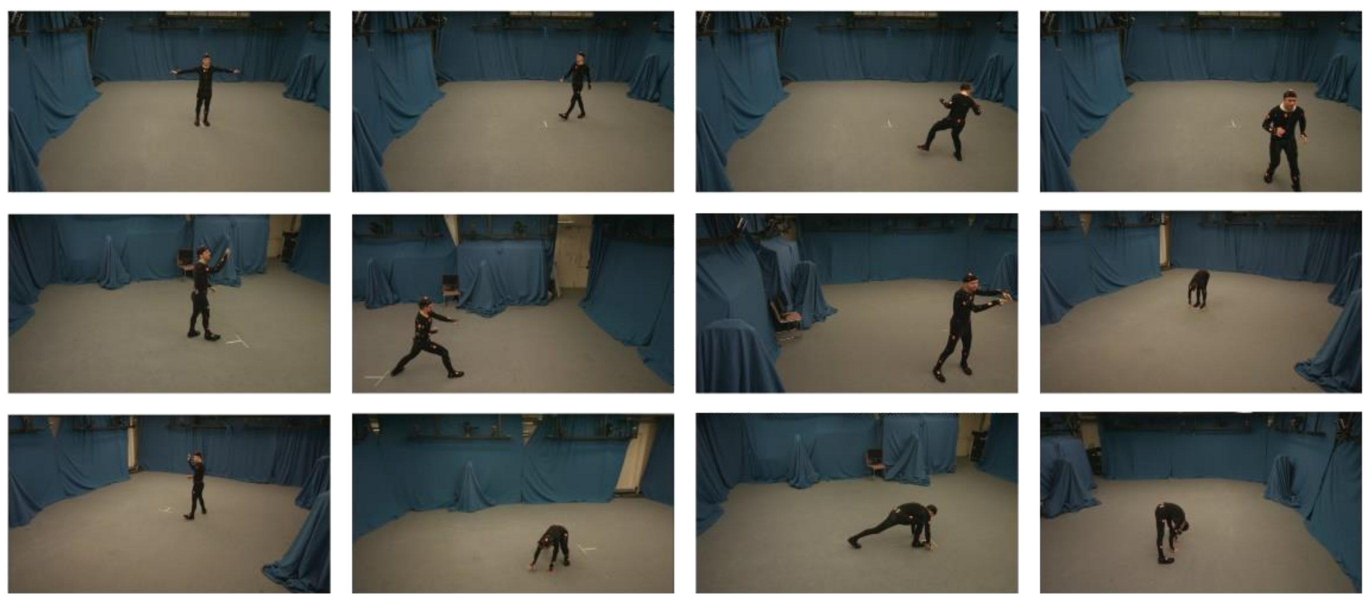
- Markerless motion capture – no special suit required for performer
- Unconstrained environments – remove need for dedicated motion capture shoots
- Fuse video and inertial sensors – overcome limitations of individual sources

Contributions

- Novel 3D human pose estimation fusing multi-view video and inertial signals
- Multiple views incorporated into fully 3D convolutional neural network
- Releasing new hybrid dataset including video, IMU and 3D ground truth

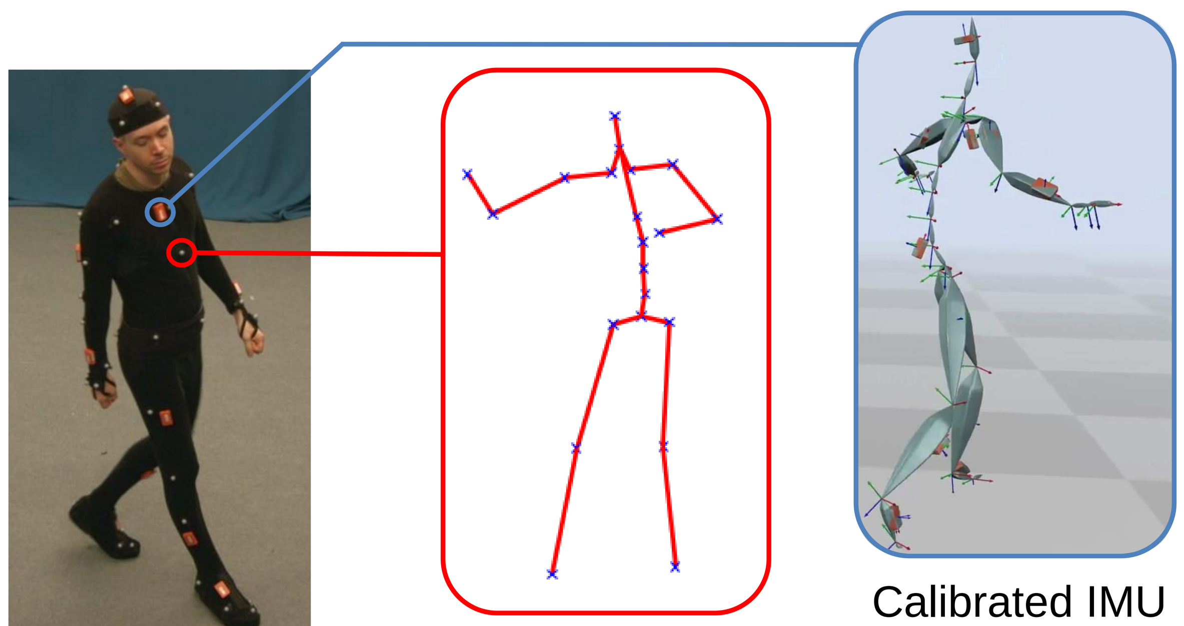
Total Capture Dataset

Multi-view video, inertial measurement unit (IMU) and vicon ground truth 3D pose data



- 8 x 1080p60 video cameras
- 13 IMU sensors
- Vicon ground truth labelling
- 5 subjects x 12 sequences
- Over 1hour of 60Hz footage
- Freely available at:

http://cvssp.org/data/totalcapture



3D ground truth pose

Calibrated IMU rotation and acceleration

Network Training

Video branch

- Geometric proxy (PVH) constructed from MVV
- Passed as input into 3D CNN
- 100k unique training poses / 50K test
- Augmented with rotation around vertical axis

IMU branch

- Manually calibrated to initial T-pose
- Joint angles inferred by forward kinematics

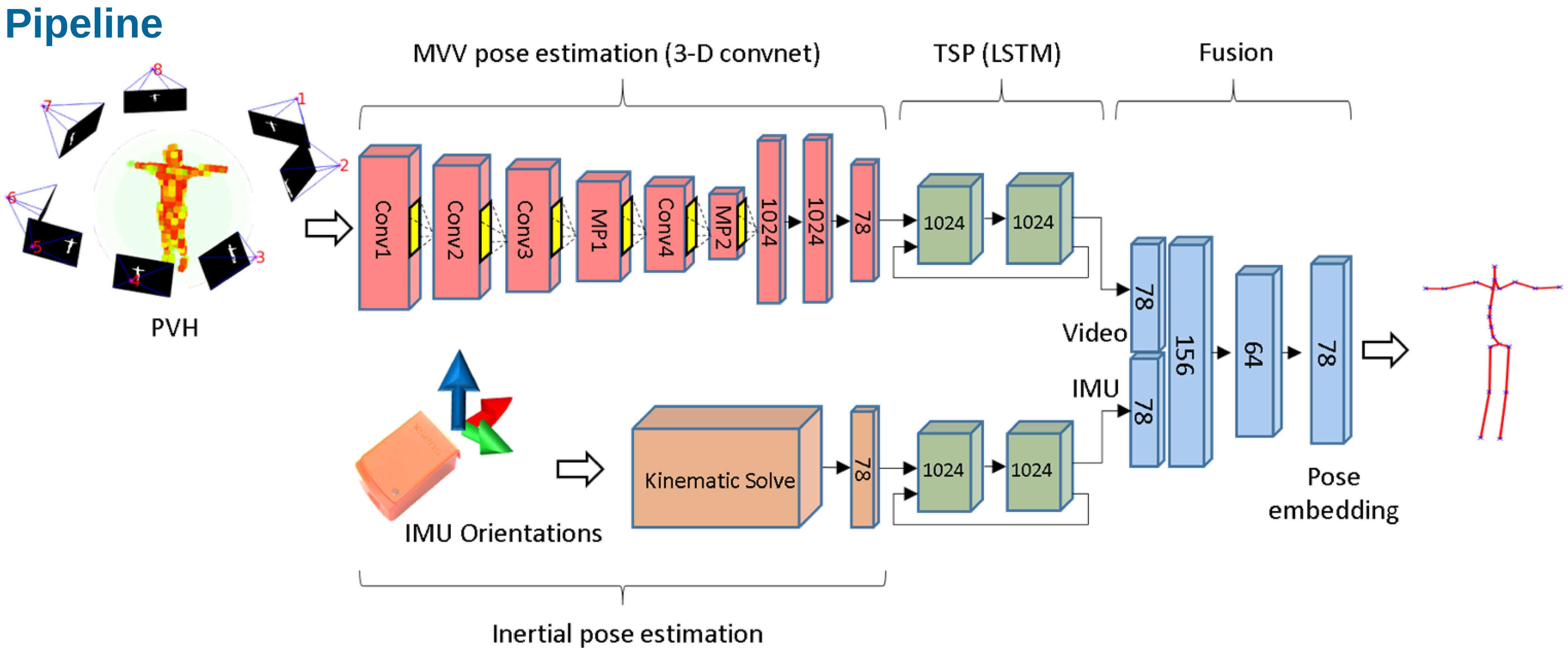
Temporal sequence prediction model (TSP)

- Independent model trained for each modality
- Enforces temporal consistency with memory cells

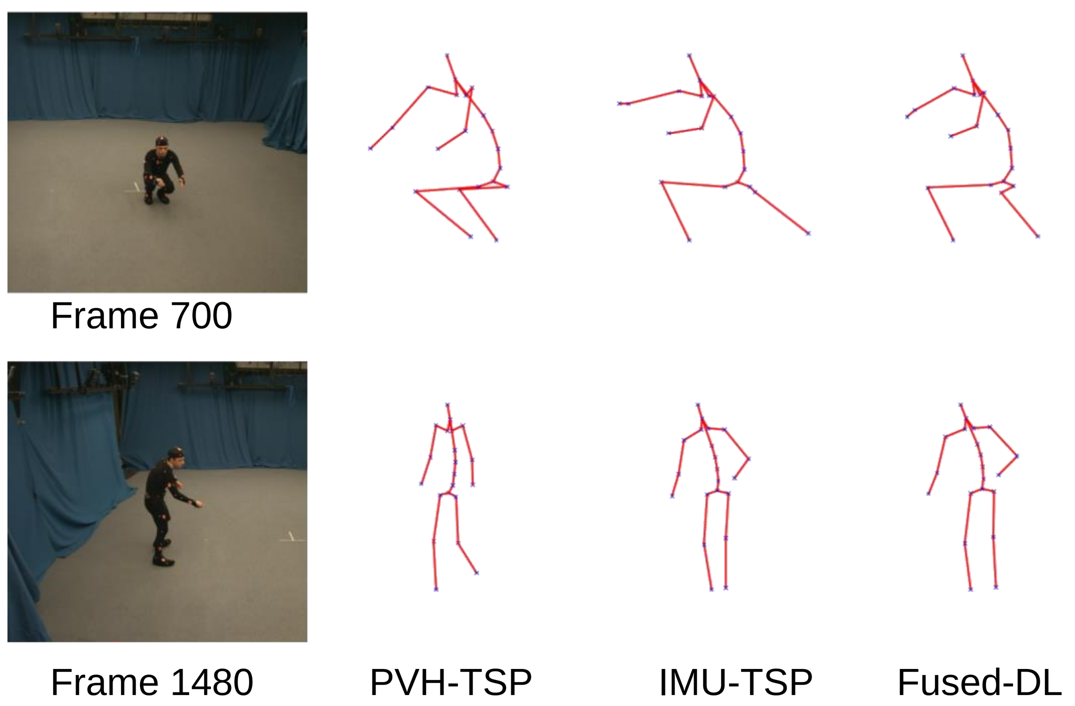
Fusion layer

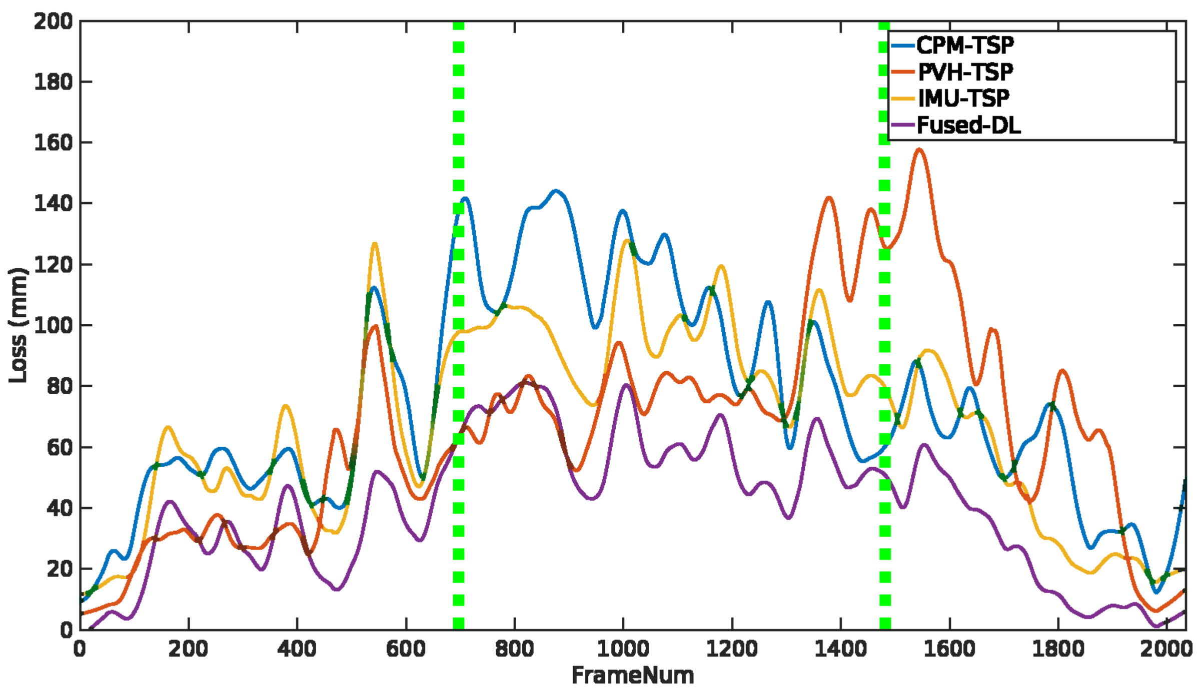
- Output from branches concatenated
- Passed through 2 fully connected neural layers

Pipeline

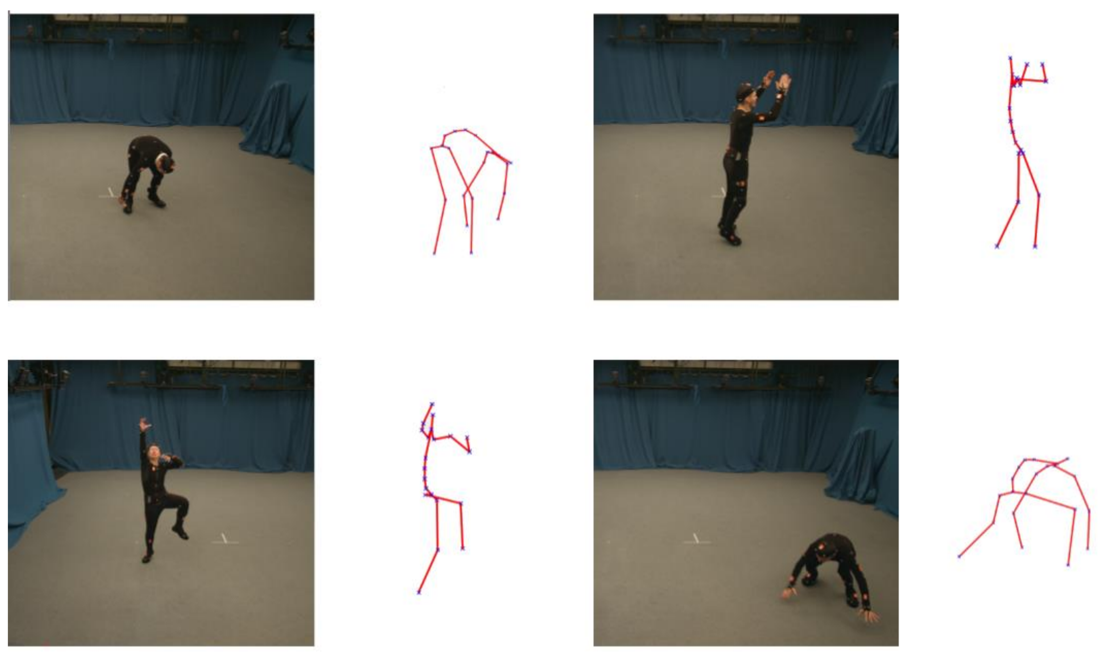


Results – Total Capture





Per frame accuracy on sequence FS3 Subject 4



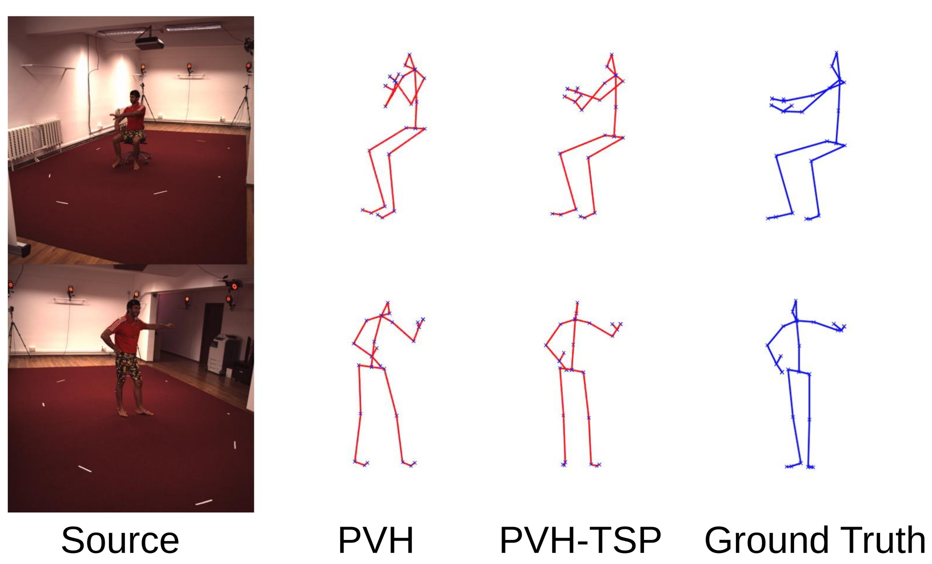
Additional results across diverse poses within Total Capture

Approach	SeenSubjects(S1,2,3) W2	FS3	A3	UnseenSubjects(S4,5) W2	FS3	A3	Mean
Tri-CPM [1]	79.0	112.1	106.5	79.0	149.3	73.7	99.8
Tri-CPM-TSP [1]	45.7	102.8	71.9	57.8	142.9	**59.6**	80.1
2D Matte [2]	104.9	155.0	117.8	161.3	208.2	161.3	142.9
2D Matte-TSP [2]	94.1	128.9	105.3	109.1	168.5	120.6	121.1
3D PVH	48.3	122.3	94.3	84.3	168.5	154.5	107.3
3D PVH-TSP	38.8	**86.3**	72.6	69.1	112.9	119.5	81.1
Solved IMU	62.4	129.5	78.7	68.0	162.5	146.0	107.9
Solved IMU-TSP	39.4	118.7	52.8	58.8	141.1	135.1	91.0
Fused-Mean IMU+3D PVH	37.3	113.8	61.3	45.2	156.7	136.5	91.8
Fused-DL IMU+3D PVH	**30.0**	90.6	**49.0**	**36.0**	**112.1**	109.2	**70.0**

Average per joint error in mm

[1] Wei et al. Convolutional Pose Machines, CVPR 2016
[2] Trumble et al. Deep convolutional networks for maker-less human pose estimation from multiple views, CVMP 2016

Results – Human 3.6M



- 4 x MVV camera input only
- No IMU sensors
- Evaluation on vision branch only
- Tri-CPM: triangulation of per camera 2D joint estimates using Convolutional Pose Machines (Wei et al. CVPR 2016)

Approach	Direct.	Discus	Eat	Greet.	Phone	Photo	Pose	Purch.
Tri-CPM[1]	125.0	111.4	101.9	142.2	125.4	147.6	109.1	133.1
Tri-CPM-TSP[1]	67.4	71.9	65.1	108.8	88.9	112.0	55.6	77.5
PVH-TSP	92.7	85.9	72.3	93.2	86.2	101.2	75.1	78.0
	Sit.	Sit D	Smke	Wait	W.Dog	walk	W. toget.	Mean
Tri-CPM[1]	135.7	142.1	116.8	128.9	111.2	105.2	124.2	124.0
Tri-CPM-TSP[1]	92.7	110.2	80.3	100.6	71.7	57.2	77.6	88.1
PVH-TSP	83.5	94.8	85.8	82.0	114.6	94.9	79.7	87.3

Average per joint error in mm

Acknowledgements

The work was supported by an EPSRC doctoral bursary and InnovateUK via the Total Capture project, grant agreement #102685. The work was supported in part by the Visual Media project (EU H2020 grant #687800) and through donation of GPU hardware by Nvidia corporation.