

Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera

Timo von Marcard¹, Roberto Henschel¹, Michael J. Black², Bodo Rosenhahn¹,
and Gerard Pons-Moll³

¹ Leibniz Universität Hannover, Germany

² MPI for Intelligent Systems, Tübingen, Germany

³ MPI for Informatics, Saarland Informatics Campus, Germany

{marcard,henschel,rosenhahn}@tnt.uni-hannover.de, black@tue.mpg.de,
gpons@mpi-inf.mpg.de

Abstract. In this work, we propose a method that combines a single hand-held camera and a set of Inertial Measurement Units (IMUs) attached at the body limbs to estimate accurate 3D poses in the wild. This poses many new challenges: the moving camera, heading drift, cluttered background, occlusions and many people visible in the video. We associate 2D pose detections in each image to the corresponding IMU-equipped persons by solving a novel graph based optimization problem that forces 3D to 2D coherency within a frame and across long range frames. Given associations, we jointly optimize the pose of a statistical body model, the camera pose and heading drift using a continuous optimization framework. We validated our method on the TotalCapture dataset, which provides video and IMU synchronized with ground truth. We obtain an accuracy of $26mm$, which makes it accurate enough to serve as a benchmark for image-based 3D pose estimation in the wild. Using our method, we recorded *3D Poses in the Wild (3DPW)*, a new dataset consisting of more than 51,000 frames with accurate 3D pose in challenging sequences, including walking in the city, going up-stairs, having coffee or taking the bus. We make the reconstructed 3D poses, video, IMU and 3D models available for research purposes at <http://virtualhumans.mpi-inf.mpg.de/3DPW>.

Keywords: Human Pose, Video, IMUs, Sensor Fusion, 2D to 3D, People Tracking, 3D Pose Dataset

1 Introduction

This paper addresses two inter-related goals. First, we propose a method capable of accurately reconstructing 3D human pose in outdoor scenes, with multiple people interacting with the environment, see Fig. 1. Our method combines data coming from IMUs (attached at the person’s limbs) with video obtained from a hand-held phone camera. This allows us to achieve the second goal, which is collecting the first dataset with accurate 3D reconstructions in the wild. Since

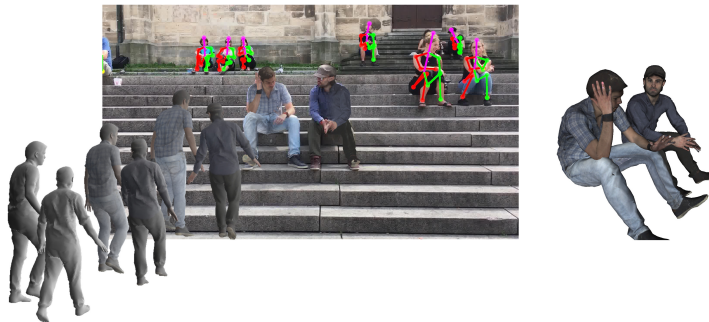


Fig. 1. We propose Video Inertial Poser (VIP), which enables accurate 3D human pose capture in natural environments. VIP combines video obtained from a hand-held smartphone camera with data coming from body-worn inertial measurement units (IMUs). With VIP we collected *3D Poses in the Wild*, a new dataset of accurate 3D human poses in natural video, containing variations in person identity, activity and clothing.

our system works with a moving camera, we can record people in their everyday environments, for example, walking in the city, having coffee or taking the bus.

3D human pose estimation from un-constrained single images and videos has been a longstanding goal in computer vision. Recently, there has been a significant progress, particularly in 2D human pose estimation [23, 4]. This progress has been possible thanks to the availability of large training datasets and benchmarks to compare research methods. While obtaining manual 2D pose annotations in the wild is fairly easy, collecting 3D pose annotations manually is almost impossible. This is probably the main reason there exist very limited datasets with accurate 3D pose in the wild. Datasets such as HumanEva [32] and H3.6M [8] have facilitated progress in the field by providing ground truth 3D poses obtained using a marker-based motion capture system synchronized with video. These datasets, while useful and necessary, are restricted to indoor scenarios with static backgrounds, little variation in clothing and no environmental occlusions. As a result, evaluations of 3D human pose estimation methods in challenging images have been made mainly qualitatively, so far. There exist several options to record humans in outdoor scenes, none of which is satisfactory. Marker-based capture outdoors is limited. Depth sensors like Kinect do not work under strong illumination and can only capture objects near the camera. Using multiple cameras as in [21] requires time consuming set-up and calibration. Most importantly, the fixed recording volume severely limits the kind of activities that can be captured.

IMU-based systems hold promise because they are not bound to a fixed space since they are worn by the person. In practice, however, accuracy is limited by a number of factors. Inaccuracies in the initial pose introduce sensor-to-bone misalignments. In addition, during continuous operation, IMUs suffer from heading drift, see Fig. 2. This means, that after some time, each IMU does

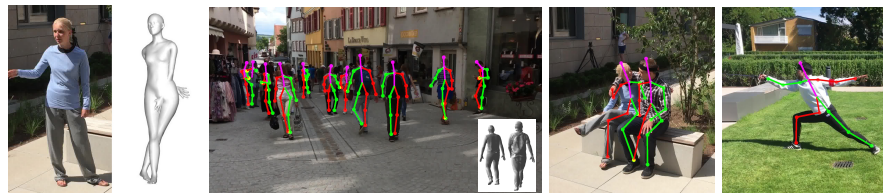


Fig. 2. For accurate motion capture in the wild we have to solve several challenges: IMU heading drift has accumulated after a longer recording session and the obtained 3D pose is completely off (left image pair). In order to estimate the heading drift, we combine IMU data and 2D poses detected in the camera view. This requires the association of 2D poses to the person wearing IMUs, which is difficult when several people are in the scene (middle image). Also, 2D pose candidates might be inaccurate and should be automatically rejected during the assignment step (right image pair).

not measure relative to the *same* world coordinate frame. Rather, each sensor provides readings relative to *independent* coordinate frames that slowly drift away from the world frame. Furthermore, global position can not be accurately obtained due to positional drift. Moreover, IMU systems do not provide 3D pose synchronized and aligned with image data.

Therefore, we propose a new method, called *Video Inertial Poser (VIP)*, that jointly estimates the pose of people in the scene by using 6 to 17 IMUs attached at the body limbs and a single hand-held moving phone camera. Using IMUs makes the task less ambiguous but many challenges remain. First, the persons need to be detected in the video and associated with the IMU data, see Fig. 2. Second, IMUs are inaccurate due to heading drift. Third, the estimated 3D poses need to align with the images of the moving camera. Furthermore, the scenes we tackle in this work include complete occlusions, multiple people, tracked persons falling out of the camera view and camera motion. To address these difficulties, we define a novel graph-based association method, and a continuous pose optimization scheme that integrates the measurements from all frames in the sequence. To deal with noise and incomplete data, we exploit SMPL [14], which incorporates anthropometric and kinematic constraints.

Specifically, our approach has three steps: initialization, association and data fusion. During initialization, we compute initial 3D poses by fitting SMPL to the IMU orientations. The association step automatically associates the 3D poses with 2D person detections for the full sequence by solving a single binary quadratic optimization problem. Given those associations, in the data fusion step, we define an objective function and jointly optimize for the 3D poses of the full sequence, the per-sensor heading errors, the camera pose and translation. Specifically, the objective is minimized when (i) the model orientation and acceleration is close to the IMU readings and (ii) the projected 3D joints of SMPL are close to 2D CNN detections [4] in the image. To further improve results, we repeat association and joint optimization once.

With VIP we can accurately estimate 3D human poses in challenging natural scenes. To validate the accuracy of VIP, we use the recently released 3D dataset Total Capture [39] because it provides video synchronized with IMU data. VIP obtains an average 3D pose error of $26mm$, which makes it accurate enough to benchmark methods that tackle in-the-wild data. Using VIP we created *3D Poses in the Wild (3DPW)*: a dataset consisting of hand-held video with ground-truth 3D human pose and shape in natural videos.

We make 3DPW publicly available for research purposes, including 60 video sequences (51,000 frames or 1700 seconds of video captured with a phone at 30Hz), IMU data, 3D scans and 3D people models with 18 clothing variations, and the accurate 3D pose reconstruction results of VIP in all sequences. We anticipate that the dataset will stimulate novel research by providing a platform to quantitatively evaluate and compare methods for 3D human pose estimation.

2 Related Work

Pose Estimation using IMUs. There exist commercial solutions for MoCap with IMUs. The approach of [30] integrates 17 IMUs in a Kalman Filter to estimate pose. The seminal work of [41] uses a custom made suit to capture pose in everyday surroundings. These approaches require many sensors and do not align the reconstructions with video; therefore they suffer from drift. The approach of [42] fits the SMPL body model to 5-6 IMUs over a full sequence, obtaining realistic results. The method, however, is applied to only 1 person at a time and the motion is not aligned with video. To compensate for drift, 4-8 cameras and 5 IMUs are combined in [17, 25]. Using particle-based optimization, in [24] they use 4 cameras and IMUs to sample from a manifold of constrained poses. Other works combine depth data with IMUs [6, 47]. In [39] a CNN-based approach fuses information from 8 camera views and IMU data to regress pose. Since these approaches also use multiple static cameras, recordings are restricted to a fixed recording volume. A recent approach [16] also combines IMUs and 2D poses detected in one or two cameras but expects only a single person visible in the cameras and does not account for heading drift.

3D Pose Datasets. The most commonly used datasets for 3D human pose evaluation are HumanEva [32] and H3.6M [8], which provide synchronized video with MoCap. These datasets however are limited to indoor scenes, static backgrounds and limited clothing and activity variation. Recently, a dataset of single people, including outdoor scenes, has been introduced [19]. The approach uses commercial marker-less motion capture from multiple cameras (the accuracy of the marker-less MoCap software used is not reported). The sequences show variation in clothing, but again, since it uses a multi-camera setup, the activities are restricted to a fixed recording volume. Another recent dataset is TotalCapture [39], which features synchronized video, marker-based ground-truth poses and IMUs. In order to collect 3D poses in the wild, in [11] they ask users to pick “acceptable” results obtained using an automatic 3D pose estimation method. The problem is that it is difficult to judge a correct pose visually and it is not

clear how accurate automatic methods are with in-the-wild images. We do not see our proposed dataset as an alternative to existing datasets; rather 3DPW complements existing ones with new, more challenging, sequences.

3D Human Pose. Several works lift 2D detections to 3D using learning or geometric reasoning [18, 29, 35, 9, 26, 49, 33, 48, 44, 34, 13, 43, 45]. These works aim at recovering the missing depth dimension in single-person images, whereas we focus on directly associating the 3D to the 2D poses in cluttered scenes. For multiple people, the work [1] infers the 3D poses using a tracking formulation that is based on short tracklets of 2D body parts. Recently 2D annotations have been leveraged to train networks for the task of 3D pose estimation [21, 28, 36, 38, 50]. Such works typically predict only stick figures or bone skeletons. Some approaches directly predict the parameters of a body model (SMPL) from a single image using 2D supervision [10, 22, 40]. Closer to our method are the works [2, 11], which fit SMPL [14] to 2D detections. The optimization problem we solve, even though it integrates more sensors, is much more involved. Very few approaches tackle multiple-person 3D pose estimation [31, 20]. 3DPW allows a quantitative evaluation of all these approaches for in-the-wild images.

3 Background

SMPL Body Model. We utilize the Skinned Multi-Person Linear (SMPL) body model [14], which is a statistical body model parameterized by identity-dependent shape parameters and the skeleton pose. We optimize the shape parameters to the person to be tracked by fitting SMPL to a 3D scan. Holding shape fixed, our aim is to recover the pose $\theta \in \mathbb{R}^{75}$, consisting of 6 parameters for global translation and rotation, and 23 relative rotations represented by axis-angle for each joint. We use the standard forward kinematics to map pose θ to the rigid transformation $\mathbf{G}^{GB}(\theta) : \mathbb{R}^{75} \rightarrow SE(3)$ of bone B . The bone transformation comprises the rotation and translation $\mathbf{G}^{GB} = \{\mathbf{R}^{GB}, \mathbf{t}^{GB}\}$ to map from the local bone coordinate frame F^B to the global SMPL frame F^G .

Coordinate Frames. Ultimately, we want to find the pose θ that produces bone orientations close to the IMU readings. IMUs measure the orientation of the local coordinate frame F^S (of the sensor box) relative to a global coordinate frame F^I . However, this frame F^I is different from the coordinate frame F^G of SMPL, see Fig. 5. The offset $\mathbf{G}^{GI} : F^I \rightarrow F^G$ between coordinate frames is typically assumed constant, and is calibrated at the beginning of a recording session – but that is not enough. We also need to know the offset \mathbf{R}^{BS} from the sensor to the SMPL bone where it is placed. The SMPL bone orientation $\mathbf{R}^{GB}(\theta_0)$ can be obtained in the first frame assuming a known pose θ_0 . Using this bone orientation $\mathbf{R}^{GB}(\theta_0)$ and the raw IMU reading $\mathbf{R}^{IS}(0)$ in the first frame, we can trivially find the offset relating them as

$$\mathbf{R}^{BS} = (\mathbf{R}^{GB}(\theta_0))^{-1} \mathbf{R}^{GI} \mathbf{R}^{IS}(0) \quad (1)$$

where the raw IMU reading $\mathbf{R}^{IS}(0)$ needs to be mapped to the SMPL frame first using \mathbf{R}^{GI} . We assume that sensors do not move relative to the bones, and

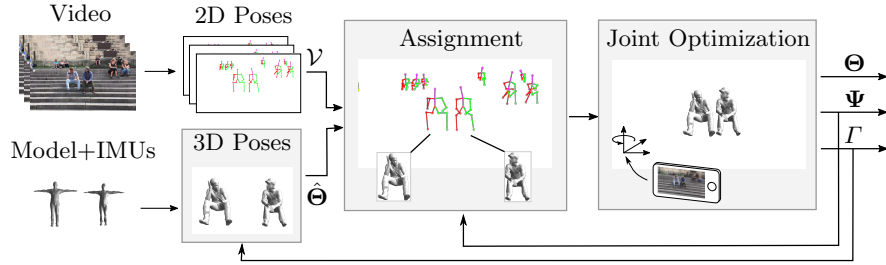


Fig. 3. Method overview: By fitting the SMPL body model to the measured IMU orientations we obtain initial 3D poses Θ . Given all 2D poses \mathcal{V} detected in the images we search for a globally consistent assignment of 2D to 3D poses. We jointly optimize camera poses Ψ , heading angles Γ and 3D poses Θ with respect to associated IMU and image data. In a second iteration we feed back camera poses and heading angles which provides additional information further improving the assignment and tracking results.

hence compute \mathbf{R}^{BS} from the initial pose θ_0 and IMU orientations in the first frame.

Heading Drift. Unfortunately, the orientation measurements of the IMUs are deteriorated by magnetic disturbances, which introduce a time-varying rotational offset to \mathbf{G}^{GI} , also commonly known as heading error or heading drift. This drift ($\mathbf{G}^{I'I} : F^I \rightarrow F^{I'}$) shifts the original global inertial frame F^I to a disturbed inertial frame $F^{I'}$. What is even worse, the drift is different for every sensor. While most previous works ignore heading drift or treat it as noise, we model it explicitly and recover it as part of the optimization. Concretely, we model it as a one-parameter rotation $\mathbf{R}(\gamma) \in SO(3)$ about the vertical axis, where γ is the rotation angle. The collection of all angles, one per IMU sensor, is denoted as Γ . Since the heading error commonly varies slowly, we assume it is constant during a single tracking sequence. Recovering heading orientation was crucial in order to be able to perform long recordings without time-consuming re-calibration.

4 Video Inertial Poser (VIP)

In order to perform accurate 3D human motion capture with hand-held video and IMUs we perform three subsequent steps: initialization, pose candidate association and video-inertial fusion. Fig. 3 provides an overview of the pipeline and we describe each step in more detail in the following.

4.1 Initialization

We obtain initial 3D poses by fitting the SMPL bone orientations to the measured IMU orientations. For an IMU, the *measured bone orientation* $\hat{\mathbf{R}}^{GB}$ is given by

$$\hat{\mathbf{R}}^{GB} = \mathbf{R}^{GI'} \mathbf{R}^{I'I}(\gamma) \mathbf{R}^{IS} (\mathbf{R}^{BS})^{-1}, \quad (2)$$

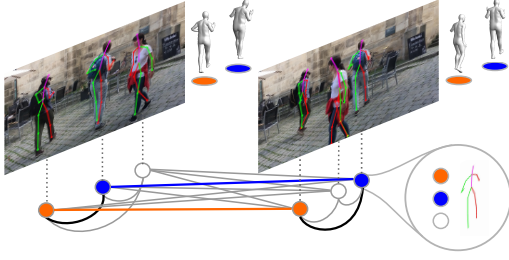


Fig. 4. Every 2D pose represents a node in the graph which can be assigned to a 3D pose corresponding to person 1 or 2 (represented by colors orange and blue). The graph has intra-frame edges (shown in black) activated if two nodes are assigned in a single frame and inter-frame edges (shown in blue and orange) activated for the same person across multiple frames.

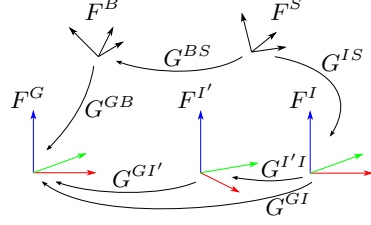


Fig. 5. Coordinate frames: Global tracking frame F^G , global inertial frame F^I , shifted inertial frame $F^{I'}$, bone coordinate frame F^B and IMU sensor coordinate frame F^S .

where \mathbf{R}^{BS} represents the constant bone to sensor offset (Eq. (1)), and the concatenation of $\mathbf{R}^{GI'}$, $\mathbf{R}^{I'I}$ and \mathbf{R}^{IS} describes the rotational map from sensor to global frame, see Fig. 5. We define the rotational discrepancy between actual bone orientation $\mathbf{R}^{GB}(\theta)$ and measured bone orientation $\hat{\mathbf{R}}^{GB}$ as

$$\mathbf{e}^{\text{rot}}(\theta) = \log \left(\mathbf{R}^{GB}(\theta) \left(\hat{\mathbf{R}}^{GB} \right)^{-1} \right)^\vee, \quad (3)$$

where the log-operation recovers the skew-symmetric matrix from the relative rotation between $\mathbf{R}^{GB}(\theta)$ and $\hat{\mathbf{R}}^{GB}$, and the \vee -operator extracts the corresponding axis-angle parameters. We find the 3D initial poses at frame t that minimize the sum of discrepancies for all IMUs

$$\theta_t^* = \arg \min_{\theta} \frac{1}{N_s} \sum_{s=1}^{N_s} \|\mathbf{e}_{s,t}^{\text{rot}}(\theta_t)\|^2 + w_{\text{prior}} E_{\text{prior}}(\theta_t), \quad (4)$$

where $E_{\text{prior}}(\theta)$ is a pose prior weighted by w_{prior} . $E_{\text{prior}}(\theta)$ is chosen as defined in [42], enforcing θ to remain close to a multivariate Gaussian distribution of model poses and to stay within joint limits. During the first iteration, we have no information about the heading angles γ . To initialize them, we use the IMU placement as a proxy to know how local sensor axes are aligned with respect to the body. This gives us a rough estimate of the sensor to bone offset $\hat{\mathbf{R}}^{BS}$, which we use to compute initial heading angles by solving Eq. (1) for γ .

In the following, we will refer to this tracking approach simply as the inertial tracker (IT), which outputs initial 3D pose candidates $\theta_{t,l}^*$ for every tracked person l . Such initial 3D poses need to be associated with 2D detections in the video in order to effectively fuse the data – this poses a challenging assignment problem.

4.2 Pose Candidate Assignment

Using the CNN method of Cao *et al.* [4], we obtain 2D pose detections v , which comprise the image coordinates of $N_{\text{joints}} = 18$ landmarks along with corresponding confidence scores. In order to associate each 2D pose v to a 3D pose candidate, we create an undirected weighted graph $G = (\mathcal{V}, \mathcal{E}, c)$, with \mathcal{V} comprising all detected 2D poses in a recording sequence. An assignment hypothesis, denoted as $\mathcal{H}(l, v) = (\theta_t^l, v)$, links the 3D pose θ_t^l of person $l \in \{1, \dots, P\}$ to the 2D pose $v \in \mathcal{V}$ in the same frame t . We introduce indicator variables x_v^l , which take value 1 if hypothesis $\mathcal{H}(l, v)$ is selected, and 0 otherwise. The basic idea is to assign costs to each hypothesis, and select the assignments for the sequence that minimize the total costs. We cast the selection problem as a graph labeling problem by minimizing the following objective

$$\arg \min_{x \in \mathcal{F} \cap \{0,1\}^{|\mathcal{V}|P}} \sum_{\substack{v \in \mathcal{V} \\ l \in \{1, \dots, P\}}} c_v^l x_v^l + \sum_{\substack{\{v, v'\} \in \mathcal{E} \\ l, l' \in \{1, \dots, P\}}} c_{v, v'}^{l, l'} x_v^l x_{v'}^{l'}, \quad (5)$$

where the feasibility set \mathcal{F} is subject to:

$$(a) \sum_{l=1}^P x_v^l \leq 1 \quad \forall v \in \mathcal{V}; \quad (b) \sum_{v \in \mathcal{V}_t} x_v^l \leq 1 \quad \forall t, \forall l \in \{1, \dots, P\}. \quad (6)$$

The edge set \mathcal{E} contains all pairs of 2D poses $\{v, v'\}$ that are considered for the assignment decision. Eq. (6)(a) ensures that a 2D pose v is assigned to at most 1 person, and Eq. (6)(b) ensures that each person is assigned to at most one of the 2D pose detections $v \in \mathcal{V}_t \subset \mathcal{V}$ in frame t . The objective in (5) consists of unary costs c_v^l , measuring 2D to 3D consistency, and pairwise costs $c_{v, v'}^{l, l'}$ measuring consistency across different hypothesis. Our formulation automatically outputs a globally consistent assignment and does not require manual initialization.

Next we describe the unaries and pairwise potentials – specifically, we introduce consistency features which are mapped to the costs $c_v^l, c_{v, v'}^{l, l'}$ of the objective in (5) via logistic regression. Details about the training process are described in Section 5.1. Fig. 4 visualizes the graph for two example frames and also illustrates the corresponding labeling solution.

Unary Costs. To measure 2D to 3D consistency of a hypothesis $\mathcal{H} := \mathcal{H}(l, v)$, we obtain a *hypothesis camera* $\mathbf{M}_{\mathcal{H}}$ by minimizing the re-projection error between 3D landmarks of θ_t^l and the 2D detected ones v . The per landmark re-projection error, denoted by $\mathbf{e}_{\text{img}, k}(\mathcal{H}, \mathbf{M}_{\mathcal{H}})$, is weighted by the confidence scores w_k . The consistency is then measured as the average of all weighted residuals $\mathbf{e}_{\text{img}, k}(\mathcal{H}, \mathbf{M}_{\mathcal{H}})$, denoted by $\mathbf{e}_{\text{img}}(\mathcal{H}, \mathbf{M}_{\mathcal{H}})$. This measure depends heavily on the distance to the camera. To balance it, we scale it by the average 3D joint distance to the camera center $\mathbf{e}_{\text{cam}}(\mathbf{M}_{\mathcal{H}})$ and obtain the feature:

$$f_{\text{un}}(\mathcal{H}) = \mathbf{e}_{\text{img}}(\mathcal{H}, \mathbf{M}_{\mathcal{H}}) \mathbf{e}_{\text{cam}}(\mathcal{H}, \mathbf{M}_{\mathcal{H}}). \quad (7)$$

Pairwise Costs. We define features to measure the consistency of two hypothesis $\mathcal{H} = (\theta_t^l, v)$ and $\mathcal{H}' = (\theta_{t'}^{l'}, v')$ in frames t and t' . In particular, two kinds of edges connect hypothesis: (a) *inter-frame*, and (b) *intra-frame*.

a) *Inter-frame:* Consider two hypothesis $\mathcal{H}, \mathcal{H}'$ corresponding to the *same person* and separated by fewer than 30 frames. Then, the respective root joint position $r(\theta_t^l)$ and orientation $\mathbf{R}(\theta_t^l)$ in camera hypothesis ($\mathbf{M}_{\mathcal{H}}$) coordinates should not vary too much. This variation depends on the temporal distance $|t - t'|$. Consequently, we introduce the following features

$$f_{\text{trans}}(\mathcal{H}, \mathcal{H}') = \|\mathbf{M}_{\mathcal{H}}r(\theta_t^l) - \mathbf{M}_{\mathcal{H}'}r(\theta_{t'}^{l'})\|^2, \quad (8)$$

$$f_{\text{ori}}(\mathcal{H}, \mathcal{H}') = \left\| \log \left((\mathbf{R}_{\mathcal{H}}\mathbf{R}(\theta_t^l))^{-1}(\mathbf{R}_{\mathcal{H}'}\mathbf{R}(\theta_{t'}^{l'})) \right) \right\|^2, \quad (9)$$

$$f_{\text{time}}(\mathcal{H}, \mathcal{H}') = \|t - t'\|^2, \quad (10)$$

where f_{trans} and f_{ori} measure root joint translation and orientation consistency, and f_{time} is a feature to accommodate for temporal distance. Here, $\mathbf{R}_{\mathcal{H}}$ is the rotational part of $\mathbf{M}_{\mathcal{H}}$, and f_{rot} computes the geodesic distance between $\mathbf{R}(\theta_t^l)$ and $\mathbf{R}(\theta_{t'}^{l'})$, similar to Eq. (3).

b) *Intra-frame:* Now consider two hypothesis $\mathcal{H}, \mathcal{H}'$ for *different persons* in the same frame. The resulting camera hypothesis centers should be consistent. To measure coherency, we compute a meta-camera hypothesis $\mathbf{M}_{\underline{\mathcal{H}}}$ by minimizing the re-projection error of both hypothesis at the same time. Then the feature

$$f_{\text{intra}}(\mathcal{H}, \underline{\mathcal{H}}) = \|\mathbf{c}(\theta_t^l, \mathbf{M}_{\mathcal{H}}) - \mathbf{c}(\theta_t^l, \mathbf{M}_{\underline{\mathcal{H}}})\|^2 \quad (11)$$

measures the camera $\mathbf{c}(\theta_t^l, \mathbf{M}_{\mathcal{H}})$ to meta-camera center $\mathbf{c}(\theta_t^l, \mathbf{M}_{\underline{\mathcal{H}}})$ difference. Accordingly, we also use the feature $f_{\text{intra}}(\mathcal{H}', \underline{\mathcal{H}})$ for intra-frame edges.

Graph Optimization. Although the presented graph labeling problem in (5) is NP-Hard, it can be solved efficiently in practice [7, 12]. We use the binary LP solver Gurobi [5] by applying it to the linearized formulation of (5), where we replace each product $x_v^l x_{v'}^{l'}$ by a binary auxiliary variable $y_{v,v'}^{l,l'}$ and add corresponding constraints such that $x_v^l x_{v'}^{l'} = y_{v,v'}^{l,l'}$ for all $v, v' \in \mathcal{V}$, for all $l, l' \in \{1, \dots, P\}$.

4.3 Video-Inertial Data Fusion

Once the assignment problem is solved we can utilize the associated 2D poses to jointly optimize model poses, camera poses and heading angles by minimizing the following energy:

$$E(\Theta, \Psi, \Gamma) = E_{\text{ori}}(\Theta, \Gamma) + w_{\text{acc}}E_{\text{acc}}(\Theta, \Gamma) + w_{\text{img}}E_{\text{img}}(\Theta, \Psi) + w_{\text{prior}}E_{\text{prior}}(\Theta), \quad (12)$$

where Θ is a vector containing the pose parameters for each actor and frame, Γ is the vector of IMU heading correction angles and Ψ contains the camera poses for each frame. $E_{\text{ori}}(\Theta, \Gamma)$, $E_{\text{acc}}(\Theta, \Gamma)$ and $E_{\text{img}}(\Theta, \Psi)$ are energy terms related to IMU orientations, IMU accelerations and image information, respectively. $E_{\text{prior}}(\Theta)$ is an energy term related to pose priors. Finally, every term is weighted by a corresponding weight w .

Orientation Term The orientation term simply extends Eq. (4) by considering all frames N_T of a sequence according to

$$E_{\text{ori}}(\Theta, \Gamma) = \frac{1}{N_T N_s} \sum_{t=1}^{N_T} \sum_{s=1}^{N_s} \|\mathbf{e}_{s,t}^{\text{rot}}(\theta_t, \gamma_s)\|^2. \quad (13)$$

This term also includes the camera IMU, where the camera rotation mapping from camera coordinate system F^C to the global coordinate frame F^G is given by the inverse rotational part of the camera pose M .

Acceleration Term The acceleration term enforces consistency of the measured IMU acceleration and the acceleration of the corresponding model vertex to which the IMU is attached. The IMU acceleration in world coordinates for sensor s at time t is given by

$$\mathbf{a}_{s,t}^G(\gamma) = \mathbf{R}^{GI'} \mathbf{R}^{I'I}(\gamma_s) \mathbf{R}^{IS} \mathbf{a}_{s,t}^S - \mathbf{g}^G, \quad (14)$$

where \mathbf{g}^G is gravity in global coordinates. The corresponding SMPL vertex acceleration $\hat{\mathbf{a}}(\theta_t)$ is approximated by finite differences. Finally, the acceleration term contains the quadratic norm of the deviation of measured and estimated acceleration for all N_S IMUs over all frames N_T :

$$E_{\text{acc}}(\Theta, \Gamma) = \frac{1}{N_T N_S} \sum_{t=1}^{N_T} \sum_{s=1}^{N_S} \|\hat{\mathbf{a}}_s(\theta_t) - \mathbf{a}_{s,t}(\gamma_s)\|^2. \quad (15)$$

This term also contains the measured acceleration of the camera IMU and the corresponding acceleration of the camera center in global coordinates.

Image Term The image term simply accumulates the re-projection error over all N_{joints} landmarks and all frames N_T according to

$$E_{\text{img}}(\Theta, \Psi) = \frac{1}{N_T N_{\text{coco}}} \sum_{t=1}^{N_T} \sum_{i=k}^{N_{\text{joints}}} w_k \|\mathbf{e}_{\text{img},k}(\theta_t, \mathbf{M}_t)\|^2, \quad (16)$$

where w_k is the confidence score associated with a landmark.

Prior Term The prior term is the same as in Eq. (4), now accumulated for all poses Θ and scaled by the number of poses N_{Θ} .

4.4 Optimization

In order to solve the optimization problems related to obtaining initial 3D poses in Eq. (4), obtaining camera poses to minimize re-projection error and to jointly optimize all variables in Eq. (12), we apply gradient-based Levenberg-Marquardt.

5 Results

To validate our approach quantitatively (Section 5.1 and Section 5.2), we use the recent TotalCapture [39] dataset, which is the only one including IMU data and video synchronized with ground-truth. In Section 5.3 we then provide details of the newly recorded 3DPW dataset, demonstrate 3D pose reconstruction of VIP in challenging scenes, and evaluate the accuracy of automatic 2D to 3D pose assignment in multiple-person scenes.

5.1 Tracker Parameters

Pose assignment: In the graph G , edges $e \in \mathcal{E}$ are created between any two nodes that are at most 30 frames apart. The weights mapping from features to costs are learned using 5 sequences from 3DPW dataset, which have been manually labeled for this purpose. Given the features \mathbf{f} defined in Section 4.2 and learned weights α from logistic regression, we turn features into costs via $c = -\langle \mathbf{f}, \alpha \rangle$, making the optimization problem (5) probabilistically motivated [37].

Video Inertial Fusion: Different weighting parameters in Eq. (4) and Eq. (12) produce good results as long as they are balanced. However, rather than setting them by hand, we used Bayesian Optimization [3] in the proposed training set of TotalCapture (seen subjects). The values found are $w_{\text{acc}} = 0.2$, $w_{\text{img}} = 0.0001$ and $w_{\text{prior}} = 0.006$ and are kept fixed for all experiments. Note, that these are very few parameters and therefore, there is very little risk of over-fitting, which is also reflected in the results.

5.2 Tracking Accuracy

We quantitatively evaluate tracking accuracy on the TotalCapture dataset. The dataset consists of 5 subjects performing several activities such as walking, acting, range of motions and freestyle motions – which are recorded using 8 calibrated, static RGB-cameras and 13 IMUs attached to head, sternum, waist, upper arms, lower arms, upper legs, lower legs and feet. Ground-truth poses are obtained using a marker-baser motion capture system. All data is synchronized and operates at a framerate of 60Hz. The ground truth poses are provided as joint positions, which do not contain information about pronation and supination angles; i.e. rotations about the bone’s long axis. To obtain full degree of freedom pose, we fit the SMPL model to the raw ground-truth markers using a method similar to [15].

Approach	[39]	[16]	IT	VIP-2D	VIP-Cam	VIP-IMU6	VIP-IT	VIP
MPJPE	70.0	(62)	55.0	15.1	25.3	39.6	28.2	26.0
MPJAE	-	-	16.9	10.1	12.1	15.3	12.0	12.1

Table 1. Mean Joint Position Error (MPJPE) in mm and Mean Per Joint Angular Error (MPJAE) in degrees evaluated on TotalCapture.

Error Metrics: We report: Mean Per Joint Position Error (MPJPE) and Mean Per Joint Angular Error (MPJAE). MPJPE is the average Euclidean distance between ground-truth and estimated joint positions of hips, knees, ankles, neck, head, shoulders, elbows and wrists; MPJAE is the average geodesic distance between ground-truth and estimated joint orientations for hips, knees, neck, shoulders and elbows. In order to evaluate pose accuracy independently of absolute camera position and orientation, we align our estimates with the ground-truth. This is standard practice in existing benchmarks [8]. Thus, in our case MPJPE is a measure of pose accuracy independent of global position and orientation.

Results: Our tracking results on TotalCapture are summarized in Table 1. We used only 1 camera and the 13 IMUs provided. The cameras in TotalCapture are rigidly mounted to the building and are not equipped with an IMU – hence we assumed a static camera with *unknown* pose. VIP achieves a remarkably low average MPJPE of 26mm and a MPJAE of only 12.1°.

Comparisons to state-of-the-art: We outperform the learning-based approach introduced in the TotalCapture dataset [39] by 44mm – the approach uses *all 8 cameras* and fuses IMU data with a probabilistic visual hull. We also outperform [16], who report a mean MPJPE of 62mm using 8 cameras and all 13 IMUs. Admittedly, it is difficult to compare approaches, since [39] and [16] process the data in a frame-by-frame manner which is an advantage w.r.t. VIP, which jointly optimizes over all frames simultaneously. However, VIP uses only a single camera with unknown pose whereas the *competitors use 8 fully calibrated cameras*. To understand better the influence of components of VIP we also report the tracking accuracy for five tracker variants in Table 1.

Comparison to IMU only: The Inertial tracker (IT) corresponds to the single frame approach of Section 4.1. It uses only raw IMU orientations and is the initialization for VIP. Over all sequences, IT achieves a MPJPE of 55mm. VIP decreases this error by more than 50%. This demonstrates the usefulness of fusing image information and optimizing heading angles.

Heading drift and misalignments: We report results of VIP-IT to demonstrate the influence of optimizing heading angles, and sensor-to-bone misalignments originating from an inaccurate initial pose. VIP-IT is identical to IT, but uses the heading angles and initial pose obtained with VIP. VIP-IT is only slightly less accurate than VIP validating the importance of inferring drift and accurate initial pose. More evaluations are shown in the supplementary material.

Robustness to 2D pose accuracy: VIP-2D is identical to the VIP but utilizes ground-truth 2D poses obtained by projecting ground-truth joint posi-



Fig. 6. We show example results obtained using VIP for some challenging activities. With VIP we get accurate 3D poses aligning well with the images using the estimated camera poses.

tions to the images. VIP-2D achieves a MPJPE of 15.1mm which indicates how much VIP can improve if 2D pose estimation methods keep improving.

Robustness to camera pose: VIP-Cam is also almost identical to VIP, but uses the ground-truth camera pose instead of estimating it. The MPJPE of VIP-Cam is 25.3mm, which is only 0.7mm better compared to VIP.

Fewer sensors: We report the error of VIP using 6 IMUs similar to [42], denoted as VIP-IMU6. The combination of only 6 IMUs and 2D pose information achieves a MPJPE of 39.6mm, which is 13.6mm higher than VIP-13 IMUs but still very accurate. This demonstrates our approach could be used for applications where a minimal number of sensors is required.

This quantitative evaluation demonstrates the accuracy of VIP. Ideally, we would evaluate VIP quantitatively also in challenging scenes, like the ones in 3DPW. However, there exists no dataset with a comparable setting and ground-truth, which was one of the main motivations of this work.

5.3 3D Poses in the Wild Dataset

VIP allowed us to achieve the second goal of this work: recording a dataset with accurate 3D pose in challenging outdoor scenes with a moving camera. A hand-held smartphone camera was used to record one or two IMU-equipped actors performing various activities such as shopping, doing sports, hugging, discussing, capturing selfies, riding bus, playing guitar, relaxing. The dataset includes 60 sequences, more than 51,000 frames and 7 actors in a total of 18 clothing styles. We also scanned subjects and non-rigidly fitted SMPL to obtain 3D models similar to [27, 46]. For single subject tracking, we attached 17 IMUs to all major bone segments. We used 9-10 IMUs per person to simultaneously track up to 2 subjects. During all recordings one additional IMU was attached to the smartphone. Video and inertial data was automatically synchronized by a clapping motion at the beginning of a sequence as in [24]. For every sequence, the subjects were asked to start in an upright pose with closed arms. In Fig. 6 we show tracking results illustrating the 3D model alignment with the images. Fig. 7 shows more tracking results, where we animated the 3D models with the reconstructed poses. 3DPW is the most challenging dataset (with 3D pose annotation) for state-of-the-art 3D pose estimation methods as evidenced by the results reported in the supplemental material.

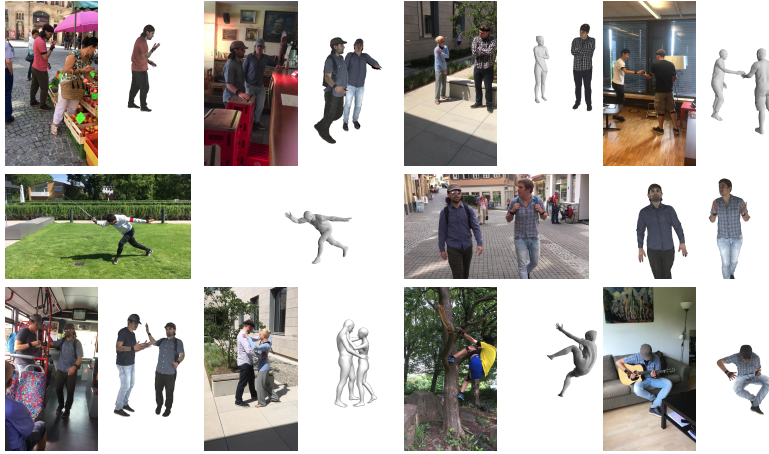


Fig. 7. We show several example frames of sequences in the 3DPW. The dataset contains large variations in person identity, clothing and activities. For a couple of cases we also show animated, textured SMPL body models.

Assignment Accuracy: In comparison to TotalCapture, the additional challenges in 3DPW originate from multiple people in the scene. Hence, we assessed the accuracy of our automatic assignment of 2D poses to 3D poses using manually labelled 2D pose candidate IDs. VIP achieves an assignment precision of 99.3% and a recall rate of 92.2% demonstrating the method correctly identifies the tracked persons for the vast majority of frames. This is a strong indication that VIP achieves a 3D pose accuracy on 3DPW comparable to the MPJPE of 26mm reported for TotalCapture.

6 Conclusions

Combining IMUs and a moving camera, we introduced the first method that can robustly recover pose in challenging scenes. The main challenges we addressed are: person identification and tracking in cluttered scenes, and joint recovery of 3D pose for 2 subjects, camera and IMU heading drift. We combined discrete optimization to find associations, with continuous optimization to effectively fuse the sensor information. Using our method, we collected the *3D Poses in the Wild* dataset, including challenging sequences with accurate 3D poses that we make available for research purposes. With VIP it is possible to record people in natural video easily and we plan to keep adding to the dataset. We anticipate the proposed dataset will provide the means to quantitatively evaluate monocular methods in difficult scenes and stimulate new research in this area.

Acknowledgements. We thank Jorge Márquez, Senya Polikovsky, Matvey Safroshkin and Andrea Keller for the technical support.

References

1. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 623–630 (2010)
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: European Conference on Computer Vision (ECCV) (2016)
3. Bull, A.D.: Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research* **12**(Oct), 2879–2904 (2011)
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
5. Gurobi Optimization, I.: Gurobi optimizer reference manual (2016)
6. Helten, T., Baak, A., Bharaj, G., Muller, M., Seidel, H.P., Theobalt, C.: Personalization and evaluation of a real-time depth-based full body tracker. In: 3D Vision (3DV) (2013)
7. Henschel, R., Leal-Taixé, L., Cremers, D., Rosenhahn, B.: Fusion of head and full-body detectors for multi-object tracking. In: Computer Vision and Pattern Recognition Workshops (CVPRW) (2018)
8. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **36**(7), 1325–1339 (2014)
9. Jahangiri, E., Yuille, A.L.: Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In: IEEE International Conference on Computer Vision (ICCV) Workshops (PeopleCap) (2017)
10. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
11. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2 (2017)
12. Levinkov, E., Uhrig, J., Tang, S., Omran, M., Insafutdinov, E., Kirillov, A., Rother, C., Brox, T., Schiele, B., Andres, B.: Joint graph decomposition & node labeling: Problem, algorithms, applications. In: CVPR. vol. 7. IEEE (2017)
13. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3d human pose estimation. In: IEEE International Conference on Computer Vision (ICCV). pp. 2848–2856 (2015)
14. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. *ACM Trans. Graphics* **34**(6), 248:1–248:16 (2015)
15. Loper, M.M., Mahmood, N., Black, M.J.: MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* **33**(6), 220:1–220:13 (2014)
16. Maleson, C., Volino, M., Gilbert, A., Trumble, M., Collomosse, J., Hilton, A.: Real-time full-body motion capture from video and imus. In: 2017 Fifth International Conference on 3D Vision (3DV) (2017)
17. von Marcard, T., Pons-Moll, G., Rosenhahn, B.: Human pose estimation from video and IMUs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **38**(8), 1533–1547 (2016)

18. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: IEEE International Conference on Computer Vision (ICCV) (2017)
19. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3D Vision (3DV). IEEE (2017)
20. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d body pose estimation from monocular rgb input. arXiv preprint arXiv:1712.03453 (2017)
21. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG) **36**(4), 44 (2017)
22. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3D human pose and shape from a single color image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
23. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
24. Pons-Moll, G., Baak, A., Gall, J., Leal-Taixé, L., Muller, M., Seidel, H.P., Rosenhahn, B.: Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In: Proceedings of the 2011 International Conference on Computer Vision (ICCV). pp. 1243–1250 (2011)
25. Pons-Moll, G., Baak, A., Helten, T., Müller, M., Seidel, H.P., Rosenhahn, B.: Multisensor-fusion for 3d full-body human motion capture. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 663–670 (2010)
26. Pons-Moll, G., Fleet, D.J., Rosenhahn, B.: Posebits for monocular human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2337–2344 (2014)
27. Pons-Moll, G., Pujades, S., Hu, S., Black, M.: ClothCap: Seamless 4D clothing capture and retargeting. ACM Transactions on Graphics, (Proc. SIGGRAPH) **36**(4) (2017)
28. Popa, A.I., Zanfir, M., Sminchisescu, C.: Deep multitask architecture for integrated 2d and 3d human sensing. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
29. Rhodin, H., Spörri, J., Katircioglu, I., Constantin, V., Meyer, F., Müller, E., Salzmann, M., Fua, P.: Learning monocular 3d human pose estimation from multi-view images. In: CVPR (2018)
30. Roetenberg, D., Luinge, H., Slycke, P.: Moven: Full 6dof human motion tracking using miniature inertial sensors. Xsen Technologies, December (2007)
31. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2d and 3d pose detection in natural images. arXiv preprint arXiv:1803.00455 (2018)
32. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision (IJCV) **87**(1-2) (2010)
33. Simo-Serra, E., Quattoni, A., Torrass, C., Moreno-Noguer, F.: A joint model for 2d and 3d pose estimation from a single image. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3634–3641 (2013)
34. Simo-Serra, E., Ramisa, A., Alenyà, G., Torrass, C., Moreno-Noguer, F.: Single image 3d human pose estimation from noisy observations. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2673–2680 (2012)

35. Sminchisescu, C., Triggs, B.: Kinematic jump processes for monocular 3d human tracking. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2003)
36. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. arXiv preprint arXiv:1704.00159 (2017)
37. Tang, S., Andres, B., Andriluka, M., Schiele, B.: Subgraph decomposition for multi-target tracking. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5033–5041 (2015)
38. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
39. Trumble, M., Gilbert, A., Malleison, C., Hilton, A., Collomosse, J.: Total capture: 3d human pose estimation fusing video and inertial sensors. In: Proceedings of 28th British Machine Vision Conference. pp. 1–13 (2017)
40. Tung, H.Y., Tung, H.W., Yumer, E., Fragkiadaki, K.: Self-supervised learning of motion capture. In: NIPS (2017)
41. Vlastic, D., Adelsberger, R., Vannucci, G., Barnwell, J., Gross, M., Matusik, W., Popović, J.: Practical motion capture in everyday surroundings. *ACM Transactions on Graphics (TOG)* **26**(3), 35 (2007)
42. von Marcard, T., Rosenhahn, B., Black, M., Pons-Moll, G.: 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) pp. 349–360 (2017)
43. Wandt, B., Ackermann, H., Rosenhahn, B.: 3d reconstruction of human motion from monocular image sequences. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **38**(8), 1505–1516 (2016)
44. Wang, C., Wang, Y., Lin, Z., Yuille, A.L., Gao, W.: Robust estimation of 3d human poses from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2361–2368 (2014)
45. Zell, P., Wandt, B., Rosenhahn, B.: Joint 3d human motion capture and physical analysis from monocular videos. In: The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)
46. Zhang, C., Pujades, S., Black, M., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3D scan sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
47. Zheng, Z., Yu, Tao, L.H., Guo, K., Dai, Q., Fang, L., Liu, Y.: Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In: European Conference on Computer Vision (ECCV) (2018)
48. Zhou, F., De la Torre, F.: Spatio-temporal matching for human detection in video. In: European Conference on Computer Vision (ECCV). pp. 62–77 (2014)
49. Zhou, X., Leonardos, S., Hu, X., Daniilidis, K.: 3D shape estimation from 2D landmarks: A convex relaxation approach. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4447–4455 (2015)
50. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: A weakly-supervised approach. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 398–407 (2017)