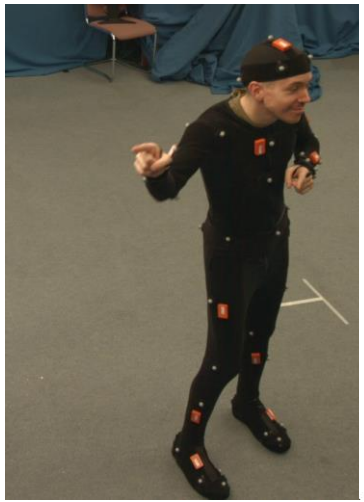
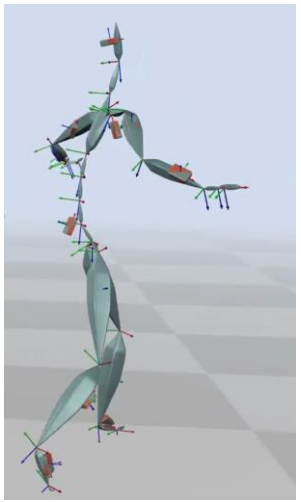
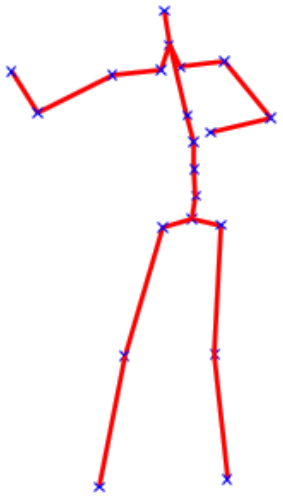
Total Capture

3D Human Pose Estimation Fusing Video and Inertial Sensors

M. Trumble, A. Gilbert, C. Malleson, A. Hilton and J. Collomosse
Centre for Vision, Speech and Signal Processing

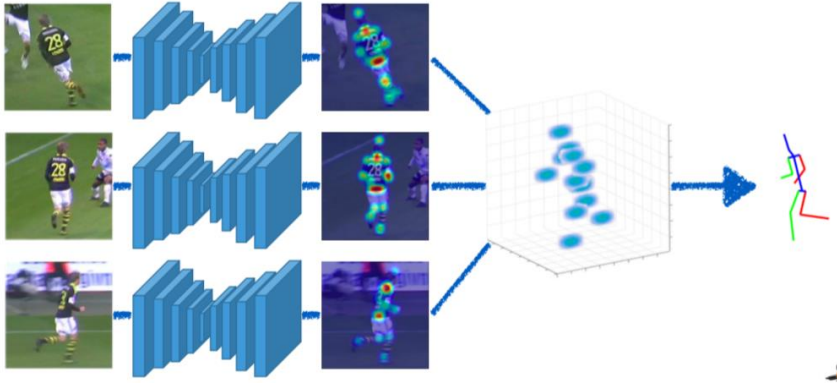


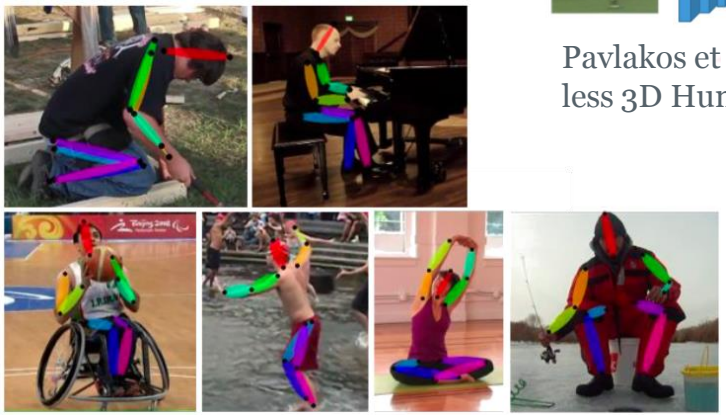
Motivation

Image credit: Electronic Arts

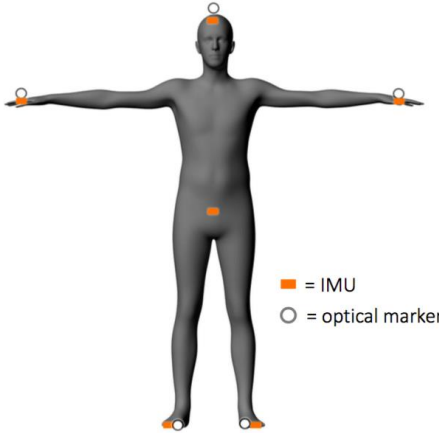
Related Work



Pavlakos et al. Harvesting Multiple Views for Marker-less 3D Human Pose Annotations, CVPR 2017

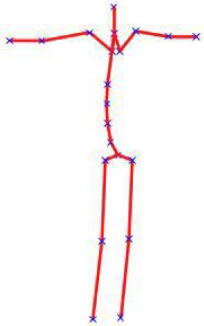


Wei et al. Convolutional Pose Machines, CVPR 2016



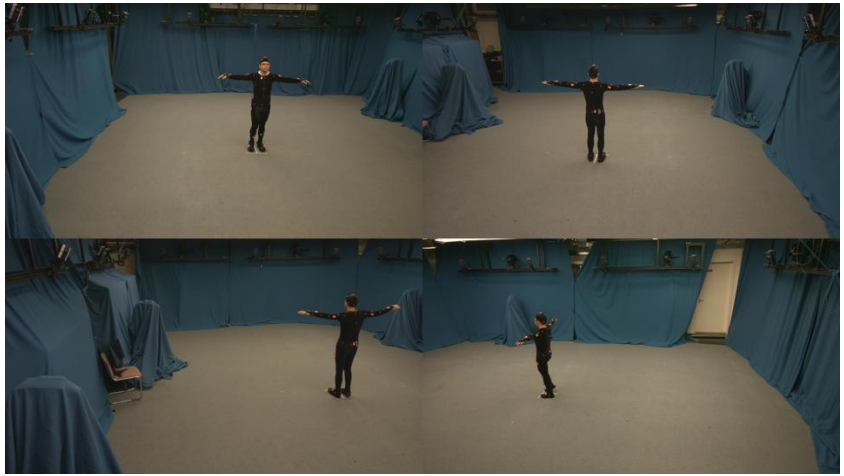
Andrews et al. Real-time Physics-based Motion Capture with Sparse Sensors, CVMP 2016

Contributions



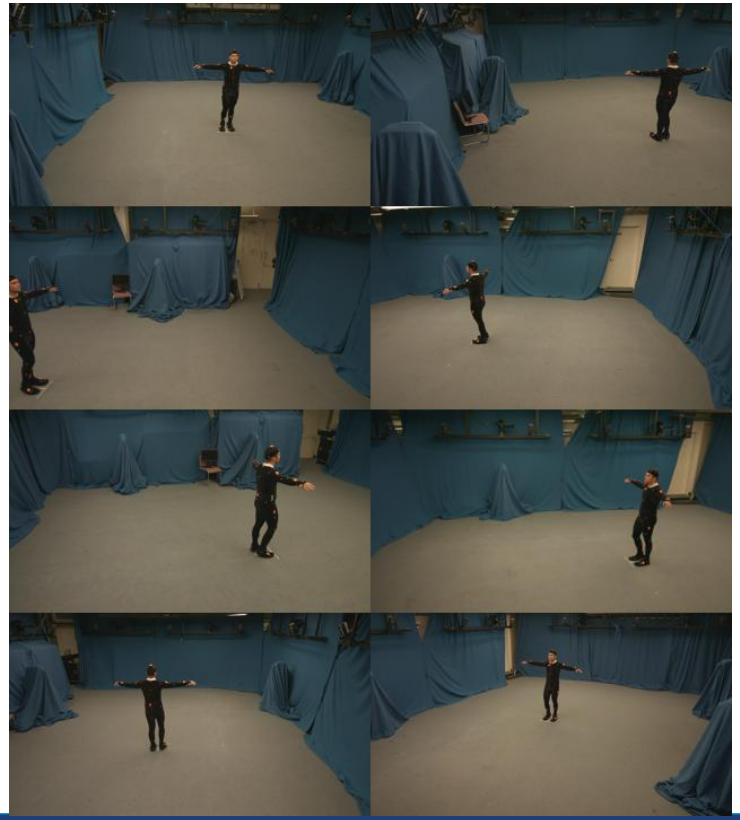
- Fusion of video and IMUs
- New multi-modal dataset

- Accurate 3D human pose estimation
- 3D Convolutional Neural Network



Contributions

Total Capture Dataset



- 4 x 6 metre capture volume
- 8 x 1080p60 video cameras
- 13 IMU sensors
- Vicon ground truth labelling
- 5 subjects x 12 sequences

http://cvssp.org/data/totalcapture

Contributions

Total Capture Dataset

Xsens MTw Awinda wireless motion trackers

- Calibrated orientation and acceleration per unit at 60Hz

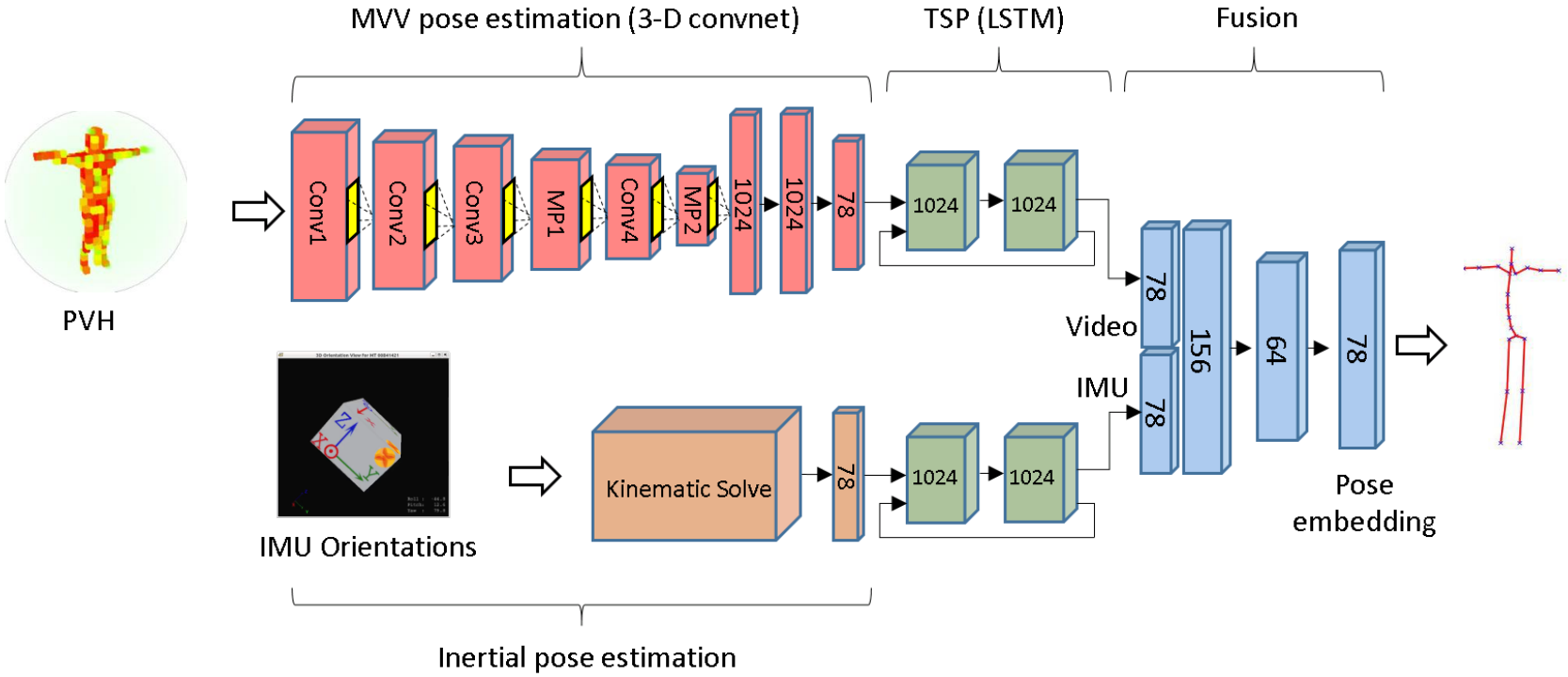
Vicon motion capture for testing

- Solved skeleton provided in BVH format, also 60Hz

http://cvssp.org/data/totalcapture

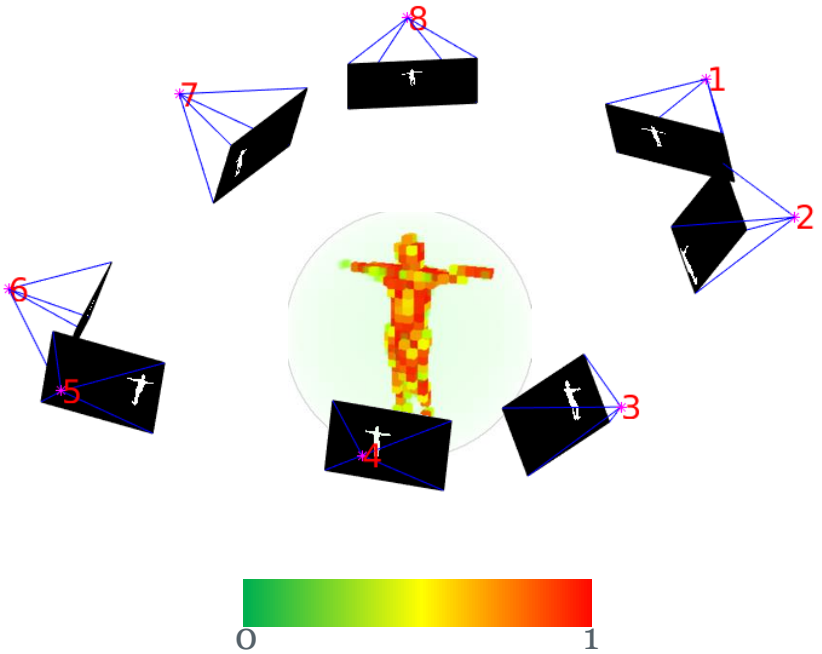
Pipeline

Overview



Pipeline

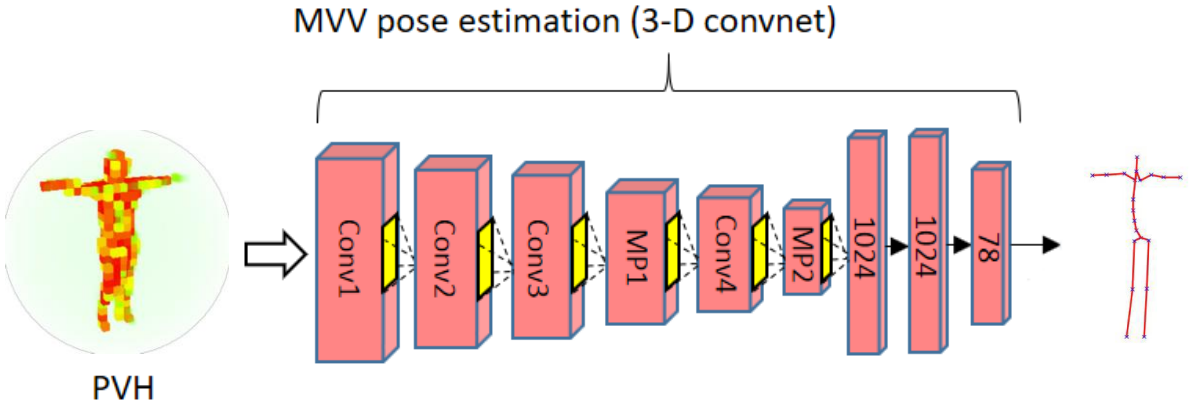
Volumetric Pose Estimation – Probabilistic Visual Hull (PVH)



- Geometric proxy constructed from MVV
- Capture volume decimated into $1cm^3$ grid
- Voxels assigned probability of occupancy
- Downsampled to 30x30x30 grid for CNN

Pipeline

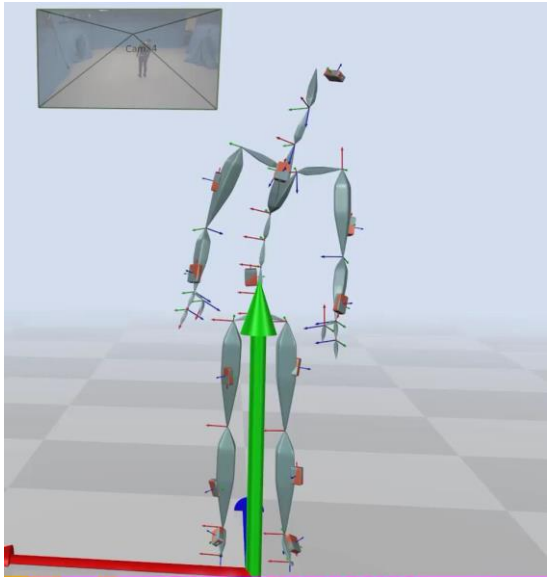
Volumetric Pose Estimation – 3D CNN Training



- Trained with stochastic gradient descent to minimize mean squared error over 26 3D joint positions
- 100K unique training poses / 50K test from Total Capture dataset
- Augmented during training with random rotation around vertical axis

Pipeline

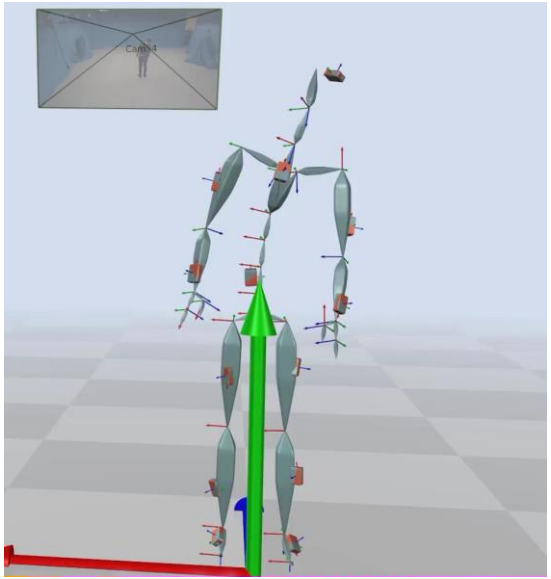
Inertial Pose Estimation



- 13 inertial measurement units (IMUs)
- Arms and legs, feet, head, sternum and pelvis
- Manual calibration to an initial T-pose
- Joint angles inferred by forward kinematics

Pipeline

Inertial Pose Estimation – forward kinematics



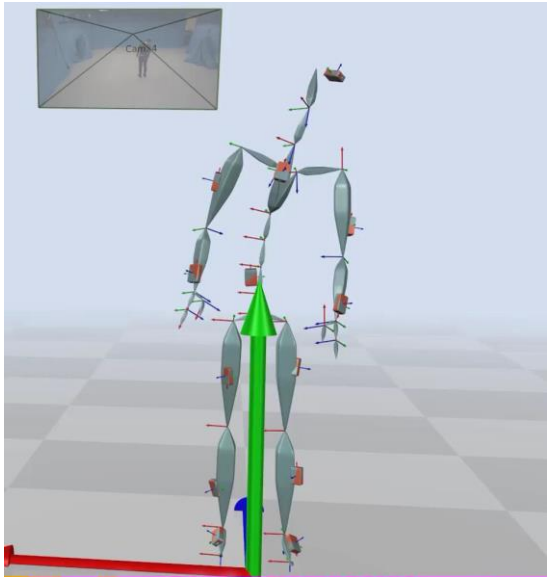
Assume fixed relative orientation between each IMU ($k \in [1,13]$) and bone: R_{ib}^{k}

Global bone orientation $\boldsymbol{R_b^k = (R_{ib}^k)^{-1}\, R_{iw}^k R_{im}^k}$

where R_{iw}^{k} is IMU reference frame in global coordinates and local IMU measurement R_{im}^{k}

Pipeline

Inertial Pose Estimation – forward kinematics

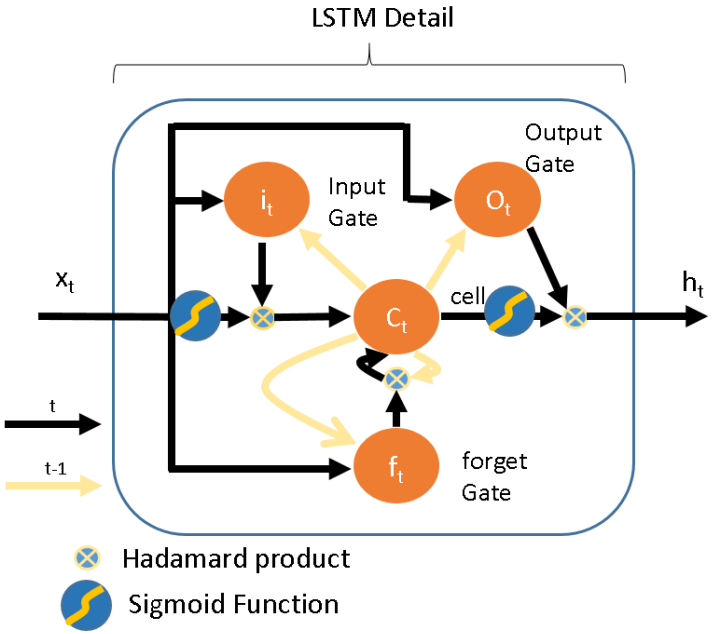


Local joint rotation $\boldsymbol{R_h^i = R_b^i (R_b^{par(i)})^{-1}}$

Inferred from parent bone, $par(i)$
by forward kinematics beginning at root node

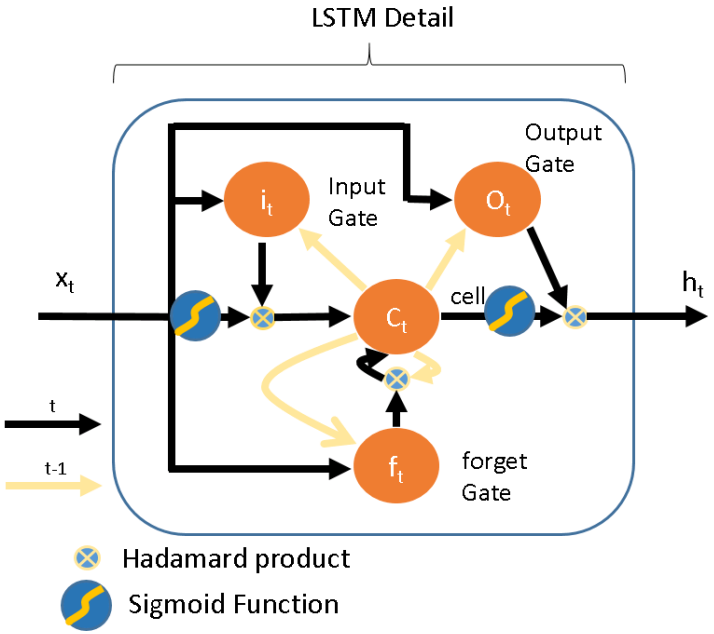
Pipeline

Temporal Sequence Prediction (TSP)



- Long Short Term Memory RNN (LSTM)
- Exploits temporal nature of motion
- Independent model for each modality
- Learns joint locations based on previous 5 frames

Temporal Sequence Prediction (TSP) – LSTM details



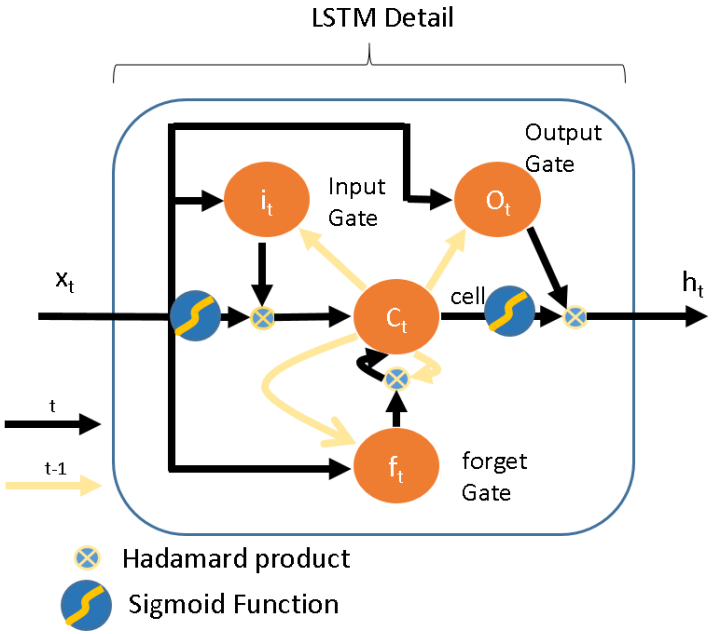
Input vector x_t, output vector $h_t = o_t \circ \sigma_h(c_t)$, learnt weights W and U

Memory cell,

$$\boldsymbol{c_t = f_t \circ c_{t-1} + i_t \circ \sigma_h(W_x x_t + U_c h_{t-1} + b_c)}$$

sigmoid function σ_g,hyperbolic tangent σ_h, vector constant b

Temporal Sequence Prediction (TSP) – LSTM details



Input vector x_t, output vector $h_t = o_t \circ \sigma_h(c_t)$, learnt weights W and U

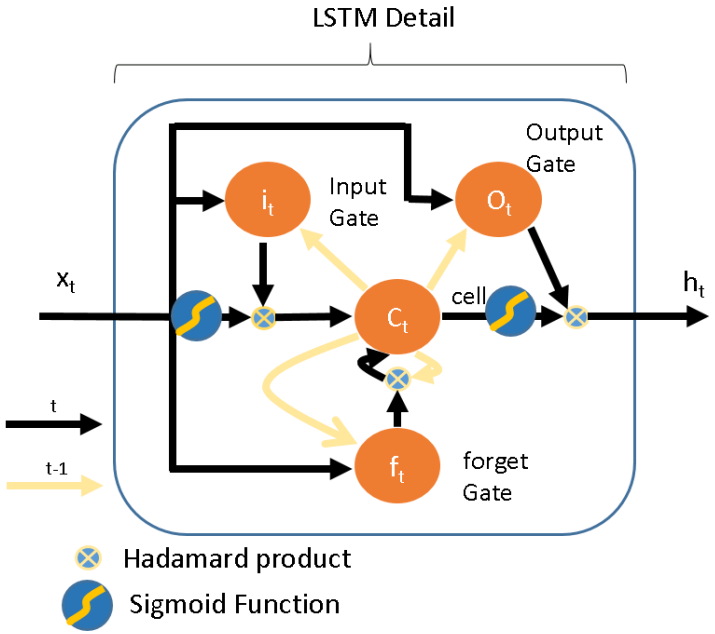
Memory cell,

$$\boldsymbol{c_t = f_t \circ c_{t-1} + i_t \circ \sigma_h(W_x x_t + U_c h_{t-1} + b_c)}$$

Input gate $\boldsymbol{i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)}$

sigmoid function σ_g,hyperbolic tangent σ_h, vector constant b

Temporal Sequence Prediction (TSP) – LSTM details



Input vector x_t, output vector $h_t = o_t \circ \sigma_h(c_t)$, learnt weights W and U

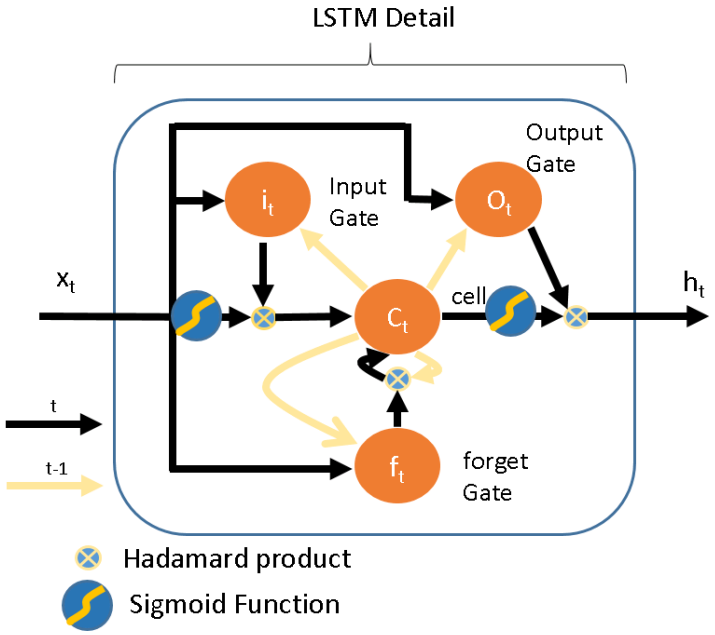
Memory cell,

$$\boldsymbol{c_t = f_t \circ c_{t-1} + i_t \circ \sigma_h(W_x x_t + U_c h_{t-1} + b_c)}$$

Forget gate $\boldsymbol{f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)}$

sigmoid function σ_g,hyperbolic tangent σ_h, vector constant b

Temporal Sequence Prediction (TSP) – LSTM details



Input vector x_t, output vector $h_t = o_t \circ \sigma_h(c_t)$, learnt weights W and U

Memory cell,

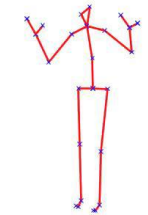
$$\boldsymbol{c_t = f_t \circ c_{t-1} + i_t \circ \sigma_h(W_x x_t + U_c h_{t-1} + b_c)}$$

Output gate $\boldsymbol{o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)}$

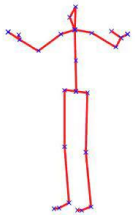
sigmoid function σ_g,hyperbolic tangent σ_h, vector constant b

Evaluation – video branch

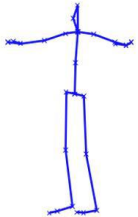
Human 3.6M



PVH Only



PVH + TSP



Ground Truth

Source

Evaluation – video branch

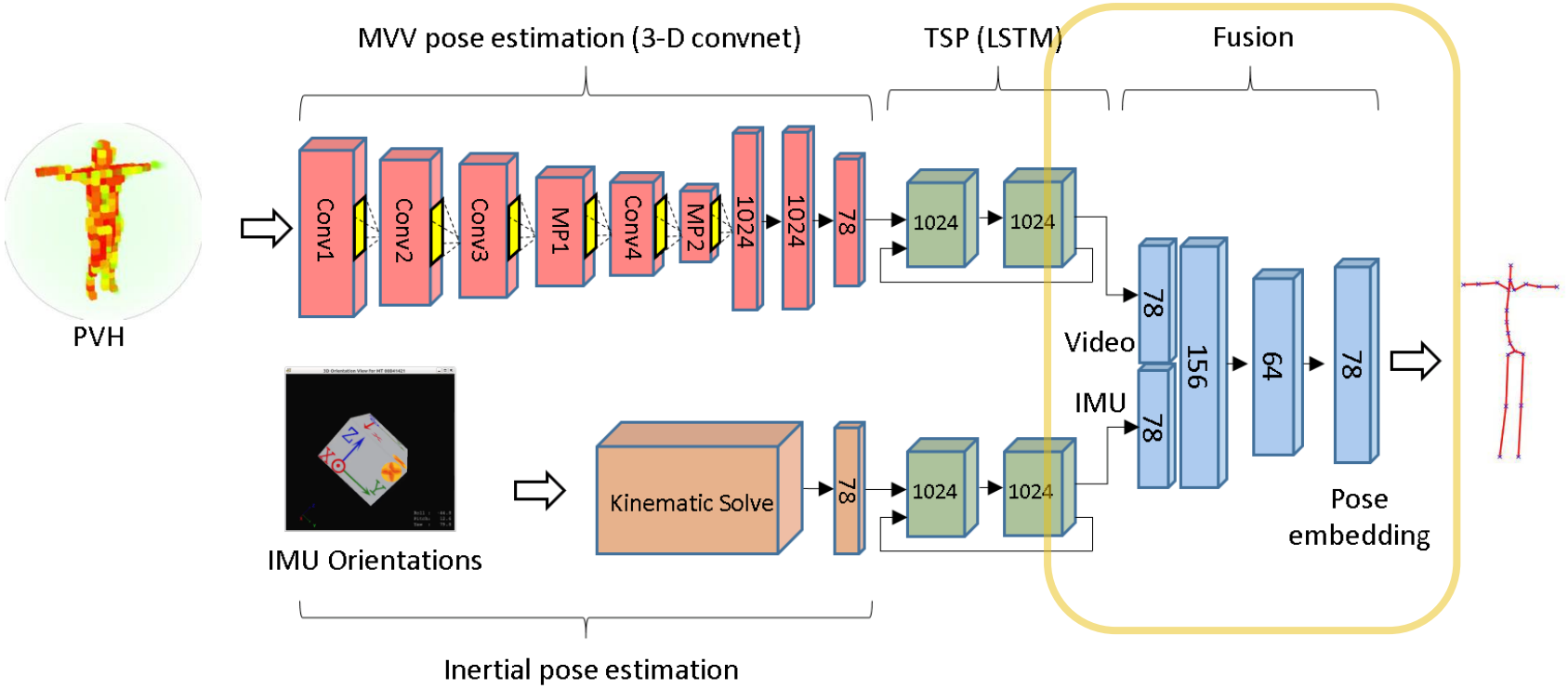
Human 3.6M

Approach	Direct.	Discus	Eat	Greet.	Phone	Photo	Pose	Purch.
Tri-CPM	125.0	111.4	101.9	142.2	125.4	147.6	109.1	133.1
Tri-CPM-TSP	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	135.7	142.1	116.8	128.9	111.2	105.2	124.2	124.0
Tri-CPM-TSP	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 millimetres

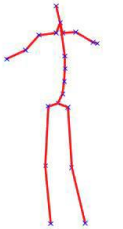
Pipeline

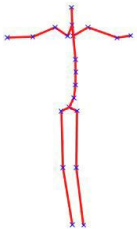
Fusion layer

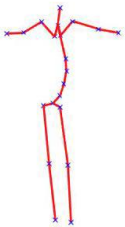


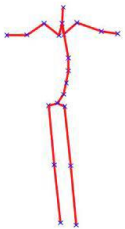
Evaluation – full pipeline

Total Capture Dataset – Full Pipeline









PVH + TSP

IMU + TSP

Fusion

Source

Evaluation – full pipeline

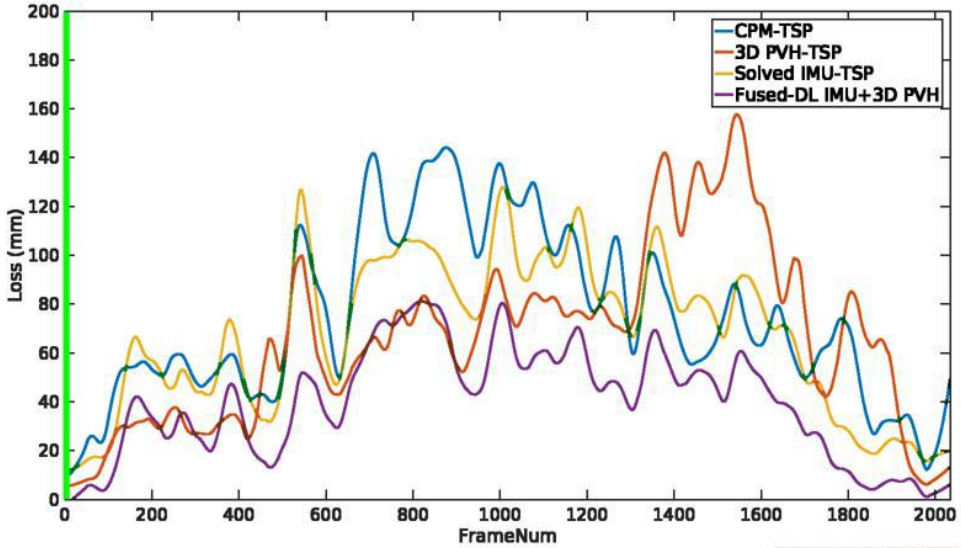
Total Capture Dataset

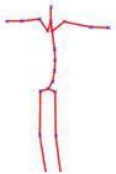
Approach	SeenSubjects(S1,2,3)			UnseenSubjects(S4,5)			Mean
	W2	FS3	A3	W2	FS3	A3	
Tri-CPM	79.0	112.1	106.5	79.0	149.3	73.7	99.8
Tri-CPM-TSP	45.7	102.8	71.9	57.8	142.9	**59.6**	80.1
3D PVH	48.3	122.3	94.3	84.3	168.5	154.5	107.3
3D PVH-TSP	38.8	**86.3**	72.6	69.1	112.9	119.5	81.1
Solved IMU	62.4	129.5	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 millimetres

Evaluation – full pipeline

Total Capture Dataset – Full Pipeline





PVH + TSP

IMU + TSP

Fusion

Source

Evaluation

Training data volume

Training Data Volume	Relative Accuracy
20%	87.1%
40%	90.4%
60%	96.7%
80%	99.4%

Training data randomly sampled from ~100k MVV frames

PVH resolution

PVH Dimensions	Per joint error (mm)
16x16x16	111
30x30x30	107
48x48x48	110

16x16x16

48x48x48

Evaluation

Camera ablation study

Num Cams	SeenSubjects(S1,2,3)			UnseenSubjects(S4,5)			Mean
	W2	FS3	A3	W2	FS3	A3	
4	93.8%	90.8%	95.3%	91.6%	89.5%	93.5%	90.4%
6	94.3%	99.3%	97.4%	96.0%	98.2%	98.1%	96.2%
8	100%	100%	100%	100%	100%	100%	100%

Relative accuracy change (mm/joint)

4 cameras

6 cameras

8 cameras

Conclusion

- Novel 3D human pose estimation fusing MVV and IMU signals
- Demonstrates high accuracy and complementary nature of the two modalities
- New hybrid MVV dataset including video, IMU and 3D ground truth

http://cvssp.org/data/totalcapture