International Journal of Computer Vision (2019) 127:381–397
https://doi.org/10.1007/s11263-018-1118-y

# Fusing Visual and Inertial Sensors with Semantics for 3D Human Pose Estimation

**Andrew Gilbert**[1] · **Matthew Trumble**[1] · **Charles Malleson**[1] · **Adrian Hilton**[1] · **John Collomosse**[1]

Received: 23 November 2017 / Accepted: 27 August 2018 / Published online: 8 September 2018


**Abstract**

We propose an approach to accurately estimate 3D human pose by fusing multi-viewpoint video (MVV) with inertial measurement unit (IMU) sensor data, without optical markers, a complex hardware setup or a full body model. Uniquely we use a multi-channel 3D convolutional neural network to learn a pose embedding from visual occupancy and semantic 2D pose estimates from the MVV in a discretised volumetric probabilistic visual hull. The learnt pose stream is concurrently processed with a forward kinematic solve of the IMU data and a temporal model (LSTM) exploits the rich spatial and temporal long range dependencies among the solved joints, the two streams are then fused in a final fully connected layer. The two complementary data sources allow for ambiguities to be resolved within each sensor modality, yielding improved accuracy over prior methods. Extensive evaluation is performed with state of the art performance reported on the popular *Human 3.6M dataset* (Ionescu et al. in Intell IEEE Trans Pattern Anal Mach 36(7):1325–1339, 2014), the newly released *TotalCapture* dataset and a challenging set of outdoor videos *TotalCaptureOutdoor*. We release the new hybrid MVV dataset (TotalCapture) comprising of multi-viewpoint video, IMU and accurate 3D skeletal joint ground truth derived from a commercial motion capture system. The dataset is available online at http://cvssp.org/data/totalcapture/.

**Keywords** 3D pose estimation · Sensor fusion · Deep neural networks · Multi viewpoint video · Inertial measurement units

## 1 Introduction

Although challenging, marker-less real time 3D human pose estimation is attracting increasing research interest as it will deliver step changes to a wide range of fields, from biomechanics, psychology, animation, human computer interaction and computer vision. The desire is to regress and estimate a 3D location based limb skeleton of a human in a range of environments as shown in Fig. 1. However, 3D pose estimation suffers from a large number of challenges including large variation in appearance, arbitrary viewpoints and obstructed visibilities due to external entities and self-occlusions. To resolve these challenges effectively, marker based systems such as Vicon (http://www.vicon.com) or OptiTrack (http://www.optitrack.com) are commonly used to provide sufficient joint accuracy.

Communicated by Andreas Geiger.

✉ Andrew Gilbert
a.gilbert@surrey.ac.uk

[1] CVSSP, University of Surrey, Guildford GU2 7XH, UK

However, the requirement to wear a special suit or a large number of physical markers is intrusive and restricts both the performance environment and the range of motions the subject can perform. Also, heavy occlusion from other actors or props in the scene, or adverse illumination can cause these approaches to fail in practical deployments. Therefore approaches have tried to remove these constraints through the use of elaborate prior terms and body modelling (von Marcard et al. 2017), or with the use of depth cameras (Yub et al. 2016), or extending 2D estimation to 3D (Tome et al. 2017; Tan et al. 2017).

Nevertheless, such systems based purely upon computer vision, suffer from inaccuracies or are restricted by using complex priors. We propose a compromise, via the fusion of vision based 3D pose estimation and Inertial Measurement Units (IMUs) (Roetenberg et al. 2009, http://www.neuronmocap.com) to estimate pose accurately. IMUs are small boxes placed on key body parts that don't suffer from illumination or occlusion failures, IMUs, however, do suffer from drift and therefore cannot provide the full solution without the visual component. Given the complementary nature of the two modalities, we fuse vision and IMU to estimate

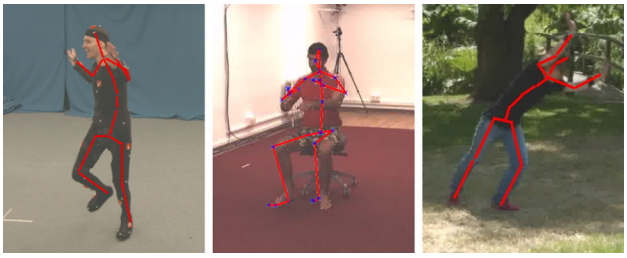


**Fig. 1** Our approach regresses 3D estimates for varied pose, subjects and environment

the 3D joint skeleton of human subjects. We show that by incorporating both cues, we can migiate the limitations of the drift and lack of spatial positional information in IMU data and the requirement of learnt complex human models for the vision. The complementary modalities mutually reinforce one another during inference; as rotational and occlusion ambiguities are mitigated by the IMUs while global positional drift and context are reduced by the vision.

Our proposed solution combines foreground occupancy mattes and semantic 2D pose estimates from a number of wide baseline video cameras to form a multi channel probabilistic visual hull (PVH) (Grauman et al. 2003). A coarse discretisation of the 3D space around the performer is then used to train a 3D convolutional network to predict 3D joint estimates from the volumetric PVH data. The contextual frame-wise temporal consistency of the 3D pose estimates is learnt with a variant of a Recurrent Neural Network (RNN) using LSTM layers. The LSTM learns a predictive model given a small number of previous frames. Concurrently IMUs are used to solve a simple kinematic model to provide a further 3D joint estimation, and both are then fused in an additional dense neural layer. The two data modes are illustrated in Fig. 2.

It is well known that training deep networks from scratch requires a large amount of data, and this requirement is heightened given the use of 3D convolutional layers in our work. Also, there is no single dataset available containing IMU and MVV video with a high-quality ground truth. Therefore we release a multi-subject, multi-action dataset as a further contribution to this work. The initial solution of this work was presented at BMVC 2017 (Gilbert et al. 2017). In this paper, we make several additional contributions. First, we enhance our initial 3D convolutional network for pose estimation through the incorporation of semantic pose information encoded in additional channels within volumetric data. We show that this information delivers a significant step-up in performance, resulting in an improved state of the art performance in both the public *TotalCapture* and *Human36M* datasets. In addition to deep analysis of these networks, we also introduce a novel dataset *TotalCaptureOutdoor* (Malleson et al. 2017) upon which we evaluate our system. The additional analysis within the experimental section (Sect. 4.5) allows greater insight into the contribution of the individual components while the methodology is expanded, allowing the reader further insight into our implementation.

## 2 Related Work

Human pose estimation 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.

Top-down approaches to 2D pose estimation fit an articulated limb model to data incorporating kinematics into the optimisation to bias toward possible configurations. Lan and Huttenlocher (2005) 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 Jiang (2009) using linear relaxation but performs well only on uncluttered scenes. The fusion of pictorial structures with Ada-Boost shape classification was explored in Andriluka et al. (2009). Agarwal and Triggs used non-linear regression to estimate pose in 2D silhouette images (Agarwal et al. 2004). The *SMPL* model (Loper et al. 2015) provides a rich statistical body model that can be fitted to incomplete data and von Marcard et al. (2017) incorporated IMU measurements with it to provide pose estimation without visual data. While (Tan et al. 2017) employs the SMPL model to estimate the 3D pose from 2D images in a decoder/encoder framework. Then, Huang et al. (2017) combines the SMPL body model with 2D joint estimates to reinforce and improve the 3D pose.

Bottom-up pose estimation is driven by image parsing to isolate components, Srinivasan and Shi (2007) used graph-cuts to parse a subset of salient shapes from an image and group these into a model of a person. Ren et al. (2005) recursively splits Canny edge contours into segments, classifying each as a putative body part using cues such as parallelism. Ren and Collomosse (2012) also used Bag of Visual Words for implicit pose estimation as part of a pose similarity system for dance video retrieval. More recently studies have begun to leverage the power of convolutional neural networks, following in the wake of the eye-opening results of Krizhevsky et al. (2012) on image recognition. Toshev and Szegedy (2014), in the DeepPose system, used a cascade of convolutional neural networks to estimate 2D pose in images. Descriptors learned by a CNN have also been used in 2D pose estimation from very low resolution images (Park and Ramanan 2015). Elhayek et al. (2015) used MVV with a Convnet to produce 2D pose estimations while Rhodin et al. (2016) minimised the edge energy inspired by volume ray casting to deduce the

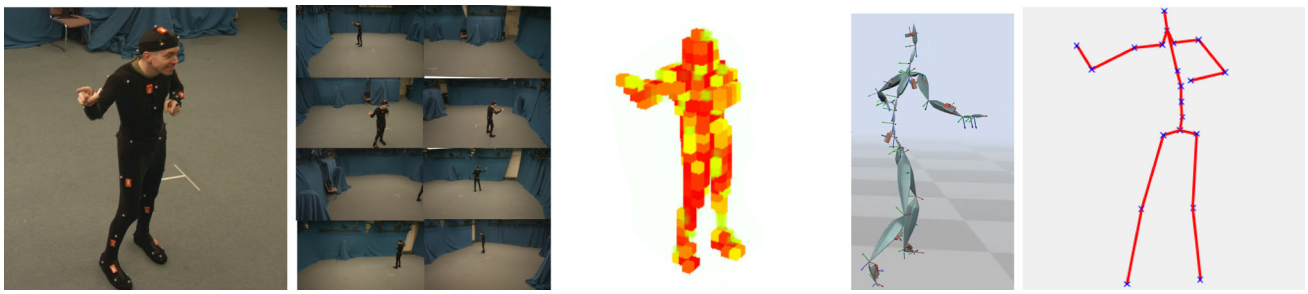



**Fig. 2** Our two-stream network fuses IMU data with volumetric (PVH) data derived from multiple viewpoint video (MVV) to learn an embedding for 3D joint locations (human pose)

3D pose. More recently given the success and accuracy of 2D joint estimation (Cao et al. 2016), an increasing number of works have been introduced to transfer those predictions into 3D, using a post processing optimisation step. Sanzari et al. (2016) estimates the location of 2D joints, before predicting 3D pose using appearance and probable 3D pose of the discovered parts with a hierarchical Bayesian model. While Zhou et al. (2016) integrates 2D, 3D and temporal information to account for uncertainties in the data. The challenge of estimating 3D human pose from MVV is currently less explored, although 3D pose estimation is generally cast as a coordinate regression task, with the target output being the spatial $x$, $y$, $z$ coordinates of a joint with respect to a known root node such as the pelvis. Trumble et al. (2016) used a flattened MVV based spherical histogram with a 2D convnet to estimate pose. While Pavlakos et al. (2017a) used a simple volumetric representation in a 3D convnet for pose estimation and Wei et al. (2016) performed related work in aligning pairs of joints to estimate 3D human pose. Differently, Huang et al. (2015) constructed a 4-D mesh of the subject from video reconstruction to estimate the 3D pose. While Tekin et al. (2016a) included a pretrained autoencoder within the network to enforce structural constraints.

Another challenge of MVV is the labelling of the training data, therefore Rogez and Schmid (2016) artificially augments a dataset of real images with 2D human pose annotations using 3D Motion Capture data. Given a candidate 3D pose, the algorithm selects for each joint an image whose 2D pose locally matches the projected 3D pose. Similarly Lassner et al. (2017) uses the SMPL (Loper et al. 2015) body model to generate training data without motion capture.

To predict temporal sequences, RNNs and their variants including LSTMs (Hochreiter and Schmidhuber 1997) and Gated Recurrent Units (Chung et al. 2014) have recently shown to learn and generalise the properties of temporal sequences successfully. Graves (2013) was able to predict isolated handwriting sequences, while in Natural language processing (NLP) Graves and Jaitly (2014) combines an LSTM model with Connectionist Temporal Classification objective function, directly transcribing audio data with text. Alahi et al. (2016) was also able to predict human trajectories of crowds by modelling each human with an LSTM and jointly predicting the paths.

In the field of IMUs, there has been a number of works that have used IMUs to estimate pose, Roetenberg Roetenberg et al. (2009), used 17 IMUs with 3D accelerometers, gyroscopes and magnetometers fused with a Kalman filter to define the pose of a subject. Slyper and Hodgins (2008) reconstructs pose using 5 accelerometers to retrieve pre-recorded poses with similar accelerations via a lookup process from a database. Acceleration data is however very noisy and the search space of possible accelerations is under constrained making the learning a very difficult task. While (Schwarz et al. 2009) directly regresses full pose using only 4 IMUs with a Gaussian Process regression, with good results when the test motions are present in the database. Similarly Pons-Moll et al. (2011) uses a particle filter framework to optimise the orientation constrained by IMU samples taken from a manifold of poses, to solve for outdoor sequences. Also, Liu et al. (2011) regress to a full pose querying a database of online local models based on the response of 6 IMUs.

The initial work to fuse IMU and video was by Pons-Moll et al. (2010), combining limb orientations from the inertial sensors, with stable and drift-free accurate position information from video data. While Marcard et al. (2016) fused video and IMU data to improve and stabilise full body motion capture. Helten et al. (2013) used a single depth camera with IMUs to track the entire body, with the IMUs identifying similar candidate poses and the depth data being used to obtain the full body estimate. Andrews et al. (2016) used a sparse set of labelled optical markers, IMUs, and a motion prior in an inverse dynamics formulation. While Malleson et al. (2017) used IMUs with a full kinematic solve to effectively estimate 3D pose indoor and outdoor.

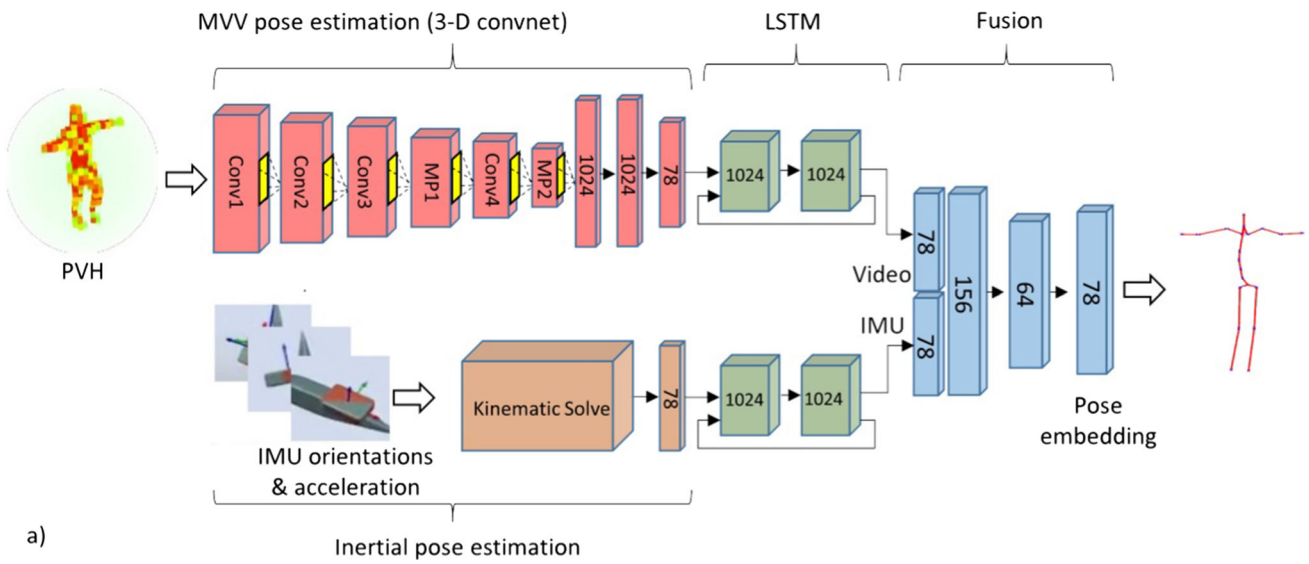



**Fig. 3** Network architecture comprising two streams: a 3D Convnet for **MVV pose embedding**, and kinematic solve from IMUs. Both streams pass through **LSTM** before the **Fusion** of the concatenated estimates in a further FC layer

## 3 Methodology

An overview of the approach is shown in Fig. 3, a 3D volumetric geometric proxy of the performer is formed from 2D foreground occupancy and 2D semantic heat maps, with a multi-channel probabilistic visual hull. This coarse visual hull is fed into a 3D convnet that directly regresses an embedding that encodes 3D spatial joint locations of the performer's body. A temporal model from a recurrent neural network is trained on the embedding to enforce temporal consistency to the 3D pose detections. Uniquely for this work, IMU data on key body parts is used to enable a forward kinematic solve of the pose that is smoothed with a learnt temporal RNN model. Given the complementary nature of the two data modes, a dense layer fuses both to provide a joint based embedding of the joint locations.

### 3.1 Volumetric Pose Embedding

Figure 3 shows a diagram of our architecture; it is based on a deep, multilayer neural network that consists of successive 3D convolutional and pooling layers. The goal of CNN pose regression is to obtain 3D Cartesian coordinates of $J$ joints given the multi-channel 3D probabilistic visual hull volume. The target of the network is $3 * J$-dimensional vector comprised of the concatenation of the $x, y, z$ coordinates of the $J$ joints of the human body, for our work $J = 17$, resulting in 51 final layer embedding $(3 * 17)$.

The detailed filter parameters are listed in Table 1 for each layer in Fig. 3. By using 3D convolution filters, we are able to encode information from all cameras as a volume simultaneously. In training, the network is supervised with an L2 regression loss:

$$\mathcal{L} = \sum_{j=1}^{J} \| p_{gt}^{j} - p_{pr}^{j} \|_2^2. \tag{1}$$

where $p_{gt}^{j}$ is the groundtruth location for joint $j$ and $p_{pr}^{j}$ is the predicted location for joint j. The location of each joint is expressed globally, normalised to a root joint or node at the pelvis. To further encourage pose invariance with respect to the facing direction of the performer, the training data is augmented by applying a random rotation about the central vertical axis, $\theta = [0, 2\pi]$.

### 3.2 Visual Channels

Two visual channels are employed, a 2D occupancy matte, and semantic 2D joints. The occupancy is a soft probability of foreground occupancy formed from the comparison of the current frame $I$ and a clean-plate $P$ taken before the recorded sequence. The thresholded L2 distance between the two images in the HSV colour domain provides the soft occupancy probability for the 1st channel. The second semantic channel consists of a human joint belief labels estimated by OpenPose (Wei et al. 2016; Cao et al. 2017), a multi-stage process that iteratively refines 2D pose estimations of joint positions using a mixture of knowledge of the image and the estimates of joint locations of the previous stage. At each stage $s$ and for each joint label $j$ the algorithm returns dense per pixel belief maps $m_s^j$, which provides the confidence of a joint centre for any given pixel $(x, y)$, and given stage $s$. Much of the algorithm's power is that in stages $s \in 2, ..., S$

**Table 1** Parameters of the 3D Convnet used to infer the MVV pose embedding

| 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 |

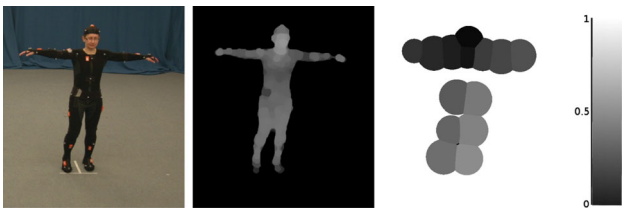



**Fig. 4** An example of the foreground occupancy and 2D joint label belief map (white indicates high probability of occupancy)

the belief maps are a function of not just the information contained in the image but also the information computed by the previous stage. For this work we transform these per joint belief maps into a single label image $M$, by maximising over the confidence of all possible joint labels on a per pixel basis.

$$M(x, y) = \underset{J}{\operatorname{argmax}}\, m_S^j(x, y) \tag{2}$$

Figure 4 shows an example of the soft occupancy and joint labels for an example image.

### 3.3 Volumetric Representation of Proxy

Many recent approaches use multiple 2D views (Pavlakos et al. 2017b) or infer 3D from a learnt 2D lookup (Tome et al. 2017; Chen and Ramanan 2017). However, we propose to simultaneously use multiple 2D views to produce a crude but accurate 3D representation of the human body. Integrating the multiple views into a 3D shape overcomes the unavoidable ambiguities and occlusions present in individual 2D images. However, the cost is the exponential increase in dimensionality over 2D, and also the lack of a pre-trained imagenet based model (Krizhevsky et al. 2012). Therefore to allow the training to be tractable and still provide the increase in detail over 2D, we propose to use a multi-channel based probabilistic visual hull (PVH) (Grauman et al. 2003) to infer the 3D occupancy shape from multiple camera views. A PVH quantises the volume occupancy in a soft probabilistic computation that greatly reduces the dimensionality while maintaining the detail. The volumetric representation is agnostic to the source of the data, and for this work, we propose to use both 2D foreground occupancy mattes and semantic 2D joint labels. Both are noisy and contain failure cases as a single view. However, the probabilistic nature of the PVH ensures that noise is ignored and only a consistent signal is propagated to the 3D volume.

Given a set of $C$ wide baseline cameras, $c = [1, \ldots, C]$, where $C > 3$ surrounding a performance volume, and calibrated with a known orientation, $\mathbf{R}_c$, focal point $COP_c$, focal length $f_c$ and optical centre $o_c^x, o_c^y$, the camera parameters for a given camera $c$ are

$$\{\mathbf{R}_c, COP_c, f_c, o_c^x, o_c^y\} \tag{3}$$

The 3D Capture Volume is finely decimated into voxels $v = [1, \ldots, V]$ approximately 10 mm$^3$ in size. Then given an 2D image denoted as $I_c$, with $\Phi = [1, \ldots, \phi]$ channels the voxel occupancy from a given camera view $c$ is defined as the probability:

$$p(V|c) = I_c(x[v_i], y[v_i], \phi) \tag{4}$$

where given a 2D image coordinate position $(x, y)$ the voxel $v_i$ projects to a real world 3D position of:

$$x[v_i] = \frac{f_c v_i^x}{v_i^z} + o_c^x \quad \text{and} \quad y[v_i] = \frac{f_c v_i^y}{v_i^z} + o_c^y, \tag{5}$$

$$\text{where} \quad \begin{bmatrix} x & y & z \end{bmatrix} = COP_c + R_c^{-1} v_i. \tag{6}$$

where $\begin{bmatrix} x & y & z \end{bmatrix}$ is the 3D real world global coordinate location. Therefore the overall probability of occupancy for a given voxel $p(v, \phi)$ is the product over all views:

$$p(v_i, \phi) = \prod_{i \in C}^{C} p(v|c), \tag{7}$$

this is then computed for all voxels in the volume

$$\sum_{i \in V} \sum_{j \in \Phi} p(v_i, \phi_j) \tag{8}$$

The fine grained voxel occupancy approximation is then down sampled via a weighted Gaussian filter to the coarse input shape and size of the first layer in the convnet, $30 \times 30 \times 30$, this roughly approximates with the same number of pixels as a $150 \times 150$ 2D image, where each voxel approximates a $67 \times 67 \times 67$ mm volume in the real world.

### 3.4 Inertial Pose Estimation

To estimate the pose from joint orientations, Xsens IMUs (Roetenberg et al. 2009) are placed on key body parts to estimate the pose. The end rigid joints provide the most discriminative data and will constrain the pose parameters effectively when fused later with the vision. The pose optimisation of Malleson et al. (2017) is used, this aims to minimize the energy of the following Equation:

$$E(\theta) = \overbrace{E_R(\theta) + E_A(\theta)}^{Data} + \overbrace{E_{PP}(\theta) + E_{PD}(\theta)}^{Prior} \tag{9}$$

where $E_R(\theta)$, and $E_A(\theta)$ are orientation and acceleration constraints, respectively and $E_{PP}(\theta)$ and $E_{PD}(\theta)$ are the pose projection and pose deviation priors, respectively.

For each IMU, $k \in [1, 13]$, we assume rigid attachment to a bone and calibrate the relative orientation, $\mathbf{R}^k_{kb}$, between the IMU $k$ and the bone $b$. The reference frame of the IMUs, $\mathbf{R}_{kw}$, is also calibrated approximately against the global world $w$ coordinates. Each local IMU orientation measurement, $\mathbf{R}^k_m$, is transformed to a global bone orientation, $\mathbf{R}^k_b$ as follows:

$$\mathbf{R}^k_b = (\mathbf{R}^k_{kb})^{-1}\mathbf{R}^k_{kw}\mathbf{R}^k_m \tag{10}$$

Then the local (hierarchical) joint rotation, $\mathbf{R}^k_h$, for a given bone $b$ in the skeleton is inferred by the kinematic chain:

$$\mathbf{R}^k_h = \mathbf{R}^k_b(\mathbf{R}^{par(b)}_b)^{-1} \tag{11}$$

where $par(b)$ is the parent of bone $b$. The forward kinematics begins at the root and proceeds down the joint tree (with unmeasured bones kept fixed).

In addition to orientation, the IMUs provide local acceleration measurements and a window of three frames, $t$ (current frame), and previous two frames $t1$ and $t2$ is used. For each IMU, a constraint is added which seeks to minimize the difference between the measured and solved acceleration of the track target site. The solved acceleration is computed using central finite differences using the solved pose from previous two frames along with the current frame being solved. The local accelerations from the previous frames of IMU data are converted to global coordinates in a similar method to Eq. 10 but gravity is also removed.

We use two priors based on the PCA of the pose: PCA projection ($E_{PP}$) and PCA deviation ($E_{ED}$). The projection prior encourages the solved body pose to lie close to the reduced dimensionality subspace of prior poses (a soft reduction in the degrees of freedom of the joints), while the deviation prior discourages deviation from the prior observed pose variation (soft joint rotation limits). Together these terms produce soft constraints that yield plausible motion while not strictly enforcing a reduced dimensionality on the solved pose, thus allowing novel motion to be more faithfully reproduced at run time. For full details of the cost functions used please see Malleson et al. (2017).

These joint orientations in conjunction with the calibrated performer's skeleton allow for joints locations to be inferred to a concatenated joint vector $\mathbf{J}_i$. For a more detailed description of relating inertial data to other sensor model coordinate systems the work by Baak et al. (2010) can provide further details. To temporally align the IMU and video data an initial foot stamp was performed by the subject, which was visible in the video and produces a strong peak in acceleration in the IMU data. The inertial reference frame of each IMU, $\mathbf{R}^k_{kw}$ is assumed to be consistent between IMUs and in alignment with the world coordinates through the global up direction and magnetic north. The IMU-bone positions $t_{kb}$ are specified by manual visual alignment and the IMU-bone orientations Rib are calibrated using the measured orientations with the subject in a known pose (the T-pose, facing the direction of a given axis).

### 3.5 Learnt Temporal Consistency

Given the temporal nature of human pose sequences, it is desirable to learn and enforce temporal consistency on the two streams of per frame pose estimation. Thus allowing the rich temporal motion patterns between frames and joints to be effectively incorporated into the 3D pose prediction. Long Short Term Memory (LSTM) layers (Hochreiter and Schmidhuber 1997) have provided excellent performance in exploiting longer term temporal correlations compared to standard recurrent neural networks on many tasks, e.g. speech recognition (Sak et al. 2014) and video description (Donahue et al. 2015). LSTM layers can store and access information over long periods of time but mitigate the vanishing gradient problem common in RNNs through a specialised gating mechanism.

Given an input vector $\mathbf{J}_i(t)$ at time $t$ consisting of concatenated joint spatial coordinates and resulting output joint vector $\mathbf{J}_o(t)$. The aim is to learn the function that minimises the loss between the input vector and the output vector $\mathbf{J}_o = o_t \circ tanh(c_t)$ ($\circ$ denotes the Hadamard product), $o_t$ is the output gate, and $c_t$ is the memory cell, a combination of the previous memory $c_{t-1}$ multiplied by a *forget gate*, and the input gate as shown in Fig. 5. Thus, intuitively it is a combination of the previous memory and the new input. For example, the old memory could be completely ignored (forget gate all 0s) or ignore the newly computed state completely (input gate all 0s), but in practice it is of course between those two extremes. The memory cell $c_t$ is shown in Eq. 12.

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(\mathbf{J}_i(t)U_g + \mathbf{J}_i(t-1)W_g) \tag{12}$$

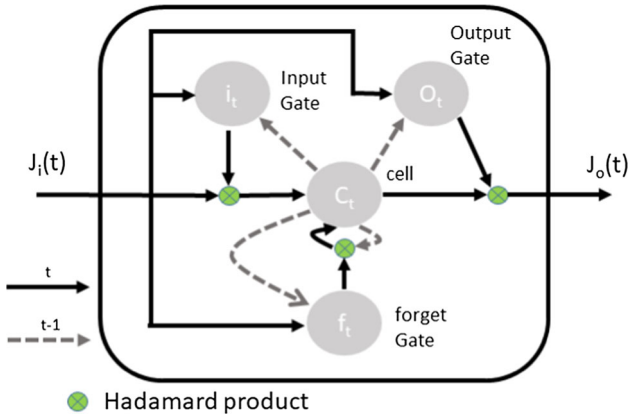



**Fig. 5** The design and connections of an LSTM layer.

Within each gates in there are two weights that are learnt, $W$ and $U$. The input gate $i_t$ defines the extent to which the newly computed state for the current input $\mathbf{J}_i(t)$ is kept in the memory,

$$i_t = W_i \mathbf{J}_i(t) + U_i \mathbf{J}_i(t-1) \tag{13}$$

A forget gate $f_t$ defines how much of the previous state remains in memory,

$$f_t = W_f \mathbf{J}_i(t) + U_f \mathbf{J}_i(t-1) \tag{14}$$

and an output gate $o_t$ defines how much of the internal state is exposed to the external network (higher layers and the next time step).

$$o_t = W_o \mathbf{J}_i(t) + U_o \mathbf{J}_i(t-1) \tag{15}$$

To learn the weights, they are trained using back propagation employing the loss function from Eq. 1 Each data modality has a distinct layer, with the temporal consistency using the previous $f$ frames to predict the current frame joint vector for both the visual and IMU pose based estimation. With two layers both with 1024 memory cells, a look back of $f = 5$ and a learning rate of $10^{-3}$ trained with RMS-prop (Dauphin et al. 2015).

### 3.6 Modality Fusion

The vision and IMU sensors both independently provide a 3D coordinate per joint estimate to reconstruct the performer's pose. Therefore, it would make sense to incorporate both modes into the final estimate, given their complementary nature. Naively, an average pool of the two joint estimates could be used; this would be fast and efficient assuming both modalities have small errors. However, it is likely that often significant errors will be present on one of the modes due to their different measurement approaches. We, therefore, propose to fuse the two modes with a further fully connected layer. We are able to utilise the idea of using a dense layer to fuse our visual and IMU joint skeleton predictions, that can combine both measurements in a more meaningful way than simply taking the average. This allows errors in the pose from the vision and IMU to be identified and corrected by the combined fused model. This fully connected dense layer consists of 64 units and was trained with an RMS-prop optimiser (Dauphin et al. 2015) with a learning rate of $10^{-4}$ to provide the feedback to reinforce the prediction. All stages of the model are implemented using Tensorflow.

## 4 Evaluation

To provide an evaluation of our approach we employ three different datasets. First we present results on the multichannel vision stream only of the approach on the *Human3.6M* (Ionescu et al. 2014) dataset in Sect. 4.1 without the IMU fusion. We then introduce our new dataset called *TotalCapture* (Gilbert et al. 2017) in Sect. 4.2, which contains both video and IMU with the associated GT joint skeleton. We evaluate our full fused vision and IMU approach on the TotalCapture dataset and we also perform an ablation study in Sect. 4.4 to examine the individual contributions of our work. Finally, we evaluate the ability of our approach to generalise to new sequences by evaluating on the challenging *TotalCaptureOutdoor* (Malleson et al. 2017) in Sect. 5 a challenging collection of sequences of MVV and IMU captured in a challenging outdoor environment.

### 4.1 Human 3.6M

We evaluate 3D pose estimation on the Human 3.6M dataset (Ionescu et al. 2014) where 3D ground truth key points are available from a marker-based motion capture system. It consists of 3.6 million video frames captured on four camera viewpoints in a 360-degree arrangement. There are five female and six male subjects, performing typical activities such as posing, sitting and giving directions. There is no IMU data within the dataset, and so we only evaluate the visual component, the PVH + LSTM. This is the upper red and green layers from Fig. 3 without the fusion of the IMU kinematic solve. To allow comparison to other approaches we follow the same data partition protocol as in previous works (Ionescu et al. 2014; Li et al. 2015; Tekin et al. 2016, a; Tome et al. 2017; Gilbert et al. 2017). The training data consists of subjects S1, S5, S6, S7, S8 and it is tested on unseen subjects S9, S11. The standard 3D Euclidean error metric is used to evaluated accuracy, it calculates the Euclidean error averaged over all frames and 17 joints (in human 3.6M) in millimetres (mm). The Results of our multi-channel 3D volumetric

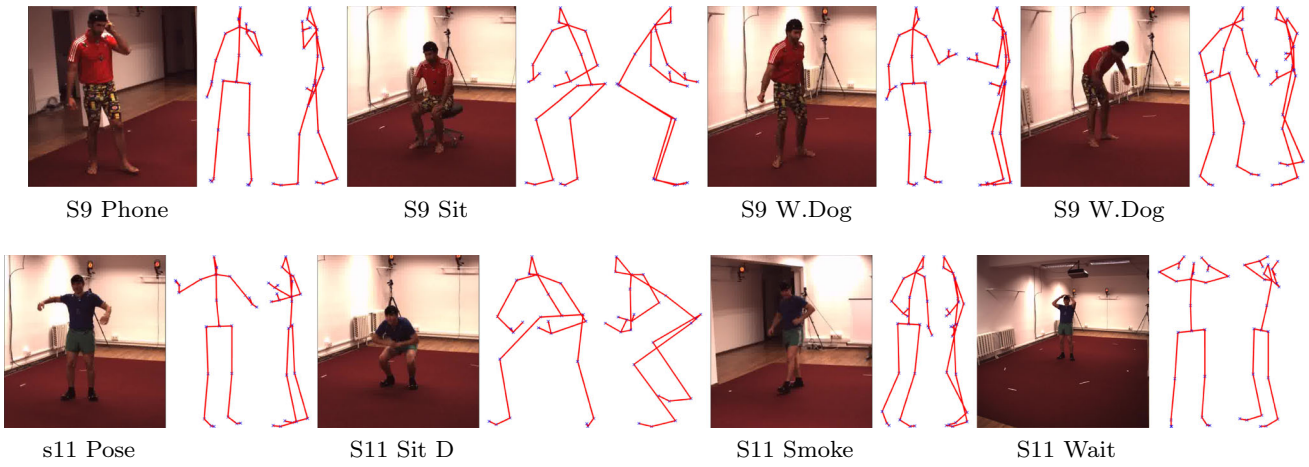



**Fig. 6** Example pose estimates from the Human 3.6M dataset from two viewpoints

approach with the temporal consistency are evaluated qualitatively in Fig. 6 and quantitatively in Table 2, in particular we compare to the approach of Mude Lin Liang Lin and Cheng (2017) who use 2D joint estimates with a 3D recurrent network, Tome et al. (2017), which infers 3D probabilistic estimates from monocular 2D joint predictions. Also we compare to a baseline approach *Tri CPM LSTM*, a 3D triangulated version of the 2D pose estimation (Cao et al. 2016) with error rejection. In this approach per camera 2D joint estimates

$$\mathbf{J}_{cpm} = \underset{x,y}{\operatorname{argmax}}\, m_S^j(x, y) \tag{16}$$

are triangulated into a 3D point, using an error rejection method that maximises the number of 2D estimates with the lowest 3D re-projection error. This is a frame wise detection based approach, and therefore temporal consistency is introduced with two learnt LSTM layers as described in Sect. 3.5, Tri CPM LSTM.

As one can see from Table 2, our proposed approach outperforms all compared methods at time of publication [the newer works of Martinez et al. (2017) and Trumble et al. (2018)] indicate the speed of improvement in field of 3D pose estimation) despite excluding the fusion with the kinematic based IMU, with the mean error reduced by 15% compared with (Tome et al. 2017), the Tri CPM LSTM approach and our previous method (Gilbert et al. 2017). While compared to the state of the art results by Mude Lin Liang Lin and Cheng (2017), many activities have a similar error around 5 or 6cm. However, there is marked performance improvement in our approach for the activities; dog walking and sitting down, while Lin achieves better performance for greeting and waiting. Qualitative comparison to the ground truth is shown in Fig. 6, it shows the high degree of accuracy achievable, representing complex human poses. Although as shown in the bottom right pose, some unusual poses, probably not sufficiently represented in the training data, are still poorly estimated. To validate the superiority of the proposed multi-channel and temporally consistent approach, we evaluate the Human3.6M dataset with separate parts of the approach in Table 3.

It can be seen that the single channels of Matte or CPM based PVH perform worse than the multi-channel PVH, with both channels combined. This is likely to be due to the semantic information of the CPM labels complementing the occupancy based soft mattes. Also the improvement for enforcing the temporal consistency through the LSTM can be seen to be around 25 mm on average.

### 4.2 Total Capture

In recent years, high quality labelled datasets have been a catalyst for rapid development in a number of areas including object recognition (Deng et al. 2009) and 2D human pose datasets (Andriluka et al. 2014; Lin et al. 2014). These have been hand labelled, providing excellent accuracy and detail, however, this is far harder in 3D, where the labelling still in general relies on expensive and less common optical motion capture systems such as (http://www.vicon.com). This constraint greatly reduces the quantity and variability of existing datasets; Table 4 shows the features of current 3D human pose datasets. As can be seen Human3.6M has a large amount of synchronised multi-view video and is popular, however no IMU sensor data. HumanEva, is a smaller dataset also with no IMU information. While TNT15, contains IMU data and MVV it is small in size. Given these restrictions, we propose a new dataset to address these short comings, *TotalCapture*.[1] It contains a large amount of MVV, and synchronised IMU

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

**Table 2** A comparison of our approach to other works on the Human 3.6M dataset, multiview indicates whether the approach uses multiple camera views [the works of Martinez et al. (2017) and Trumble et al. (2018) where published after the time of submission]

| Approach | Multiview | Direct. | Discus | Eat | Greet. | Phone | Photo | Pose | Purch. |
|---|---|---|---|---|---|---|---|---|---|
| Li et al. (2015) | Y | 132.7 | 183.6 | 132.4 | 164.4 | 162.1 | 205.9 | 150.6 | 171.3 |
| Tekin et al. (2016) | Y | 85.0 | 108.8 | 84.4 | 98.9 | 119.4 | 95.7 | 98.5 | 93.8 |
| Zhou et al. (2016) | N | 87.36 | 109.31 | 87.05 | 103.16 | 116.18 | 143.32 | 106.88 | 99.78 |
| Sanzari et al. (2016) | N | 48.82 | 56.31 | 95.98 | 84.78 | 96.47 | 105.58 | 66.30 | 107.41 |
| Tome et al. (2017) | N | 65.0 | 73.5 | 76.8 | 86.4 | 86.3 | 110.7 | 68.9 | 74.8 |
| Tri CPM LSTM (Cao et al. 2016) | Y | 67.4 | 71.9 | 65.1 | 108.8 | 88.9 | 112.0 | 55.6 | 77.5 |
| Gilbert et al. (2017) | Y | 92.7 | 85.9 | 72.3 | 93.2 | 86.2 | 101.2 | 75.1 | 78.0 |
| Mude Lin Liang Lin and Cheng (2017) | N | 58.0 | 68.3 | 63.3 | 65.8 | 75.3 | 93.1 | 61.2 | 65.7 |
| Martinez et al. (2017) | Y | 74.0 | 94.6 | 62.3 | 59.1 | 65.1 | 49.5 | 52.4 | 62.9 |
| Trumble et al. (2018) | Y | 61.0 | 95.0 | 70.0 | 62.3 | 66.2 | 53.7 | 52.4 | 62.5 |
| Proposed | Y | 61.2 | 63.0 | 58.6 | 91.2 | 76.3 | 91.1 | 59.7 | 68.3 |
| | Multiview | Sit. | Sit D | Smoke | Wait | W.Dog | Walk | W. toget. | Mean |
| Li et al. (2015) | Y | 151.6 | 243.0 | 162.1 | 170.7 | 177.1 | 96.6 | 127.9 | 162.1 |
| Tekin et al. (2016) | Y | 73.8 | 170.4 | 85.1 | 116.9 | 113.7 | 62.1 | 94.8 | 100.1 |
| Zhou et al. (2016) | N | 124.52 | 199.23 | 107.42 | 118.09 | 114.23 | 79.39 | 97.70 | 113.01 |
| Sanzari et al. (2016) | N | 116.89 | 129.63 | 97.84 | 65.94 | 130.46 | 92.58 | 102.21 | 93.15 |
| Tome et al. (2017) | N | 110.2 | 173.9 | 85.0 | 85.8 | 86.3 | 71.4 | 73.1 | 88.4 |
| Tri CPM LSTM (Cao et al. 2016) | Y | 92.7 | 110.2 | 80.3 | 100.6 | 71.7 | 57.2 | 77.6 | 88.1 |
| Gilbert et al. (2017) | Y | 83.5 | 94.8 | 85.8 | 82.0 | 114.6 | 94.9 | 79.7 | 87.3 |
| Mude Lin Liang Lin and Cheng (2017) | N | 98.7 | 127.7 | 70.4 | 68.2 | 73.0 | 50.6 | 57.7 | 73.1 |
| Martinez et al. (2017) | Y | 74.0 | 94.6 | 62.3 | 59.1 | 65.1 | 49.5 | 52.4 | 62.9 |
| Trumble et al. (2018) | Y | 61.0 | 95.0 | 70.0 | 62.3 | 66.2 | 53.7 | 52.4 | 62.5 |
| Proposed | Y | 76.2 | 93.4 | 71.2 | 85.0 | 64.5 | 53.1 | 67.1 | 71.9 |

**Table 3** Empirical study on the performance of the different parts of the approach on the Human 3.6M dataset

| Approach | Direct. | Discus | Eat | Greet. | Phone | Photo | Pose | Purch. |
|---|---|---|---|---|---|---|---|---|
| 3D Matte PVH | 152.8 | 171.4 | 152.6 | 189.2 | 179.7 | 210.2 | 147.1 | 167.0 |
| 3D CPM PVH | 104.9 | 108.0 | 100.5 | 156.3 | 130.7 | 156.1 | 102.3 | 117.1 |
| 3D Matte CPM PVH | 83.1 | 85.5 | 79.5 | 123.8 | 103.5 | 123.6 | 81.0 | 92.7 |
| 3D Matte CPM PVH LSTM (ours) | 61.2 | 63.0 | 58.6 | 91.2 | 76.3 | 91.1 | 59.7 | 68.3 |
| | Sit. | Sit D | Smke | Wait | W.Dog | Walk | W. toget. | Mean |
| 3D Matte PVH | 177.3 | 192.8 | 179.3 | 161.0 | 236.8 | 179.0 | 168.8 | 169.0 |
| 3D CPM PVH | 130.6 | 160.1 | 122.0 | 145.6 | 110.5 | 91.0 | 115.1 | 123.4 |
| 3D Matte CPM PVH | 103.4 | 126.7 | 96.6 | 115.2 | 87.5 | 72.0 | 91.1 | 97.7 |
| 3D Matte CPM PVH LSTM (ours) | 76.2 | 93.4 | 71.2 | 85.0 | 64.5 | 53.1 | 67.1 | 71.9 |

and Vicon labelling for ground truth. It was captured indoors in a volume measuring roughly $8 \times 4$ m with 8 calibrated HD video cameras at 60 Hz. The variation in the dataset is shown in Fig. 7. To provide accurate labelled ground truth, the optical marker based (http://www.vicon.com) system was utilised, calculating 17 3D joint positions and angles, by triangulating small ($0.5\,\text{cm}^3$) dots visible to infrared cameras, note these dots are not used explicitly by our algorithm, and their size is negligible compared to the performance volume. The IMU data is provided by 13 sensors on key body parts, head, upper/lower back, upper/lower arms and legs and feet, providing per unit orientation and acceleration. The location of the IMU sensors is shown in Fig. 8. The dataset consists of four male and one female subjects each performing four diverse performances, repeated three times: *ROM, Walking, Acting and Freestyle*, with each sequence

**Table 4** Characterising existing 3D human pose datasets and TotalCapture

| Dataset | Frames | Cams | Vicon | IMU |
|---|---|---|---|---|
| Human3.6M (Ionescu et al. 2014) | 3,136, 356 | 4 | Y | N |
| HumanEva (Sigal et al. 2010) | 40,000 | 7 | Y | N |
| TNT15 (Marcard et al. 2016) | 13,000 | 8 | N | Y |
| Total capture | 1,892,176 | 8 | Y | Y |

lasting around 3000–5000 frames. An example of each performance and subject variation is shown in Fig. 7. There is a total of 1,892,176 frames of synchronised video, IMU and Vicon data (although some are withheld as test footage for unseen subjects). The variation and body motions contained in particular within the *acting* and *freestyle* sequences are very challenging with actions such as *yoga*, *giving directions*, *bending over* and *crawling* performed in both the train and test data. The train and test partitions are performed wrt to the subjects and sequences, the training consists of ROM1, 2, 3; Walking1, 3; Freestyle1, 2 and Acting1, 2 on subjects 1, 2 and 3. The test set is the performances Freestyle3 (**FS3**), Acting (**A3**) and Walking 2 (**W2**) on subjects 1, 2, 3, 4 and 5. This split allows for a comparison of unseen and seen subjects but always unseen sequences.

### 4.3 Total Capture Evaluation

To provide a reference of our approach to other methods we compare to three state of the art approaches, the 3D triangulated CPM, Tri-CPM, described in Sect. 4.1 a flattened multi-view matte based 2D convolutional neural network approach (Trumble et al. 2016), 2D Matte, and our previously published results without the semantic 2D pose labels in the probabilistic visual hull (Gilbert et al. 2017). The results are shown with and without temporal consistency provided by the learnt LSTM model. As with Human3.6M, we show performance using the 3D Euclidean error metric over the 17 joints quantitatively in Table 5, and then qualitatively in Fig. 9 and in the accompanying video (The video is available at http://youtu.be/CLDqpze53lU). The table shows that our combined semantic and occupancy based fusion with IMU approach outperforms all other methods, including our previous work (Gilbert et al. 2017) by 6 mm, and the triangulated CPM by 13 mm, which also performed well on the Human3.6M. The ability of the LSTM layers to introduce the temporal consistency and remove failure cases, improves all approaches by around 20 mm.

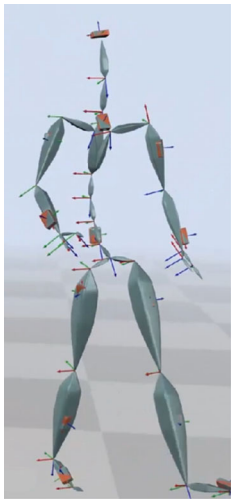


**Fig. 8** The location of the 13 orange box IMU sensors

### 4.4 Ablation Study

Our ablation study cumulatively enables each of our individual contributions on top of a classic baseline of a 3D Matte PVH. 3D pose estimation performance error is presented in Table 6 for separate parts of the approach.

The table shows how that the two channels of the PVH, 3D Matte PVH and 3D CPM PVH separately have a similar per-

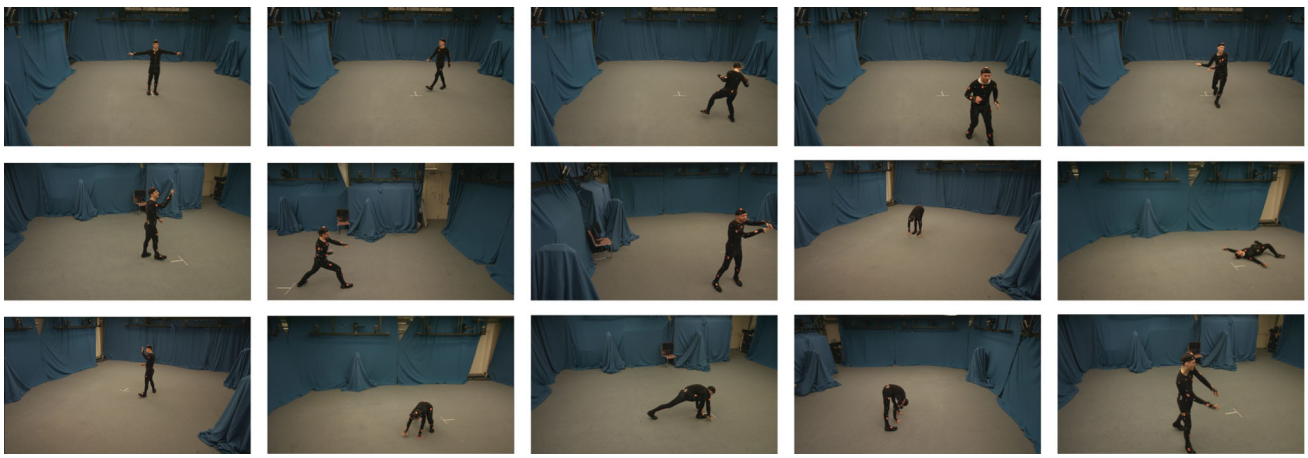


**Fig. 7** Examples of performance variation in the proposed TotalCapture dataset

**Table 5** Comparison of our approach on TotalCapture to other human pose estimation approaches, expressed as average per joint per frame error (mm)

| Approach | SeenSubjects(S1,2,3) | | | UnseenSubjects(S4,5) | | | Mean |
|---|---|---|---|---|---|---|---|
| | W2 | FS3 | A3 | W2 | FS3 | A3 | |
| Tri-CPM (Cao et al. 2016) | 79.0 | 112.1 | 106.5 | 79.0 | 149.3 | 73.7 | 99.8 |
| Tri-CPM-LSTM (Cao et al. 2016) | 45.7 | 102.8 | 71.9 | 57.8 | 142.9 | 59.6 | 80.1 |
| 2D Matte (Trumble et al. 2016) | 104.9 | 155.0 | 117.8 | 161.3 | 208.2 | 161.3 | 142.9 |
| 2D Matte-LSTM (Trumble et al. 2016) | 94.1 | 128.9 | 105.3 | 109.1 | 168.5 | 120.6 | 121.1 |
| 3D Matte PVH + IMU-LSTM (Gilbert et al. 2017) | 30.0 | 90.6 | 49.0 | 36.0 | 112.1 | 109.2 | 70.0 |
| Ours | 19.2 | 48.8 | 42.3 | 24.7 | 61.8 | 58.8 | 42.6 |


**Fig. 9** Additional results across diverse poses within TotalCapture. The two skeleton results shown the joint estimates from a different camera views

**Table 6** Mean per joint error (mm) of the approach components on the TotalCapture Dataset

| Approach | SeenSubjects(S1,2,3) | | | UnseenSubjects(S4,5) | | | Mean |
|---|---|---|---|---|---|---|---|
| | W2 | FS3 | A3 | W2 | FS3 | A3 | |
| 3D Matte PVH | 48.3 | 122.3 | 94.3 | 84.3 | 168.5 | 154.5 | 107.3 |
| 3D RGB Matte PVH | 57.0 | 133.8 | 102.2 | 90.2 | 176.3 | 157.7 | 115.2 |
| 3D CPM PVH | 85.5 | 123.1 | 88.6 | 105.7 | 142.2 | 97.7 | 105.8 |
| 3D Matte CPM PVH | 66.0 | 93.3 | 75.2 | 78.1 | 114.2 | 100.0 | 85.9 |
| 3D Matte CPM PVH-LSTM | 52.8 | 80.9 | 62.1 | 61.4 | 102.6 | 90.0 | 73.0 |
| Raw IMU-LSTM | 84.3 | 138.5 | 102.4 | 85.1 | 168.1 | 158.1 | 122.75 |
| Solved IMU | 38.5 | 60.5 | 68.7 | 48.0 | 89.5 | 80.0 | 64.2 |
| Solved IMU-LSTM | 29.8 | 50.7 | 59.8 | 32.4 | 64.1 | 74.5 | 51.9 |
| Averaged fused approach | 25.4 | 50.6 | 57.2 | 25.9 | 63.4 | 68.5 | 48.7 |
| Dense layer fused approach | 19.2 | 48.8 | 42.3 | 24.7 | 61.8 | 58.8 | 42.6 |

formance error, however by employing a two channel PVH it is possible to reduce the error by 20 mm. We also show the accuracy of using a 3 channel PVH (3D RGB Matte PVH ) with the foreground RGB pixel values instead, this performs worse, due to the increased dimensionally of the 3 channels but without the increased complementary knowledge that combining the occupancy and semantic label channels provides. With regards the IMU, the *Raw IMU LSTM*, uses the raw global orientation of the IMU units without an kinematic solve, an LSTM model is trained on the raw IMU input and this performs badly with nearly double the error of the Solved-IMU. Part of the reason for this higher error is likely to be due to sensor drift within the IMU being unable to be modelled correctly by the LSTM. However, through constraining the noisy IMU unit responses with inverse kinematics, we are able to negate the IMU sensor drift to some degree. By then fusing the **SolvedIMU** and two channel PVH, the error is further reduced. This is likely to be due to

the complementary nature of the two data sources. Also, we show the result of just averaging the two data streams as the fusion method, this produces a high error, as expected as it is unable to learn anything about how the two data stream interact. It is possible to examine the per frame error for a sequence for subject 2 and sequence Acting3, in Fig. 10. Looking at the framewise errors, it shows that the two modes of data, the 3D PVH and SolvedIMU have lower errors at times, however through the use of the fusion layer, the overall error is lower than both. At around frame 1250, the Solved IMU increases in error due to a failure, however, the overall error rate of our proposed approach is relativity unchanged. While at frame 2500, the IMU is out performing the 3D PVH allowing the fused result to maintain a low error. However, at frame 4000 both modes fail, to cause higher errors in both data modes and the fused results, qualitative results of these three frames are shown in Fig. 11. For frame 4000 the higher errors can be seen to be caused by the arms not being extended correctly. The differences between the inferred poses can be quite small, indicating the contribution of all components of the approach. Although it's important to notice that the errors in the Solved-IMU pose for frames 2800 and 4000 aren't introduced to the final fused results. Run-time performance is 25 fps, including PVH generation. The ability of the approach to generalise between datasets is an interesting topic, therefore we compared applying a model trained on the TotalCapture dataset to the Human 3.6M dataset. We used the trained TotalCapture model from Table 6, *3D Matte CPM PVH-LSTM* i.e. the input to the fusion layer (as we cant use a model that takes in IMU data on the Human3.6M dataset). Given the different number of cameras and far poorer resultant PVHs formed by human3.6M, we fine-tune the TotalCapture trained model on the human3.6M using unfixed weights with a single epoch of the Human3.6M training data (normally the model is trained with 100 epochs where an epoch is a complete pass of the training data). The new fine-tuned model was then shown all the test sequences from Human3.6M and achieved an average joint error of 75.3 mm. This is similar to the performance of our approach with exclusive training on Human3.6M of 71.9 mm as shown in Table 2. This indicates that the learnt model is similar, although a small amount of adaptation is required between the datasets due in this case to the poor PVH generalisation for the Human3.6M dataset, later in Sect. 5 we will shown results on the TotalCaptureOutdoor without any generalisation.

## 4.5 In Depth Analysis

In this section, we explore and analyse some of the parameters in the approach. We investigate the effect of the number of cameras used, the amount of training data, the number of previous frames used for the temporal consistency and the effect the size of the voxels in the PVH volume has on the overall performance.

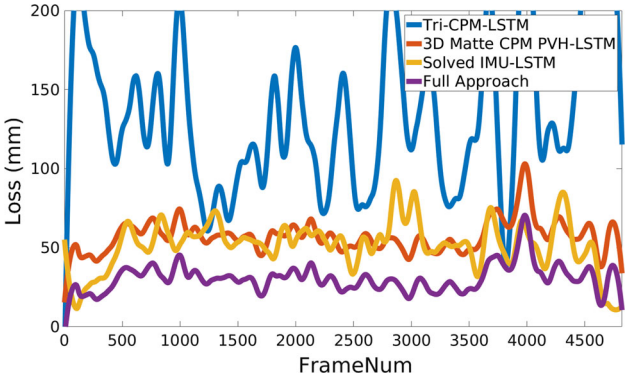



**Fig. 10** Per frame accuracy of our proposed approach on sequence A3 Subject2

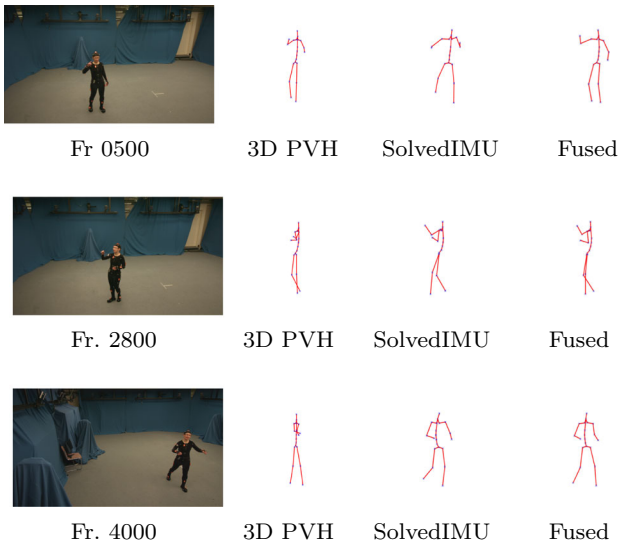
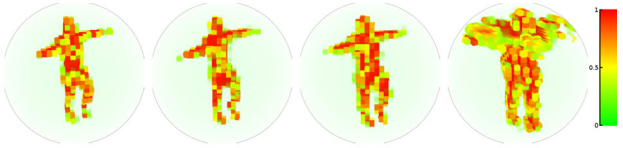



**Fig. 11** Visual comparison of poses resolved at different pipeline stages. TotalCapture: Acting3, Subject 2

### 4.5.1 Number of Cameras Used

Within the TotalCapture dataset there are 8 cameras, the greater the number of cameras, the more visually realistic the PVH is, for this work however it is possible to remove a large number of these with little or no impact on performance. The 3D PVH is constructed from the intersection of the foreground mattes and the intersection of the semantic 2D joint heat maps. With a greater number of cameras a more realistic PVH can be constructed, as can be seen comparing Fig. 12a, b, c which show the foreground matte based PVH with 8, 6, and 4 cameras respectively. While Fig. 12d shows the PVH for the Human3.6M dataset, the reason visually the Human3.6M PVH is worse in Fig. 12a–c, is probably due to the cameras being closer to the ground and also noisier fore-



**(a)** TC 8 cams **(b)** TC 6 cams **(c)** TC 4 cams **(d)** H3.6M, 4 cams

**Fig. 12** Effect on varying camera count on qualitative PVH appearance, for TotalCapture dataset (**a**–**c**) and Human3.6M (**d**)

**Table 7** Relative accuracy change (mm/joint) when varying the number of cameras

| Num | Seen(S1,2,3) | | | Unseen(S4,5) | | |
|---|---|---|---|---|---|---|
| Cams | W2 (%) | FS3 (%) | A3 (%) | W2 (%) | FS3 (%) | A3 (%) |
| 4 | 93.8 | 90.8 | 95.3 | 91.6 | 89.5 | 93.5 |
| 6 | 94.3 | 99.3 | 97.4 | 96.0 | 98.2 | 98.1 |
| 8 | 100 | 100 | 100 | 100 | 100 | 100 |

ground mattes being used, however performance isn't greatly affected. It can be seen that the PVH is visually less realistic with fewer cameras, however as shown in Table 7, which shows the relative performance for the whole fusion system with 4,6, and 8 cameras used to construct the 2 channel visual PVH, the performance is relatively unaffected despite halving the number of cameras used.

#### 4.5.2 Training Data Size

Generally for training neural networks a large amount of varied data is required, and the more data the higher the performance, especially as we use 3D convnets, which have an additional dimension and therefore additional weights to learn. We are able to investigate how the amount of training data affects the performance. The test sequences were kept consistent throughout as before, and 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. Table 8 suggests that the performance is relatively unaffected by the lower amounts of training data. This can be in part due to the use of our range of motions sequences within the training set. The approach can train with a sparse set of data and doesn't over-fit even if only 20% of the training data is used.

#### 4.5.3 Temporal Frame Length

Within the LSTM layers, there are memory cells that *remember* the previous $f$ data instances in time to provide temporal consistency. For this work $f = 5$, which is a compromise between little or no temporal memory and too long, which would fail to generalise to the test data after training. Figure 13 shows the how the performance on the regular train and test set varies for an increasing number of previous frames used. It can be seen that initially, the error is higher when little or no previous frame information is incorporated, it then increases and slows after a minimal around 5–6 frames. This is to be expected as the approach is starting to overfit to the training data and can't generalise to work well on the unseen test sequences.

**Table 8** Evaluating impact of accuracy (relative change in per joint mm error) as training data volume increases

| % Train | Seen(S1,2,3) | | | Unseen(S4,5) | | |
|---|---|---|---|---|---|---|
| Data | W2 (%) | FS3 (%) | A3 (%) | W2 (%) | FS3 (%) | A3 (%) |
| 20 | 96 | 89 | 86 | 93 | 85 | 84 |
| 40 | 97 | 91 | 87 | 94 | 86 | 86 |
| 60 | 97 | 94 | 89 | 94 | 89 | 90 |
| 80 | 99 | 95 | 93 | 97 | 91 | 93 |
| 100 | 100 | 100 | 100 | 100 | 100 | 100 |

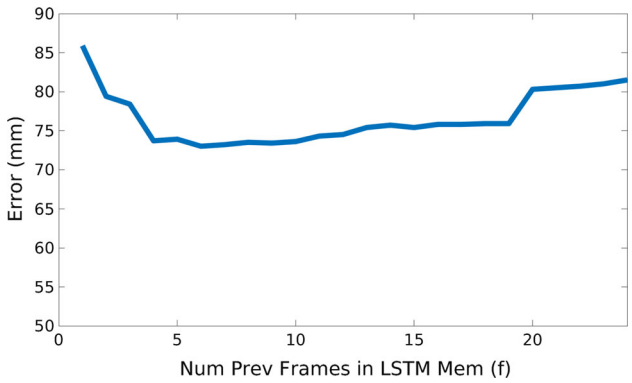



**Fig. 13** 3D Pose estimation error for increasing number of previous frames used by LSTM layers

**Table 9** Relative accuracy change (mm/joint) when varying the number of voxels in the PVH

| Voxels | Seen(S1,2,3) | | | Unseen(S4,5) | | |
|---|---|---|---|---|---|---|
| | W2 (%) | FS3 (%) | A3 (%) | W2 (%) | FS3 (%) | A3 (%) |
| $16 \times 16 \times 16$ | 85 | 84 | 82 | 86 | 87 | 82 |
| $30 \times 30 \times 30$ | 100 | 100 | 100 | 100 | 100 | 100 |
| $48 \times 48 \times 48$ | 97 | 98 | 97 | 99 | 98 | 99 |

#### 4.5.4 Voxel Resolution

Discrete voxels are used to carve up the 3D occupied volume to produce the probabilistic visual hull and then fed into the 3D convnet, with an initial resolution of $30 \times 30 \times 30$ voxels. Therefore for a $2 \times 2 \times 2$ m volume each voxel being

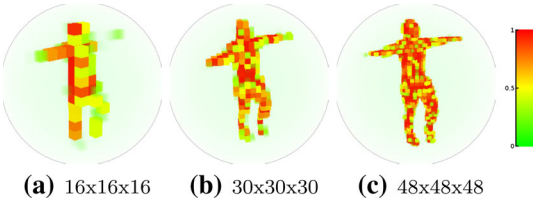



**Fig. 14** Affect of voxel sizes on qualitative PVH appearance

67 $mm^3$, which is the error measure, and therefore could be hypothesised that this is the minimum error noise threshold. We can investigate the effect of this coarse quantisation by increasing and reducing the number of voxels. Table 9 shows the relative effect in adjusting the voxel quantity, and visually in Fig. 14.

It can be seen that there is a slight reduction in performance with larger and smaller voxels 125 mm ($16 \times 16 \times 16$) and 41 mm ($48 \times 48 \times 48$) respectively however this is to be expected as with a larger voxels, the detail is reduced, and with the smaller voxels the parameter space is exponentially increased (110,000 elements for $48 \times 48 \times 48$ voxels compared to 27,000 for $30 \times 30 \times 30$), and therefore unable to effectively learn the additional weight parameter without the exponential increase in training data.

## 5 TotalCaptureOutdoor

To further demonstrate the generalisation of the approach, we test on a new challenging dataset used by (Malleson et al. 2017), This is a MVV and IMU dataset that was recorded outdoors in challenging uncontrolled conditions with a moving and changing background and varying illumination. 6 video cameras were placed in a 120 arc around the subject, with a large $8 \times 8$ m capture volume used. Examples of the camera viewpoints are shown in Fig. 15. For the TotalCaptureOutdoor sequences we uses the fully trained model (Dense Layer Fused approach from Table 6) from the TotalCapture dataset in Sect. 4.3 to predict the joints on the TotalCaptureOutdoor sequences. To indicate the generalisation ability of the approach a different camera setup (6 against the 8 on the indoor TotalCapture dataset) and the 13 Xsens IMUs were only placed in roughly similar locations to previous captures. Given the change in environment from a controlled studio to a unconstrained sunny and cloudy outdoor setting. We we able to achieve excellent qualitative performance on this more challenging dataset. There is no ground truth data is available for this dataset, however Fig. 16 shows a selection of pose estimations, for our full approach and the input image for subject 2. It can be seen that the resolved poses for our approach are able to accurately reflect the image despite all the training data being from the indoor TotalCapture dataset. Also the moving background from the tree is ignored correctly as noise by the occupancy based PVH. Finally Fig. 17, illustrates the resultant joint estimates from views taken in a 360° around the subject.

It shows that despite cameras only being present on one side, we are able to be accurately estimate full 360° joint locations.

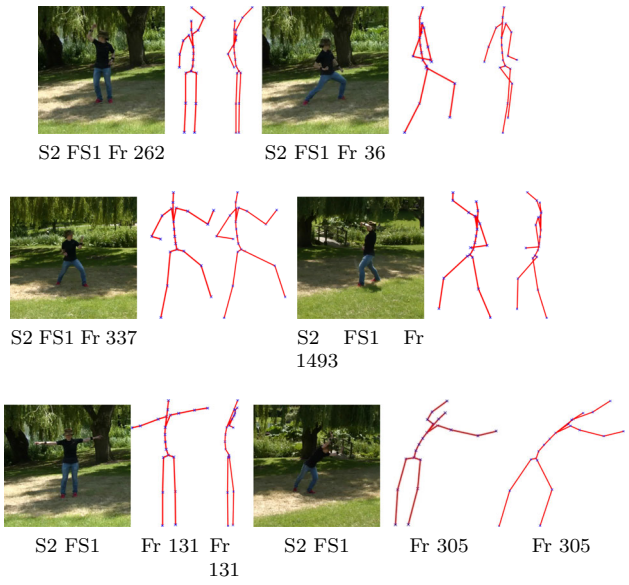



**Fig. 16** Visual comparison of poses resolved for the dataset TotalCaptureOutdoor for our proposed approach and the Tri-CPM

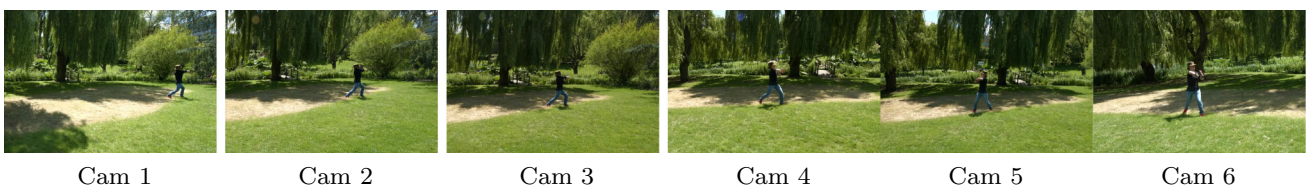



**Fig. 15** The cameras viewpoints of the TotalCaptureOutdoor dataset (Malleson et al. 2017)


**Fig. 17** 360° views of a frame from TotalCapture Outdoor

## 6 Conclusion

We have presented a novel approach for marker-less performance capture, that fuses MVV and IMU data to provide high accuracy human pose estimation in 3D. The MVV is used to produce semantic joint estimations and foreground occupancy, with a temporal model provided by LSTM layers to produce state of the art performance on the Human3.6M dataset, with a mean per joint error of 71.9 mm. Through the fusion of a forward kinematic solve from IMUs, this error can be further reduced by 10 mm beyond the state of the art. Currently the limitations of the approach are often due to poor foreground mattes that can cause the PVH to fail to accurately describe the subjects volume. Similarly, in challenging poses the 2D pose estimation can fail in a number of camera views, resulting in a poor input PVH input. However we have shown excellent qualitative results on three datasets including on a challenging outdoor dataset and are able to release 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 the donation of GPU hardware by Nvidia.



## References

Agarwal, A., & Triggs, B. (2004). 3D human pose from silhouettes by relevance vector regression. In *Proceedings of CVPR*.

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 961–971).

Andrews, S., Komura, T., Sigal, L., & Mitchell, K. (2016). Real-time physics-based motion capture with sparse sensors. In *CVMP*.

Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3686–3693).

Andriluka, M., Roth, S., & Schiele, B. (2009). Pictoral structures revisited: People detection and articulated pose estimation. In *Proceedings computer vision and pattern recognition*.

Baak, A., Helten, T., Müller, M., Pons-Moll, G., Rosenhahn, B., & Seidel, H.P. (2010). Analyzing and evaluating markerless motion tracking using inertial sensors. In *European conference on computer vision* (pp. 139–152). Springer, Berlin.

Cao, Z., Simon, T., Wei, S.E., & Sheikh, Y. (2016). Realtime multi-person 2D pose estimation using part affinity fields. In *ECCV'16*.

Cao, Z., Simon, T., Wei, S.E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*.

Chen, C.H., & Ramanan, D. (2017). 3D human pose estimation = 2D pose estimation + matching. In *CVPR*.

Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In arXiv preprint arXiv:1412.3555.

Dauphin, Y., de Vries, H., & Bengio, Y. (2015). Equilibrated adaptive learning rates for non-convex optimization. In *Advances in neural information processing systems* (pp. 1504–1512).

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR09*.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625–2634).

Elhayek, A., de Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., et al. (2015). Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3810–3818). IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7299005.

Grauman, K., Shakhnarovich, G., & Darrell, T. (2003). A bayesian approach to image-based visual hull reconstruction. In *Proceedings of CVPR*.

Graves, A. (2013). Generating sequences with recurrent neural networks. In arXiv preprint arXiv:1308.0850.

Graves, A., & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st international confernce on machine learning (ICML)*.

Helten, T., Muller, M., Seidel, H.P., & Theobalt, C. (2013). Real-time body tracking with one depth camera and inertial sensors. In *Proceedings of the IEEE international conference on computer vision* (pp. 1105–1112).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Huang, P., Tejera, M., Collomosse, J., & Hilton, A. (2015). Hybrid skeletal-surface motion graphs for character animation from 4D performance capture. *ACM Transactions on Graphics (ToG)*, *34*, 1–14.

Huang, Y., Bogo, F., Classner, C., Kanazawa, A., Gehler, P.V., Akhter, I., et al. (2017). Towards accurate markerless human shape and pose estimation over time. In *3DV*.

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). 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.

Jiang, H. (2009). Human pose estimation using consistent max-covering. In *International conference on computer vision*.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of NIPS*.

Lan, X., & Huttenlocher, D. (2005). Beyond trees: Common-factor model for 2d human pose recovery. *Proceedings of the IEEE International Conference on Computer Vision*, *1*, 470–477.

Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., & Gehler, P.V. (2017). Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*.

Li, S., Zhang, W., & Chan, A.B. (2015). Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision* (pp. 2848–2856).

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer, Berlin.

Liu, H., Wei, X., Chai, J., Ha, I., & Rhee, T. (2011). Realtime human motion control with a small number of inertial sensors. In *Symposium on interactive 3D graphics and games* (pp. 133–140). ACM.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, *34*(6), 248.

Malleson, C., Gilbert, A., Trumble, M., Collomosse, J., & Hilton, A. (2017). Real-time full-body motion capture from video and imus. In *3DV*.

Marcard, T.V., Pons-Moll, G. & Rosenhahn, B. (2016). *Multimodal motion capture dataset TNT15*. Technical Report. Hanover, Germany: Leibniz Univ. Hannover and Tübingen, Germany: Max Planck for Intelligent Systems.

Martinez, J., Hossain, R., Romero, J., & Little, J.J. (2017). A simple yet effective baseline for 3D human pose estimation. ICCV. arXiv:1705.03098.

Mude, L., Liang, L., Xiaodan, L., Keze, W., & Cheng, H. (2017). Recurrent 3D pose sequence machines. In *CVPR*.

Optitrack motive. http://www.optitrack.com. Accessed Dec 2017.

Park, D., & Ramanan, D. (2015). Articulated pose estimation with tiny synthetic videos. In *Proceedings of CHA-LEARN workshop on looking at people*.

Pavlakos, G., Zhou, X., Derpanis, K.G., & Daniilidis, K. (2017a). Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*.

Pavlakos, G., Zhou, X., Derpanis, K.G., & Daniilidis, K. (2017b). Harvesting multiple views for marker-less 3D human pose annotations. In *CVPR*.

Perception neuron. http://www.neuronmocap.com. Accessed Dec 2017.

Pons-Moll, G., Baak, A., Gall, J., Leal-Taixe, L., Mueller, M., Seidel, H.P., et al. (2011). Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *2011 IEEE international conference on computer vision (ICCV)* (pp. 1243–1250). IEEE.

Pons-Moll, G., Baak, A., Helten, T., Müller, M., Seidel, H.P., & Rosenhahn, B. (2010). Multisensor-fusion for 3D full-body human motion capture. In *2010 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 663–670). IEEE.

Ren, R., & Collomosse, J. (2012). Visual sentences for pose retrieval over low-resolution cross-media dance collections. *IEEE Transactions on Multimedia*, *14*, 1652–1661.

Ren, X., Berg, E., & Malik, J. (2005). Recovering human body configurations using pairwise constraints between parts. *Proceedings of the IEEE International Conference on Computer Vision*, *1*, 824–831.

Rhodin, H., Robertini, N., Casas, D., Richardt, C., Seidel, H.P., & Theobalt, C. (2016). General automatic human shape and motion capture using volumetric contour cues. In *European conference on computer vision* (pp. 509–526). Springer, Berlin.

Roetenberg, D., Luinge, H., & Slycke, P. (2009). Xsens mvn: Full 6D of human motion tracking using miniature inertial sensors. http://www.xsens.com.

Rogez, G., & Schmid, C. (2016). Mocap-guided data augmentation for 3D pose estimation in the wild. In *Advances in neural information processing systems* (pp. 3108–3116).

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*.

Sanzari M., Ntouskos, V., & Pirri, F. (2016). Bayesian image based 3D pose estimation. In *European conference on computer vision* (pp. 566–582). Springer, Berlin.

Schwarz, L.A., Mateus, D., & Navab, N. (2009). Discriminative human full-body pose estimation from wearable inertial sensor data. In *3D physiological human workshop* (pp. 159–172). Springer, Berlin.

Sigal, L., Balan, A. O., & Black, M. J. (2010). Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, *87*, 4–27.

Slyper, R., & Hodgins, J.K. (2008). Action capture with accelerometers. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics symposium on computer animation* (pp. 193–199). Eurographics Association.

Srinivasan, P., & Shi, J. (2007). Bottom-up recognition and parsing of the human body. In *Proceedings computer vision and pattern recognition* (pp. 1–8).

Tan, J., Budvytis, I., & Cipolla, R. (2017). Indirect deep structured learning for 3D human body shape and pose prediction. In *BMVC*.

Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., & Fua, P. (2016). Structured prediction of 3D human pose with deep neural networks. In *BMVC*. arXiv preprint arXiv:1605.05180.

Tekin, B., Márquez-Neila, P., Salzmann, M., & Fua, P. (2016). Fusing 2D uncertainty and 3D cues for monocular body pose estimation. arXiv preprint arXiv:1611.05708.

Tome, D., Russell, C., & Agapito, L. (2017). Lifting from the deep: Convolutional 3D pose estimation from a single image. arXiv preprint arXiv:1701.00295.

Toshev, A., & Szegedy, C. (2014). Deep pose: Human pose estimation via deep neural networks. In *Proceedings of CVPR*.

Trumble, M., Gilbert, A., Hilton, A., & Collomosse, J. (2018). Deep autoencoder for combined human pose estimation and body model upscaling. In *European conference on computer vision (ECCV'18)*.

Trumble, M., Gilbert, A., Hilton, A., & John, C. (2016). 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)*.

Trumble, M., Gilbert, A., Malleson, C., Hilton, A. & Collomosse, J. (2017). Total capture: 3D human pose estimation fusing video and inertial sensors. In *BMVC17*.

Vicon blade. http://www.vicon.com. Accessed Dec 2017.

von Marcard, T., Rosenhahn, B., Black, M., & Pons-Moll, G. (2017). 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).

von Marcard, T., Pons-Moll, G., & Rosenhahn, B. (2016). Human pose estimation from video and imus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(8), 1533–1547.

Wei, S.E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In *CVPR*.

Yub, H.J., Suh, Y., Moon, G., & Mu Lee, K. (2016). Sequential approach to 3D human pose estimation: Seperation of localization and identification of body joints. In *Proceedings of European conference on computer vision (ECCV16)*.

Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., & Daniilidis, K. (2016). Sparseness meets deepness: 3D human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4966–4975).