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

Matthew Trumble
matthew.trumble@surrey.ac.uk
Andrew Gilbert
a.gilbert@surrey.ac.uk
Charles Malleson
charles.malleson@surrey.ac.uk
Adrian Hilton
a.hilton@surrey.ac.uk
John Collomosse
j.collomosse@surrey.ac.uk

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, UK

Abstract

We present an algorithm for fusing multi-viewpoint video (MVV) with inertial measurement unit (IMU) sensor data to accurately estimate 3D human pose. A 3-D convolutional neural network is used to learn a pose embedding from volumetric probabilistic visual hull data (PVH) derived from the MVV frames. We incorporate this model within a dual stream network integrating pose embeddings derived from MVV and a forward kinematic solve of the IMU data. A temporal model (LSTM) is incorporated within both streams prior to their fusion. Hybrid pose inference using these two complementary data sources is shown to resolve ambiguities within each sensor modality, yielding improved accuracy over prior methods. A further contribution of this work is a new hybrid MVV dataset (TotalCapture) comprising video, IMU and a skeletal joint ground truth derived from a commercial motion capture system. The dataset is available online at http://cvssp.org/data/totalcapture/.

1 Introduction

The ability to record and understand 3-D human pose is vital to a huge range of fields, from biomechanics, psychology, animation, and computer vision. Human pose estimation aims to deduce a skeleton from data in terms of 3-D limb location/orientation or a probability map of their locations. Currently to achieve a highly accurate understanding of the human pose, commercial marker-based systems such as Vicon [3] or OptiTrack [1] are used.

However, marker-based systems are intrusive and restrict the motions and appearance of the subjects, and often fail with heavy occlusion or in high illumination. A special suit augmented with small reflective markers and, many specialist cameras (IR) are necessary, increasing cost and setup time and restricts the shooting to artificially lit areas. To remove these constraints there has been significant progress in the vision-based estimation of 3D human pose. however, a complex human body model is used to constrain the estimates [34] or depth data [37] is required. Inertial Measurement Units (IMUs) [2, 25] have been introduced

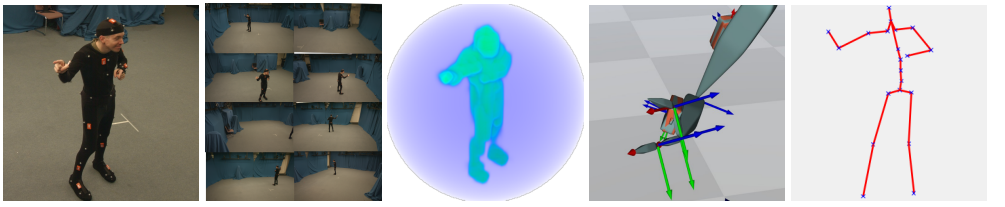


Figure 1: Our two-stream network fuses IMU data with volumetric (PVH) data derived from multiple viewpoint video (MVV) to learn an embedding for 3-D joint locations (human pose).

as a compromise, placed on key body parts and used for motion capture, without the concerns of occlusions and illumination. However, they suffer from drift over even short time periods.

Therefore we propose the fusion of vision and IMUs to estimate the 3-D joint skeleton of human subjects overcoming the limitations of the drift and lack of positional information in IMU data and the requirement of learnt complex human models. We show that the complementary modalities mutually reinforce one another during inference; rotational and occlusion ambiguities are mitigated by the IMUs whilst global positional drift is mitigated by the video. Our proposed solution combines alpha foreground mattes from a number of synchronised wide baseline video cameras to form a probabilistic visual hull (PVH), which is used to train a 3-D convolutional network to predict joint estimates. These joint estimates are fused with joint estimates from IMU data within a simple kinematic model, as illustrated in Fig 1. Taking advantage of the temporal nature of the sequences, Temporal Sequence Prediction (TSP) is employed on the video and IMU pose estimates to provide contextual frame-wise predictions using a variant of Recurrent Neural Networks (RNN) using LSTM layers. The two independent data modes are fused within a two-stream network so combining the complementary signals from the multiple viewpoint video (MVV) and IMUs. Currently, there is no dataset available containing IMU and MVV video with a high-quality ground truth. We release such a multi-subject, multi-action dataset as a further contribution of this work.

2 Related Work

Approaches can be split into two broad categories; a top-down approach to fit an articulated limb kinematic model to the source data and those that use a data driven bottom-up approach.

Lan [18] provide a top down model based approach, considering the conditional independence of parts; however Inter-Limb dependencies (*e.g.* symmetry) are not considered. A more global treatment is proposed in [17] using linear relaxation but performs well only on uncluttered scenes. The *SMPL* body model [21] provides a rich statistical body model that can be fitted to incomplete data and Marcard [35] incorporated IMU measurements with the SMPL model to provide pose estimation without visual data.

In bottom-up pose estimation, Ren [24] recursively splits Canny edge contours into segments, classifying each as a putative body part using cues such as parallelism. Ren [23] also used BoVW for implicit pose estimation as part of a pose similarity system for dance video retrieval. Toshev [31], in the DeepPose system, used a cascade of convolutional neural networks to estimate 2-D pose in images. Sanzari [26] estimates the location of 2D joints, before predicting 3D pose using appearance and probable 3-D pose of the discovered parts with a hierarchical Bayesian model. While Zhou [38] integrates 2-D, 3-D and temporal information to account for uncertainties in the data. The challenge of estimating 3D human pose from MVV is currently less explored, although initial work by Trumble [32] used MVV with a

simple 2D convolutional neural network (convnet), and Wei [36] performed related work aligning pairs of 3D human pose. While Huang [15] used a tracked 4-D mesh of a human performer from video reconstruction for estimating pose.

To predict temporal sequences, RNNs and their variants including LSTMs [13] and Gated Recurrent Units [7] have recently shown to successfully learn and generalise the properties of temporal sequences. Graves [10] was able to predict isolated handwriting sequences, and transcribe audio data with text [11]. While Alahi [4] was able to predict human trajectories of crowds by modelling each human with an LSTM and jointly predicting the paths.

In the field of IMUs, Roetenberg [25], used 17 IMUs with 3-D accelerometers, gyroscopes and magnetometers to define the pose of a subject. Marcard [33] fused video and IMU data to improve and stabilise full body motion capture. While Helten [12] used a single depth camera with IMUs to track the full body.

3 Methodology

A geometric proxy of the performer is constructed from MVV on a per frame basis and passed as input into a convnet designed to accept a 3-D volumetric representation, the network directly regresses an embedding that encodes 3-D skeletal joint positions. That estimate is then processed through a temporal model (LSTM) and fused with a similarly processed signal from a forward kinematic solve of the IMU data to learn a final pose embedding (Fig. 2).

3.1 Volumetric Representation of Proxy

Images from the MVV camera views are integrated to create a probabilistic visual hull (PVH) adapting the method of Grauman [9]. Each of the C cameras, $c = [1, C]$, where $C > 3$, is calibrated with known orientation R_c, focal point COP_c, focal length f_c and optical centre o_c^x, o_c^y, the image from which is denoted I_c. A 3D performance volume centred on the performer, is decimated into voxels $\mathcal{V} = \{V_1, \ldots, V_m\}$ approximately 1cm^3 in size. Voxel occupancy from a given view c is defined as the probability:

$$p(V|c) = B(I_c(x[V_i], y[V_i])) \tag{1}$$

Where $B(.)$ is background subtraction of I_c from a clean plate at image position (x, y) and where the voxel V_i projects to:

$$x[V_i] = \frac{f_c v_x}{v_z} + o_c^x \quad \text{and} \quad y[V_i] = \frac{f_c v_y}{v_z} + o_c^y, \tag{2}$$

$$\text{where} \quad \begin{bmatrix} v_x & v_y & v_z \end{bmatrix} = COP_c - R_c^{-1} V_i. \tag{3}$$

The overall probability of occupancy for a given voxel $p(V)$ is the product over all views:

$$p(V_i) = \prod_{i=1}^{C} p(V|c), \tag{4}$$

calculated for all $V_i \in \mathcal{V}$ to create the initial PVH. This is down sampled via a Gaussian filter to a volume of dimensions $30 \times 30 \times 30$, the input size for our CNN.

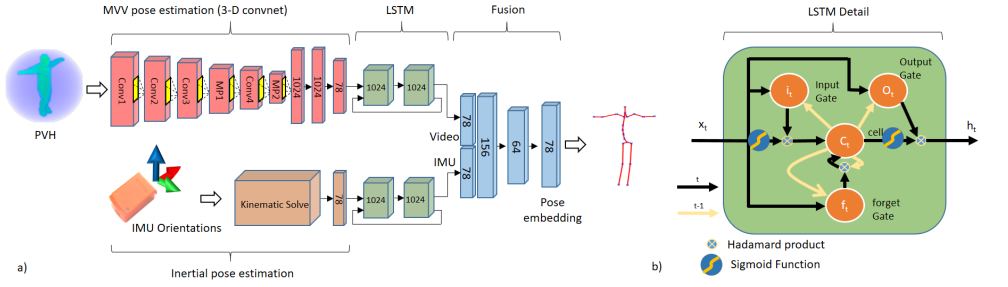


Figure 2: Network architecture (a) comprising two streams: a 3D Convnet for MVV/PVH pose embedding, and kinematic solve from IMUs. Both streams pass through LSTM (b) before fusion of the concatenated estimates in a further FC layer.

3.2 Network Architecture

3.2.1 Volumetric Pose Estimation

The MVV processes volumetric input through a series of 3-D convolution and max-pooling layers to a series of fully connected (fc) layers terminating in 78-D output layer (3×26 encoding Cartesian coordinates of 26 joints). Table 1 lists the filter parameters for each layer (Fig. 2a, red stream). Both max-pooling layers are followed by a 50% dropout layer and ReLu activation is used throughout. A training set comprising exemplar PVH volumes $V = \{v_1, v_2, ..., v_n\}$ downsampled to $30 \times 30 \times 30$ and corresponding ground truth poses $P = \{p_1, p_2, ..., p_n\}$ are used to learn pose embedding $E(V) \mapsto P$ minimising:

$$\mathcal{L}(P,V) = \sum_{i=1}^{n} \|p_i - f(v_i)\|_2^2. \tag{5}$$

During training V is augmented by applying a random rotation about the central vertical axis, $\theta = [0, 2\pi]$ encouraging pose invariance with respect to the direction the performer.

Layer	Conv1	Conv2	Conv3	MP1	Conv4	MP2	FC1	FC2	FC3
Filter dim.	5	3	3	2	3	2	1024	1024	1024
Num. filters	64	96	96	-	96	-	1024	1024	78
Stride	2	1	1	2	1	2	1	1	1

Table 1: Parameters of the 3-D Convnet used to infer the MVV pose embedding.

3.2.2 Inertial Pose Estimation

We use orientation measurements from 13 Xsens IMUs [25] to estimate the pose. The IMU sites are the upper and lower limbs, feet, head, sternum and pelvis. For each IMU, $k \in [1, 13]$, we assume rigid attachment to a bone and calibrate the relative orientation, $\mathbf{R}_{ib}^k$, between them. The reference frame of the IMUs, $\mathbf{R}_{iw}$, is also calibrated approximately against the global coordinates. Using this calibration, a local IMU orientation measurement, $\mathbf{R}_m^k$, is transformed to a global bone orientation, $\mathbf{R}_b$ as follows: $\mathbf{R}_b^k = (\mathbf{R}_{ib}^k)^{-1}\mathbf{R}_{iw}^k\mathbf{R}_{im}^k$. The local (hierarchical) joint rotation, $\mathbf{R}_h^i$, for bone i in the skeleton is inferred by forward kinematics: $\mathbf{R}_h^i = \mathbf{R}_b^i(\mathbf{R}_b^{par(i)})^{-1}$, where $par(i)$ is the parent of bone i. The forward kinematics begins at the root and proceeds down the joint tree (with unmeasured bones kept fixed).

3.2.3 LSTM Temporal Prediction

Both the image and inertial sensors estimate on a per frame basis, however it is desirable to exploit the temporal nature of the signal. Following the success of RNNs for sequence prediction, we propose a Temporal Sequence Prediction (TSP) model to learn previous contextual joint estimations to provide the ability to generalise and predict future joint locations. We use Long Short Term Memory (LSTM) layers [13] that are able to store and access information over long periods of time but mitigate the vanishing gradient problem common in RNNs (Fig. 2, right). Given an input vector x_t and resulting output vector h_t, there are two learnt weights W and U, to learn the function that minimises the loss between the input vector and the output vector $h_t = o_t \circ \sigma_h(c_t)$ ($\circ$ denotes the Hadamard product), where c_t is the memory cell

$$\mathrm{c}_t = f_t \circ c_{t-1} + i_t \circ \sigma_h(W_x x_t + U_c h_{t-1} + b_c) \tag{6}$$

which is formed by three gates shown in Fig 2 (b), an input gate i_t controls the extent to which a new input vector x_t is kept in the memory,

$$\mathrm{i}_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i). \tag{7}$$

A forget gate f_t controls the extent to which a value remains in memory,

$$\mathrm{f}_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{8}$$

and an output gate o_t controls the extent to which the value in memory is used to compute the output activation of the block,

$$\mathrm{o}_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{9}$$

Where the activation functions are as follows; σ_g a sigmoid function, σ_h is a hyperbolic tangents, and b is a vector constant. The weights are trained with back-propagation using the same euclidean loss function as in equation 5. There is one independent model for each modality, the vision and IMU, and LSTM learns joint locations based on the previous f frames and predicts their future position. In implementation, we used two layers both with 1024 memory cells, look back $f = 5$ and a learning rate of 10^{-3} with RMS-prop [8].

3.2.4 Modality Fusion

The vision and IMU sensors both independently provide a 3D coordinate per joint estimate. Therefore, it would make sense to incorporate both modes into the final estimate, given their complementary nature. Naively, an average of the two joint estimates could be used, this would be fast and effective assuming both modalities have small errors, however it is likely that often large errors will be present on one of the modes. We therefore propose to fuse the two modes with a further fully connected layer. This learns the mapping between the predicted joint estimates of the two data sources and the actual joint locations, allowing errors in the pose from the vision and IMU to be identified and corrected for the combined fused model. The fully connected fusion layer consists of 64 units and was trained with an RMS-prop optimiser [8] with learning rate of 10^{-4}. All stages of the model are implemented using Tensorflow.

4 Evaluation

We evaluate our approach on two 3D human pose datasets. We evaluate our MVV only method (Sec 3.1) for pose estimation, *i.e.* using visual data alone, on the MVV dataset *Human3.6M* [16]. Second, we evaluate our full proposed network (using MVV and IMU data) on *TotalCapture*; a new dataset containing MVV and IMU data (plus ground truth).

4.1 Human 3.6M

The Human 3.6M dataset [16] consists of 3.6 million MVV and vicon frames, with 5 female and 6 male subjects, captured on 4 cameras.The subjects are performing typical activities such as walking, eating, etc. Given the lack of IMU data, we are only able to evaluate the performance of the vision component (3D convnet) of our proposed approach. That is from the upper (red, and red+green) branch of Fig 2 (a) without fusion of the IMU data. We use the standard evaluation protocol as followed by [16, 19, 28, 29, 30] where subjects S1, S5, S6, S7, S8 are used for training and Subjects S9, S11 provide the test sequences. We also compare the results of our proposed approach **PVH-TSP** to a 3D triangulated version of the recent Convolution Pose Machine [6] with error rejection, **Tri-CPM**. Per camera 2D joint estimates are triangulated into a 3D point, using a rejection method that maximises the number of 2D estimates with the lowest 3D re-projection error x, via a sigmoid based error metric $E_o = \frac{1}{1+exp(a*x-b)}$, where a and b are constants controlling confidence fall off. This is also presented with further training on the Temporal Sequence Predictor (TSP) model from section 3.2.3, denoted **TRI-CPM-TSP**. To evaluate performance we use the 3D Euclidean error metric, the mean Euclidean distance between the regressed 3D and ground truth, averaged over all 17 joints in millimetres (mm). Results of our 3D volumetric approach with the Temporal Sequence Prediction (TSP) compared to previous approaches is shown in Table 2. Our approach achieves excellent results despite excluding the fusion with the kinematic based

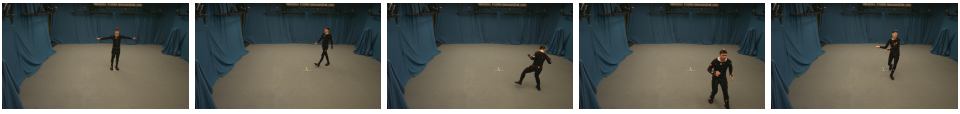
Approach	Direct.	Discus	Eat	Greet.	Phone	Photo	Pose	Purch.
Lin [19]	132.7	183.6	132.4	164.4	162.1	205.9	150.6	171.3
ekin [29]	85.0	108.8	84.4	98.9	119.4	95.7	98.5	93.8
Tome [30]	65.0	73.5	76.8	86.4	86.3	110.7	68.9	74.8
Tri-CPM [6]	125.0	111.4	101.9	142.2	125.4	147.6	109.1	133.1
Tri-CPM-TSP [6]	67.4	71.9	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
Lin [19]	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
ekin [29]	73.8	170.4	85.1	116.9	113.7	62.1	94.8	100.1
Tome [30]	110.2	173.9	85.0	85.8	86.3	71.4	73.1	88.4
Tri-CPM [6]	135.7	142.1	116.8	128.9	111.2	105.2	124.2	124.0
Tri-CPM-TSP [6]	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

Table 2: A Comparison of our approach to other works on the Human 3.6m dataset

IMU. We observe competitive performance wrt. the state of the art although some actions perform poorly; this is likely due to the limited view (4) of Human3.6M affecting the PVH quality.

4.2 Total Capture

There are a number of high-quality hand labelled 2D human pose datasets [5, 20]. However, the hand labelling of 3D human pose is far more challenging and optical motion capture



ROM Sub. 1 Walking Sub. 2 Act Sub. 3 Running Sub 4 Freestyle Sub. 5

Figure 3: Examples of performance variation in the proposed TotalCapture dataset (cam. 1).

systems such as Vicon [3] are the only reliable method for ground truth labelling. This hardware constraint greatly reduces the viability of existing datasets; Table 3 shows the trade-offs between existing 3D human pose datasets. Human3.6M has a large amount of ground

Dataset	NumFrames	NumVideoCams	Vicon GT	IMU data
Human3.6M [16]	3,136,356	4	Y	N
HumanEva [27]	40,000	7	Y	N
TNT15 [22]	13,000	8	N	Y
Total Capture(Proposed)	1,892,176	8	Y	Y

Table 3: Characterising existing 3D human pose datasets and TotalCapture

truth labelled videos, but no IMU sensor data, while TNT15 has only a small amount of video frames, and is missing true Vicon ground truth labelling. HumanEva has a low number of frames, and no IMU data. Given the compromise in each dataset, we propose and release our 3D human pose dataset *TotalCapture*[1]; the first dataset to have fully synchronised video, IMU and Vicon labelling for a large number of frames ($\sim$ 1.9M), for many subjects, activities and viewpoints. The data was captured indoors in a volume measuring roughly 4x6m with 8 calibrated full HD video cameras recording at 60Hz on a gantry suspended at approximately 2.5 metres, with examples shown in Fig 3. The Vicon high-speed motion capture system [3] provides 21 pixel-accurate 3D joint positions and angles. Obtaining this ground-truth required visible markers to be worn, *however these are not used by our algorithm*. The size of these markers (0.5cm^3) is negligible relative to the volume and are not visible in the mattes and inconspicuous in the RGB images. While the XSens IMU system [25] consists of 13 sensors on key body parts, head, upper/lower back, upper/lower limbs and feet. Clean plates allow for accurate per pixel background subtraction and this is also made available. Total Capture consists of 4 male and 1 female subjects, each performing five diverse performances, repeated 3 times: *ROM, Walking, Acting, Running and Freestyle*. An example of each performance and subject variation is shown in Fig 3 and video.

The *acting* and *freestyle* performances, in particular, are very challenging with actions such as *yoga*, *giving directions*, *bending over* and *crawling*, see Fig 3. We partition the dataset wrt subjects and performance sequence, the training consists of performances: ROM1,2,3; Walking1,3; Freestyle1,2; Acting1,2; and Running1 on subjects 1,2 and 3. The test set is the performances Freestyle3 (**FS3**), Acting (**A3**) and Walking2 (**W2**) on subjects 1,2,3,4 and 5. This setup allows for testing on unseen and seen subjects but *always* unseen performances.

4.3 Total Capture Evaluation

To fully test and evaluate our approach we use the Total Capture dataset, with the volumetric vision, IMUs and fully connected fusion layer. We compare to two state of the art approaches, the 3D triangulated CPM, **Tri-CPM**, described in section 4.1 and a multi-view matte based 2D convolutional neural network approach [32], **2D Matte**, both with and without Temporal

[1]The TotalCapture dataset is available online at http://cvssp.org/data/totalcapture/.

Sequence Predictor (TSP) training. 2D Matte uses MVV to produce a PVH from which a spherical histogram [14] is used as input to an eight layer 2D convolution neural network. The performance of our approach on the Total Capture dataset using the 3D Euclidean error metric over the 21 joints is shown in table 4.

Approach	SeenSubjects(S1,2,3)			UnseenSubjects(S4,5)			Mean
	W2	FS3	A3	W2	FS3	A3	
Tri-CPM [6]	79.0	112.1	106.5	79.0	149.3	73.7	99.8
Tri-CPM-TSP [6]	45.7	102.8	71.9	57.8	142.9	**59.6**	80.1
2D Matte [32]	104.9	155.0	117.8	161.3	208.2	161.3	142.9
2D Matte-TSP [32]	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**

Table 4: Comparison of our approach on Total Capture to other human pose estimation approaches, expressed as average per joint error (mm).

The table shows how the performance of our proposed approach **Fused-DL IMU+3D PVH** greatly outperforms the performance of the previous approaches [6, 32], across a wide range of sequences & subjects, with a reduction of over 10mm error per joint. The ability of the TSP through the LSTM layers to effectively predict the joints is visible when comparing with & without the TSP, **3D PVH** and **3D PVH-TSP**, where the error is reduced by over 20mm.

Table 4 also shows the performance of the sub parts of the approach, **Solved IMU** uses the raw IMU orientations within the kinematic model described in section 3.2.2 and **Solved IMU-TSP** learns a TSP model on the solved IMU joint positions. Examining the IMU **Solved IMU-TSP**) and vision (**3D PVH TSP**) independently illustrates that through the fusion of the two modes around 10-20mm of per joint error reduction is achievable. This is likely to be due to the complementary performance of the two data sources. With respect to the fusion of the **Solved IMU-TSP** and **3D PVH-TSP**, we contrast our proposed fully connected layer fusion **Fused-DL IMU+3D PVH** with a simple mean of the joint estimates from the two data modes **Fused-Mean IMU+3D PVH**. Fig 4 quantifies the per frame error for the key techniques over the unseen subject S4 and performance FS3. Visually it can be seen that in the initial part of the sequence, the video based 3D PVH has a lower error than the solved IMU, however, after frame 1400 the 3D PVH increases in error and the IMU performs better. By fusing both modes we are able to have a consistently low error for the human pose estimation, with a smoother error compared to the high variance of the separate data modes. Fig 5 qualitatively shows the two modes and fused result for a selected number of frames. The differences between the inferred poses can be quite small, indicating the contribution of all components of the approach. Fig 6 and the video provide additional results. Run-time performance is 25fps, including PVH generation.

4.3.1 Training Data Volume

Within CNN based systems, the amount of data required to train effectively is a key concern. Therefore, we perform an ablation to explore the effect of the amount of training data on the accuracy. With the test sequences being kept consistent throughout as before, an increasing percentage of total available training data was used from Subjects 1, 2 and 3, randomly sampled from maximum of $\sim$ 250k MVV frames. At 20%, 40%, 60%, 80% the relative

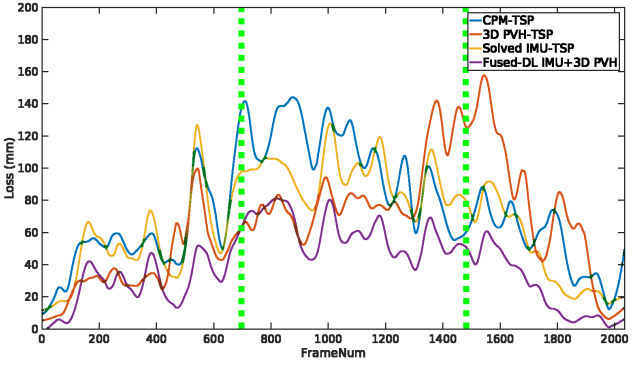


Figure 4: Per frame accuracy of our proposed approach on sequence FS3 Subject4 (Green dotted line indicates frame shown in examples in Fig 5).

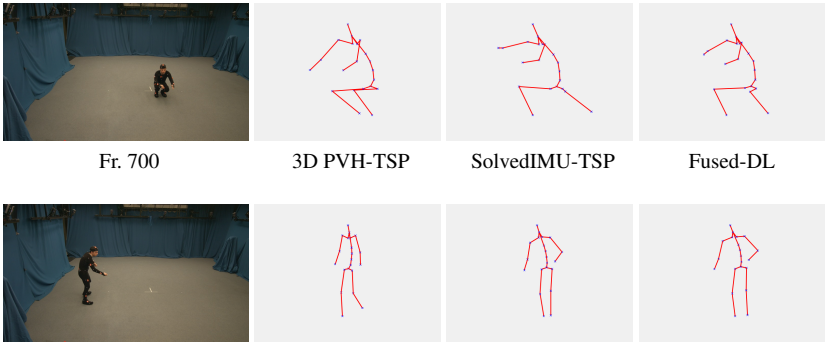


Figure 5: Visual comparison of poses resolved at different pipeline stages. TotalCapture: Freestyle3, Subject 4.

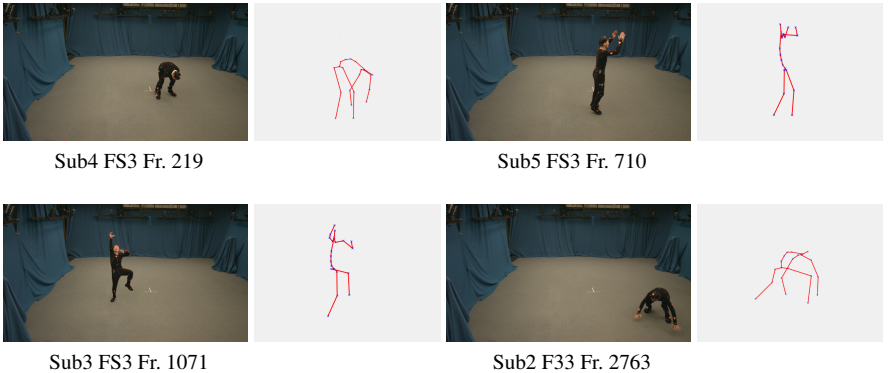


Figure 6: Additional results across diverse poses within TotalCapture. See video for more.

decrease in accuracy was 87.1%, 90.4%, 96.7% and 99.4% respectively. This suggests, for the purposes of CNN training, the range of motions in our dataset can be well represented by a relatively small sample, and that the internal model of the network can still generalise well

and without over-fitting having only seen a sparse set of ground truth poses.

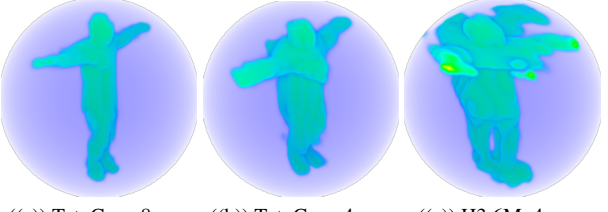
4.3.2 Analysis on Number of Cameras Used

We investigate the effect of the estimated 3D joint accuracy on the number of cameras used to construct the PVH. The experiment used 4, 6, and 8 cameras equally spaced around the volume, Table 5 shows the accuracy for the 3D PVH component, for the different subjects with increasing number of cameras. It shows there is only a minor impact on the performance

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%

Table 5: Relative accuracy change (mm/joint) when varying the number of cameras.

of the approach if the number of cameras is halved, still 90% performance with only 4 cams, despite the PVH becoming qualitatively worse in appearance, as illustrated in Fig 7. Likewise, Fig 7(c) shows a PVH for the Human3.6M dataset. It is more noisy due to the 4 cameras being closer to the ground, and noise on the mattes, however we still achieve state of the art performance.



((a)) Tot. Cap., 8 cams ((b)) Tot. Cap., 4 cams ((c)) H3.6M, 4 cams

Figure 7: Varying PVH fidelity of performer in the 'T' pose vs. camera count.

5 Conclusion

We have presented a novel algorithm for 3D human pose estimation that fuses video (MVV) and inertial (IMU) signals to produce a high accuracy pose estimate. We first outlined a 3D convnet for pose estimation from purely visual (MVV) data and showed how a temporal model (LSTM) can deliver state of the art results using this modality alone on a standard dataset (Human3.6M), with a per joint error of only 87.3mm. We next showed how the fusion of IMU data through a two-stream network, incorporating the LSTM, can further enhance accuracy with a 10mm improvement beyond state of the art, A further contribution was the TotalCapture dataset; the first publicly available dataset simultaneously capturing MVV, IMU and skeletal ground truth.

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.

References

[1] Optitrack motive. http://www.optitrack.com.

[2] Perception neuron. http://www.neuronmocap.com.

[3] Vicon blade. http://www.vicon.com.

[4] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016.

[5] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *ECCV'16*, 2016.

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *arXiv preprint arXiv:1412.3555*, 2014.

[8] Yann Dauphin, Harm de Vries, and Yoshua Bengio. Equilibrated adaptive learning rates for non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 1504–1512, 2015.

[9] K. Grauman, G. Shakhnarovich, and T. Darrell. A bayesian approach to image-based visual hull reconstruction. In *Proc. CVPR*, 2003.

[10] Alex Graves. Generating sequences with recurrent neural networks. In *arXiv preprint arXiv:1308.0850*, 2013.

[11] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *proceedings of the 31st International Confernce on Machine Learning (ICML)*, 2014.

[12] Thomas Helten, Meinard Muller, Hans-Peter Seidel, and Christian Theobalt. Real-time body tracking with one depth camera and inertial sensors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1105–1112, 2013.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural computation*, volume 9, pages 1735–1780. MIT Press, 1997.

[14] P. Huang, A. Hilton, and J. Starck. Shape similarity for 3d video sequences of people. *Intl. Journal of Computer Vision*, 2010.

[15] P. Huang, M. Tejera, J. Collomosse, and A. Hilton. Hybrid skeletal-surface motion graphs for character animation from 4d performance capture. *ACM Transactions on Graphics (ToG)*, 2015.

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[17] H. Jiang. Human pose estimation using consistent max-covering. In *Intl. Conf. on Computer Vision*, 2009.

[18] X. Lan and D. Huttenlocher. Beyond trees: common-factor model for 2d human pose recovery. In *Proc. Intl. Conf. on Computer Vision*, volume 1, pages 470–477, 2005.

[19] Sijin Li, Weichen Zhang, and Antoni B Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2848–2856, 2015.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015.

[22] Timo v Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Multimodal motion capture dataset tnt15. 2016.

[23] R Ren and J Collomosse. Visual sentences for pose retrieval over low-resolution cross-media dance collections. *IEEE Transactions on Multimedia*, 2012.

[24] X. Ren, E. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *Proc. Intl. Conf. on Computer Vision*, volume 1, pages 824–831, 2005.

[25] Daniel Roetenberg, Henk Luinge, and Per Slycke. Xsens mvn: full 6dof human motion tracking using miniature inertial sensors. In *http://www.xsens.com*, 2009.

[26] Marta Sanzari, Valsamis Ntouskos, and Fiora Pirri. Bayesian image based 3d pose estimation. In *European Conference on Computer Vision*, pages 566–582. Springer, 2016.

[27] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. In *International journal of computer vision*, volume 87, pages 4–27. Springer, 2010.

[28] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. *arXiv preprint arXiv:1605.05180*, 2016.

[29] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Fusing 2d uncertainty and 3d cues for monocular body pose estimation. *arXiv preprint arXiv:1611.05708*, 2016.

[30] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *arXiv preprint arXiv:1701.00295*, 2017.

[31] A. Toshev and C. Szegedy. Deep pose: Human pose estimation via deep neural networks. In *Proc. CVPR*, 2014.

[32] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep convolutional networks for marker-less human pose estimation from multiple views. In *Proceedings of the 13th European Conference on Visual Media Production (CVMP 2016)*, CVMP 2016, 2016.

[33] Timo von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and imus. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1533–1547, 2016.

[34] Timo von Marcard, Bodo Rosenhahn, Michael Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics)*, 2017.

[35] Timo von Marcard, Bodo Rosenhahn, Michael Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics)*, 2017.

[36] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li. Dense human body correspondences using convolutional networks. *CoRR*, abs/1511.05904, 2015.

[37] Ho Jung Yub, Yumin Suh, Gyeongsik Moon, and Kyoung Mu Lee. Sequential approach to 3d human pose estimation: Seperation of localization and identification of body joints. In *Proceedings of European Conference on Computer Vision (ECCV16)*, 2016.

[38] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4966–4975, 2016.