EUROGRAPHICS 2017 / L. Barthe and B. Benes
(Guest Editors)

Volume 36 (2017), Number 2

Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs

T. von Marcard[1] B. Rosenhahn[1] M. J. Black[2] G. Pons-Moll[2]

[1]Institut für Informationsverarbeitung (TNT), Leibniz-Universität Hannover, Germany
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

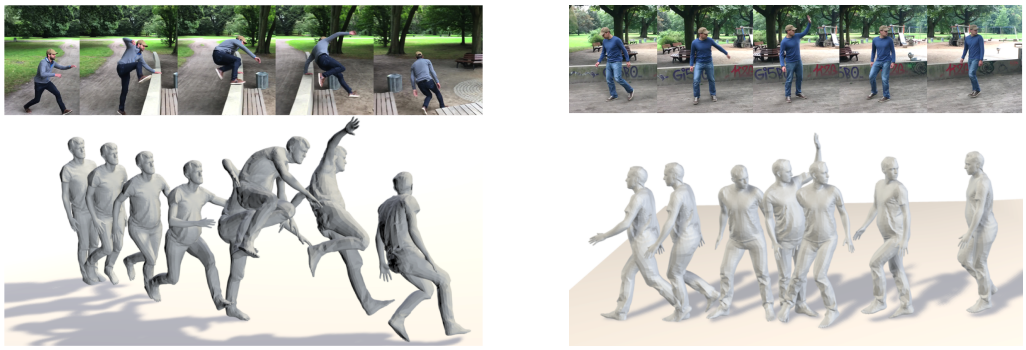


Figure 1: *Unconstrained motion capture using our new Sparse Inertial Poser (SIP). With as few as 6 IMUs attached to the body, we recover the full pose of the subject. The key idea that makes this possible is to optimise all the poses of a statistical body model for all the frames in the sequence jointly to fit the orientation and acceleration measurements captured by the IMUs. Images are shown for reference but are not used during the optimisation.*

Abstract

We address the problem of making human motion capture in the wild more practical by using a small set of inertial sensors attached to the body. Since the problem is heavily under-constrained, previous methods either use a large number of sensors, which is intrusive, or they require additional video input. We take a different approach and constrain the problem by: (i) making use of a realistic statistical body model that includes anthropometric constraints and (ii) using a joint optimization framework to fit the model to orientation and acceleration measurements over multiple frames. The resulting tracker Sparse Inertial Poser (SIP) enables motion capture using only 6 sensors (attached to the wrists, lower legs, back and head) and works for arbitrary human motions. Experiments on the recently released TNT15 dataset show that, using the same number of sensors, SIP achieves higher accuracy than the dataset baseline without using any video data. We further demonstrate the effectiveness of SIP on newly recorded challenging motions in outdoor scenarios such as climbing or jumping over a wall.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation

1. Introduction

The recording of human motion has revolutionized the fields of biomechanics, computer animation, and computer vision. Human motion is typically captured using commercial marker-based systems such as [Vic] or [Sim], and numerous recordings of human performances are now available (*e.g.,* [CMU], [Mix], [Mov]). The recording of human motion is also important for psychology and medicine, where biomechanical analysis can be used to assess physical activity and diagnose pathological conditions and monitor post-operative mobility of patients. Unfortunately, marker-based systems are intrusive and restrict motions to controlled laboratory spaces. Therefore, activities such as skiing, biking or simple daily activities like having coffee with friends cannot be recorded with such systems. The vision community has seen significant progress

in the estimation of 3D human pose from images, but this typically involves multi-camera calibrated systems, which again limit applicability. Existing methods for estimating 3D human pose from single images, e.g. [BKL*16], are still less accurate than motion capture systems. However, to record human motion in everyday situations and in natural settings one would need a dedicated camera to track a specific subject. Hence, it is unlikely that vision-based systems will be able to record large amounts of continuous daily activity data.

Systems based on Inertial Measurement Units (IMUs) do not suffer from such limitations; they can track the human pose without cameras which make them more suitable for outdoor recordings, scenarios with occlusions, baggy clothing or where tracking with a dedicated camera is simply not possible. However, inertial measurement systems such as Xsens BioMech [Xse] are quite intrusive, requiring 17 sensors worn on the body or attached to a suit. This is one of the reasons that large amounts of data have not been recorded yet. Hence, a less intrusive solution that can capture people through occlusions is needed.

In this paper, we present the Sparse Inertial Poser (SIP), a method to recover the full 3D human pose from only 6 IMUs. Six sensors, measuring orientation and acceleration are attached to the wrists, lower legs, waist and head, resulting in a minimally intrusive solution to capture human activities. Furthermore, many consumer products already have IMUs integrated, e.g., fitness and smartwatches, smartphones, Google glasses and Oculus rift. Our 6-sensor system could easily be worn with a hat or glasses, two wrist bands, a belt, and shoe or ankle sensors. However, recovering human pose from only 6 IMUs is a very difficult task. Orientation at the extremities and waist only provides a weak constraint on the human motion and incorporation of acceleration data is usually affected by drift.

To solve this problem, we exploit the rich statistical SMPL body model [LMR*15]. One key insight is that the body model can be fit to incomplete and ambiguous data because it captures information about the kinematic constraints of the human body. A similar observation has been made by [TST*15] and [TBC*16] who leveraged a statistical model for hand pose tracking. Unfortunately, this alone is not sufficient to compensate for drift. Most previous methods (e.g. [RLS07, VAV*07]) integrate acceleration frame by frame, which results in unstable estimates when using very few sensors. Optimizing frame by frame is similar to a double explicit integration scheme, which is known to be unstable and only accurate within small time intervals.

We take a different approach and optimize all the poses of all the frames of a sequence at once. Hence, our objective function enforces the coherency between the body model orientation and acceleration estimates against the IMU recordings. Effectively, the realistic body model simplifies the estimation problem, providing sufficient constraints to solve the problem from sparse measurements, even for complex movements. Some examples are shown in Fig. 1.

In several experiments we show that SIP, while simple, is very powerful and can recover all the poses of a sequence as a result of a single optimization. We report results on the recently released TNT15 dataset [MPMR16] which features 4 subjects wearing 10 IMUs performing a variety of human actions. To evaluate SIP we use 6 IMUs for tracking and 4 IMUs for validation. We compare to two baselines, namely an orientation-only tracker that uses only the orientation information and a variant of SIP that uses a different human body model. Qualitative and quantitative results demonstrate that SIP is significantly more accurate than the baselines. To further demonstrate the applicability of SIP, we present additional tracking results of two subjects wearing 6 IMUs in an outdoor setting (see Fig. 1).

In summary, SIP makes the challenging problem of human pose estimation from sparse IMU data feasible by:

- Making use of a realistic body model that incorporates anthropomorphic constraints (with a skeletal rig).
- A joint optimization framework that fits the poses of a body model to the orientation and acceleration measurements over multiple frames.

Altogether SIP is the first method that is able to estimate the 3D human pose from only 6 IMUs without relying on databases of MoCap or learning methods that make strong assumptions about the recorded motion.

2. Related Work

The literature on human pose estimation from images is vast and in this paper we focus only on methods integrating multiple sensor modalities and methods predicting full pose from sparse low dimensional control signals.

2.1. Database retrieval and learning based methods

Some work has focused on recovering full pose from sparse incomplete sensor signals. In [SH08, TZK*11] they reconstruct human pose from 5 accelerometers by retrieving pre-recorded poses with similar accelerations from a database. Acceleration data is however very noisy and the space of possible accelerations is huge which makes learning a very difficult task. A somewhat easier problem is addressed in [CH05]; they reconstruct full 3D pose from a set of sparse markers attached at the body. They build online local PCA models using the sparse marker locations as input to query the database of human poses. This approach works well since the 5-10 marker locations can constrain the pose significantly; furthermore the mapping from 3D locations to pose is much more direct than from acceleration data. Unfortunately, this approach is restricted to a lab with cameras capturing the reflective markers. Following similar ideas, in [LWC*11] they regress to full pose using online local models but using 6 IMUs to query the database. In [SMN09] they directly regress full pose using only 4 IMUs with Gaussian Process regression. Both methods report very good results when the test motions are present in the database. In [HKP*16] they extract gait parameters using deep convolutional neural networks. Although pre-recorded human motion greatly constrains the problem, methods that heavily rely on pre-recorded data are limited; in particular capturing arbitrary activities is difficult if it is missing in the databases.

2.2. Full-body IMU MoCap

There exist commercial solutions for human motion capture from IMUs; [RLS07] use 17 IMUs equipped with 3D accelerometers, gyroscopes and magnetometers and all the measurements are fused using a Kalman Filter. By achieving stable orientation measurements the 17 IMUs completely define the pose of the subject. However it is very intrusive for a subject to wear them, and long setup times are required. In the seminal work of [VAV*07] they propose a custom made system consisting of 18 sensor boards, each equipped with an IMU and acoustic distance sensors, to compensate for typical drift in the orientation estimates. While the approach is demonstrated in challenging outdoor settings like ours, the system is also very intrusive and difficult to reproduce. Other approaches have combined sparse IMUs with video input [PMBG*11, MPMR16] or sparse optical markers [AHK*16] to constrain the problem. Similarly [HMST13] combines sparse IMUs with a depth camera. IMUs are only used to query similar poses in a database and depth data is used to obtain the full pose. While powerful, using video input does not allow human movements to be captured with occlusions or in applications that require continuous activity monitoring. Hence, instead of constraining the problem using additional sensors, we constrain the problem by using a statistical body model and optimizing the pose over multiple frames. While 6 IMUs do not provide enough constraints to determine the full pose for a single frame, we find that accurate pose estimates can be obtained when integrating all orientation and acceleration measurements into a single optimization objective.

3. Background

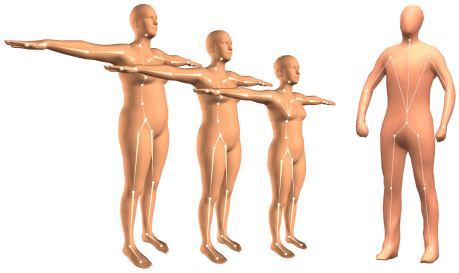
3.1. Exponential Map on SO(3) and SE(3)

In this section we quickly review the concept of exponential mapping on the Special Orthogonal Group SO(3) and the Special Euclidean Group SE(3). The exponential map representation provides a geometric and elegant treatment of rigid body motion, which we use to relate pose parameters to human body motions. Using the exponential map has some advantadges for optimization w.r.t. other representations such as Euler angles [PMR09]; for more details on the exponential mapping and a comparison to other parameterizations we refer the reader to [MLSS94, PMR11].

Both SO(3) and SE(3) are Lie groups with an associated Lie algebra. Throughout this paper we will use the cross-operator $^\times$ to construct a Lie algebra element from its coordinates and the vee-operator $^\vee$ to extract the coordinates of a Lie algebra element into a column vector. The group of rotations about the origin in 3 dimensions SO(3) is defined as $\mathrm{SO}(3) = \{\mathbf{R} \in \mathbb{R}^{3\times3} : \mathbf{R}^T\mathbf{R} = \mathbf{I}, det(\mathbf{R}) = 1\}$. Every rotation $\mathbf{R}$ can be expressed in exponential form

$$\mathbf{R} = \exp(\omega^\times), \tag{1}$$

where $\omega^\times \in so(3)$ is a skew-symmetric matrix and can be computed analytically using the Rodriguez Formula [MLSS94]. The three independent parameters $\omega \in \mathbb{R}^3$ of $\omega^\times$ are called exponential coordinates of $\mathbf{R}$ and define the axis of rotation and $||\omega||$ is the angle of rotation about this axis. The group SE(3) represents rigid body motions composed by a rotation $\mathbf{R} \in \mathrm{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$. Any rigid motion $\mathbf{G} \in \mathbb{R}^{4\times4}$ can be written in exponential form



(a) (b)

Figure 2: *(a) The joints of the skeleton in SMPL are predicted as a function of the surface. This allows us to obtain accurate joint locations which are used to predict the acceleration measurements. (b) Manually rigged models lead to worse performance fitting incomplete sensor measurements.*

$$\mathbf{G} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} = \exp(\xi^\times), \tag{2}$$

where $\xi^\times \in se(3)$ is called the associated twist action and $se(3)$ refers to the corresponding Lie algebra. The six independent parameters $\xi \in \mathbb{R}^6$ of $\xi^\times$ are called exponential coordinates of $\mathbf{G}$. They are composed of the rotational parameters $\omega \in \mathbb{R}^3$ and $\mathbf{v} \in \mathbb{R}^3$, where the latter encodes location of the axis of rotation and translation along the axis.

The inverse operation of Eq. (1) and Eq. (2) is the logarithm and recovers a Lie algebra element from a Lie group element. We also introduce the Taylor expansion of the matrix exponential given by

$$\exp(\xi^\times) = \mathbf{I} + \xi^\times + \frac{(\xi^\times)^2}{2!} + \frac{(\xi^\times)^3}{3!} + \ldots, \tag{3}$$

and the first-order approximation for the logarithm

$$\log(\exp(\delta\omega^\times)\exp(\omega^\times))^\vee \approx \delta\omega + \omega, \tag{4}$$

for a small $\delta\omega \in \mathbb{R}^3$.

3.2. SMPL Body Model

SMPL [LMR*15] is a body model that uses a learned template with $V = 6890$ vertices $\mathbf{T}$, and a learned rigged template skeleton. The actual vertex positions of SMPL are adapted according to identity-dependent shape parameters and the skeleton pose. The skeletal structure of the human body is modeled with a kinematic chain consisting of rigid bone segments linked by $n = 24$ joints. Each joint is modeled as a ball joint with 3 rotational Degrees of Freedom (DoF), parametrized with exponential coordinates ω. Including translation, the pose $\mathbf{x}$ is determined by a pose vector of $d = 3 \times 24 + 3 = 75$ parameters. The rigid motion $\mathbf{G}^{TB}(\mathbf{x})$ of a bone depends on the states of parent joints in the kinematic chain and can be computed by the

forward kinematic map $\mathbf{G}^{TB}: \mathbb{R}^d \rightarrow \mathrm{SE}(3)$:

$$\mathbf{G}^{TB}(\mathbf{x}) = \left(\prod_{j \in I(i)} \left[\begin{array}{c|c} \exp(\omega_j^{\times}) & \mathbf{j} \\ \hline \vec{0} & 1 \end{array} \right] \right) = \left(\prod_{j \in I(i)} \exp\left(\xi_j^{\times}\right) \right), \tag{5}$$

where $I(i) \subseteq \{1, \cdots, n+1\}$ is an ordered set of parent joints, $\omega_j \in \mathbb{R}^3$ are the exponential coordinates of the joint rotation, $\mathbf{j} \in \mathbb{R}^3$ is the joint location and $\xi_j^{\times} \in se(3)$ is the twist action of joint j. The initial offset between the bone and the tracking frame is the identity.

SMPL models body shape variation using shape blend shapes, that are linearly added to the template mesh. A new subject shape is typically obtained by adding a linear combination of blendshapes $\mathbf{S}_i \in \mathbb{R}^{3V}$ to the template mesh $\mathbf{T}' = \mathbf{T} + \sum_i \beta_i \mathbf{S}_i$. SMPL automatically predicts the joint locations $\mathbf{Q} = [\mathbf{j}_1^T \ldots \mathbf{j}_n^T]^T$ as a function of the surface mesh using a sparse regression matrix $\mathbf{Q} = \mathcal{J}\mathbf{T}'$. While the orientation of the limbs do not depend at all on the body joints, the linear acceleration of a particular part of the body depends on the joint locations. By using SMPL we can track any shape without having to manually edit the skeleton, see Figure 2(a).

3.3. IMUs

An Inertial Measurement Unit (IMU) is a device that is commonly equipped with 3-axes accelerometers, gyroscopes and magnetometers. It measures acceleration, rate of turn and magnetic field strength with respect to the IMU-aligned sensor coordinate system F^S. Typically, a Kalman Filter is then applied to track the sensor orientation with respect to a global inertial coordinate system F^I.

In order to utilize IMU data together with the body model we introduce several coordinate systems depicted in Figure 3(a). The body model is defined in the global tracking coordinate system F^G and each bone segment of the body has a local coordinate system F^B. The map $\mathbf{G}^{GB}: F^B \rightarrow F^G$ defines the mapping from bone to tracking coordinate system. Equivalently, $\mathbf{G}^{IS}: F^S \rightarrow F^I$ defines the mapping from the local IMU sensor coordinate system F^S to F^I. Both global coordinate systems F^G and F^I are related by the constant mapping $\mathbf{G}^{GI}: F^I \rightarrow F^G$. In the following we will assume $\mathbf{G}^{GI}$ is known and express all IMU readings in the global tracking frame F^G using the transformation rule

$$\mathbf{G}^{GS}(t) = \mathbf{G}^{GI}\mathbf{G}^{IS}(t). \tag{6}$$

For a more detailed description of relating inertial data to other sensor or model coordinate systems we refer the reader to [BHM*10]. Our aim is to find a pose trajectory such that the motion of a limb is consistent with IMU acceleration and orientation attached to it. Thus we need to know the offset between IMU and its corresponding bone coordinate system $\mathbf{G}^{BS}(t): F^S \rightarrow F^B$. We assume that it is constant as the sensors are tightly attached to the limbs and compute it at the first frame of the tracking sequence according to

$$\mathbf{G}^{BS} = \mathbf{G}^{BG}(0)\mathbf{G}^{GS}(0). \tag{7}$$

4. Sparse Inertial Poser

Recovering full pose from only $N = 6$ IMUs (strapped at lower arms, lower legs, head and waist) is highly ambiguous. Assuming no sensor noise, orientation data only constrains the full pose to lie on a lower dimensional manifold. Acceleration measurements are noisy and naive double integration to obtain position leads to unbounded exponential drift, see Figure 4. Looking at a single frame the problem is ill-posed. However, looking at the full sequence, and using anthropometric constraints from a body model, makes the problem much more constrained, see Figure 5. This motivates us to formulate the following multi-frame objective function:

$$\mathbf{x}_{1:T}^* = \arg\min_{\mathbf{x}_{1:T}} E_{\text{motion}}(\mathbf{x}_{1:T}, \mathbf{R}_{1:T}, \mathbf{a}_{1:T}), \tag{8}$$

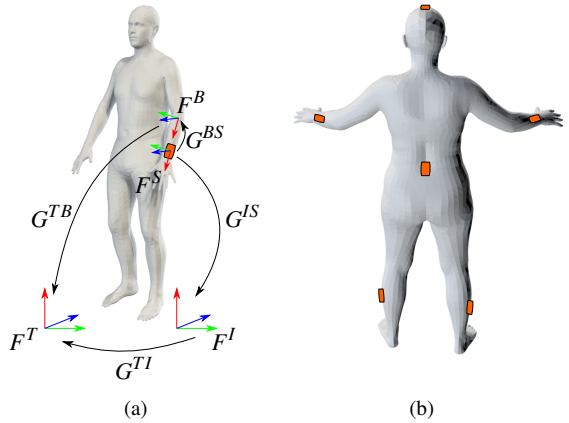


Figure 3: *(a) Coordinate frames: Global tracking coordinate frame F^G, Inertial coordinate frame F^I, Bone coordinate frame F^B and Sensor coordinate frame F^S. (b) Sensor placement at head, lower legs, wrists and back.*

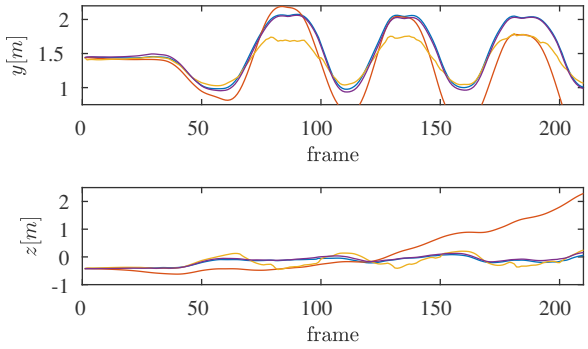


Figure 4: *Y- and Z-coordinates of the left wrist sensor position (Y pointing upwards) for a jumping jack sequence, which is also shown in Figure 7. Ground truth positions obtained by tracking with 10 IMUs, are shown in purple and are almost indistinguishable from the estimated sensor positions obtained with SIP (blue). Using only orientation (yellow) of 6 IMUs provides accurate estimates for some portions of the sequence, but cannot correctly reconstruct the extended, raised arm. Double integrating acceleration values (red) provides only reasonable estimates at the beginning of the sequences and the error accumulates over time.*

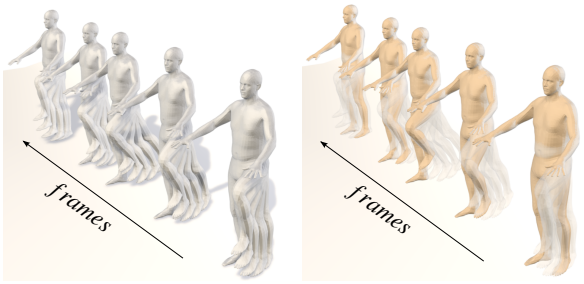


Figure 5: *SIP joint optimization: sparse IMUs give only weak constraints on the full pose. As illustrated on the left figure, multiple poses fit well the IMU orientation of the lower left leg. By optimizing all poses over the sequence we can successfully find the pose trajectory (shown in orange) that is also consistent with the acceleration data as can be seen on the right figure. The joint optimization allows the use of acceleration readings, which would produce severe drift otherwise.*

where $\mathbf{x}_{1:T} \in \mathbb{R}^{75T}$ is a vector consisting of stacked model poses for each time step $t = 1 \dots T$. $\mathbf{R}_{1:T}$ are the sensor orientations $\mathbf{R}_t \in SO(3)$ and $\mathbf{a}_{1:T}$ are the sensor acceleration measurements respectively. We define $E_{\text{motion}} : \mathbb{R}^{d \times T} \times \mathbb{R}^{3N \times T} \times \mathbb{R}^{3N \times T} \rightarrow \mathbb{R}$ as

$$\begin{aligned} E_{\text{motion}}(\mathbf{x}_{1:T}, \omega_{1:T}, \mathbf{a}_{1:T}) = & w_{\text{ori}} \cdot E_{ori}(\mathbf{x}_{1:T}, \mathbf{R}_{1:T}) \\ & + w_{\text{acc}} \cdot E_{\text{acc}}(\mathbf{x}_{1:T}, \mathbf{a}_{1:T}) \\ & + w_{\text{anthro}} \cdot E_{\text{anthro}}(\mathbf{x}_{1:T}), \end{aligned} \tag{9}$$

where E_{ori}, E_{acc} and E_{anthro} are energies related to orientation, acceleration and anthropometric consistency. The weights of Eq. (9) are fixed during all experiments, see experimental section. In the following, we detail each of the objective terms.

4.1. The Orientation Term

The sensor orientations, $\mathbf{R}^{GS}(t) : F^S \rightarrow F^G$ are related to the bone orientations by a constant rotational offset $\mathbf{R}^{BS}$. Hence, we define the estimated sensor orientation $\hat{\mathbf{R}}^{GS}(\mathbf{x}_t)$ at the current pose $\mathbf{x}_t$ as

$$\hat{\mathbf{R}}^{GS}(\mathbf{x}_t) = \mathbf{R}^{GB}(\mathbf{x}_t)\mathbf{R}^{BS}, \tag{10}$$

where $\mathbf{R}^{GB}(\mathbf{x}_t)$ is the rotational part of the forward kinematics map defined in Eq. (5) and $\mathbf{R}^{BS}$. The *orientation error* $\mathbf{e}_{ori} \in \mathbb{R}^3$ are the exponential coordinates of the rotational offset between estimated and measured sensor orientation:

$$\mathbf{e}_{ori}(\mathbf{x}_t) = \log\left(\hat{\mathbf{R}}^{GS}(\mathbf{x}_t)\left(\mathbf{R}^{GS}(t)\right)^{-1}\right)^{\vee}, \tag{11}$$

where the $\vee$-operator is used to extract the coordinates of the skew-symmetric matrix obtained from the log-operation. We define the orientation consistency E_{ori} across the sequence as

$$E_{ori} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{n=1}^{N} ||\mathbf{e}_{ori,n}(t)||^2, \tag{12}$$

which is the sum of squared L2-norm of orientation errors over all frames t and all sensors n. Actually, the squared L2-norm of $\mathbf{e}_{ori}$ corresponds to the geodesic distance between $\hat{\mathbf{R}}^{GS}(\mathbf{x}_t)$ and $\mathbf{R}^{GS}(t)$ [HTDL13, MPMR16].

4.2. The Acceleration Term

IMU acceleration measurements $\mathbf{a}^S$ are provided in the sensor coordinate system F^S shown in Figure 3(a). To obtain the corresponding sensor acceleration $\mathbf{a}^G$ in global tracking frame coordinates F^G we have to transform $\mathbf{a}^S$ by the current sensor orientation $\mathbf{R}^{GS}(t)$ and subtract gravity $\mathbf{g}^G$

$$\mathbf{a}_t^G = \mathbf{R}_t^{GS}\mathbf{a}_t^S - \mathbf{g}^G. \tag{13}$$

We aim to recover a sequence of poses such that the actual sensor acceleration matches the corresponding vertex acceleration of the body model. The corresponding vertex is manually selected; since the model has the same topology across subjects this operation is done only once. The vertex acceleration $\hat{\mathbf{a}}^G(t)$ is approximated by numerical differentiation

$$\hat{\mathbf{a}}_t^G = \frac{\mathbf{p}_{t-1}^G - 2\mathbf{p}_t^G + \mathbf{p}_{t+1}^G}{dt^2}, \tag{14}$$

where $\mathbf{p}_t^G$ is the vertex position at time instance t and dt is the sampling time. The vertex position is related to the model pose $\mathbf{x}$ by the forward kinematic map defined in Eq. (5) and equates to

$$\bar{\mathbf{p}}^G(\mathbf{x}) = \mathbf{G}^{GB}(\mathbf{x})\bar{\mathbf{p}}^B(0), \tag{15}$$

where $\bar{\mathbf{p}}$ indicates homogeneous coordinates. Hence, we define the acceleration error as the difference of estimated and measured acceleration

$$\mathbf{e}_{\text{acc}}(t) = \hat{\mathbf{a}}^G(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}) - \mathbf{a}_t^G. \tag{16}$$

Adding up the acceleration error for all T frames and N sensors defines the motion acceleration consistency E_{acc}:

$$E_{\text{acc}} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{n=1}^{N} ||\mathbf{e}_{\text{acc},n}(t)||^2. \tag{17}$$

4.3. The Anthropometric Term

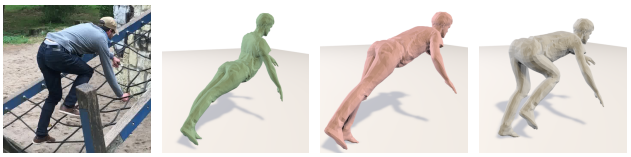
In order to constrain the skeletal joint states to human-like poses we use a multivariate Gaussian distribution of model poses with a mean pose $\mu_{\mathbf{x}}$ and covariance matrix $\Sigma_{\mathbf{x}}$ learned from the scan registrations of SMPL. While this encodes anthropometric constraints it is not motion specific as it is learned from a variety of static poses. Note that this is much less restrictive than learning based or database retrieval based approaches. We use the Mahalanobis distance that measures the likelihood of a pose $\mathbf{x}$ given the distribution $\mathcal{N}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$:

$$d_{\text{mahal}} = \sqrt{(\mathbf{x} - \mu_{\mathbf{x}})^T \Sigma_{\mathbf{x}}^{-1} (\mathbf{x} - \mu_{\mathbf{x}})}. \tag{18}$$

Additionally, we explicitly model joint limits by an error term which produces repulsive forces if a joint limit is violated. We define the joint limit error $\mathbf{e}_{\text{limit}}$ as

$$\mathbf{e}_{\text{limit}} = \min(\mathbf{x} - \mathbf{l}_{\text{lower}}, \mathbf{0}) + \max(\mathbf{x} - \mathbf{l}_{\text{upper}}, \mathbf{0}) \tag{19}$$

where $\mathbf{l}_{lower}$ and $\mathbf{l}_{upper}$ are lower and upper joint limit parameters. Altogether, the anthropometric energy term E_{antro} is a weighted



(a) (b) (c) (d)

Figure 6: *Influence of the anthropometric, orientation and acceleration consistency terms. (a) image of a climbing scene (b) using only orientation without anthropometric consistency term, (c) using orientation with anthropometric consistency term , (d) our proposed SIP using anthropometric, orientation and acceleration consistency terms.*

combination of terms

$$E_{\text{anthro}} = w_{\text{mahal}} \frac{1}{T} \sum_{t=1}^{T} d_{\text{mahal}}(t)^2 + w_{\text{limit}} \frac{1}{T} \sum_{t=1}^{T} ||\mathbf{e}_{\text{limit}}(t)||^2 \quad (20)$$

where the weighting factors w_{mahal} and w_{limit} balance the influence of the pose prior term and the joint limits term.

4.4. Energy Minimization

E_{motion} is a highly non-linear function and generally difficult to optimize. However, the exponential map formulation enables us to analytically compute gradients and since E_{motion} is composed of a sum of squared residual terms we can use the Levenberg-Marquardt algorithm. In order to compute an update-step for the Levenberg-Marquardt algorithm, we have to linearize the residual terms around the current solution with the Jacobian matrix $\mathbf{J}$. The Jacobian maps a pose increment $\delta\mathbf{x}$ to an increment of the residual according to

$$\mathbf{e}(\mathbf{x}, \delta\mathbf{x}) \approx \mathbf{e}(\mathbf{x}) + \mathbf{J}\delta\mathbf{x}. \quad (21)$$

In the following we show how to linearize the respective residual terms associated to orientation, acceleration and anthropometric consistency.

The orientation residual defined in Eq. (11) can be rewritten in terms of an incremental change of the pose $\delta\mathbf{x}$ such that

$$\mathbf{e}_{ori}(\mathbf{x}, \delta\mathbf{x}) = \log\left(\mathbf{R}(\delta\mathbf{x})\hat{\mathbf{R}}^{GS}(\mathbf{x})\left(\mathbf{R}^{GS}\right)^{-1}\right)^{\vee}, \quad (22)$$

where $\mathbf{R}(\delta\mathbf{x})$ is the rotational part of the forward kinematic map computed at the current pose $\mathbf{x}$. Using the first-order approximation for the logarithm of Eq. (4) we get a linearized expression of the orientation residual according to

$$\mathbf{e}_{ori}(\mathbf{x}, \delta\mathbf{x}) \approx \mathbf{e}_{ori}(\mathbf{x}) + \delta\mathbf{e}_{ori}(\delta\mathbf{x}). \quad (23)$$

The first term corresponds to the actual orientation residual defined in Eq. (11) and the latter term is given by

$$\delta\mathbf{e}_{ori}(\delta\mathbf{x}) = J_{ori}\delta\mathbf{x}, \quad (24)$$

where $\mathbf{J}_{ori} : \mathbb{R}^d \to \mathbb{R}^3$ is the articulated Jacobian, mapping an incremental variation of the pose vector to rotational increments in the tangent space of SO(3), see [PMR11].

In order to linearize the acceleration residual of Eq. (16), we rewrite the estimated sensor position (Eq. (15)) at a single time instance in terms of an incremental change in the pose vector $\delta\mathbf{x}$ according to

$$\bar{\mathbf{p}}(\mathbf{x}, \delta\mathbf{x}) = \mathbf{G}(\delta\mathbf{x})\mathbf{G}^{GB}(\mathbf{x})\bar{\mathbf{p}}(0) = \mathbf{G}(\delta\mathbf{x})\bar{\mathbf{p}}(\mathbf{x}), \quad (25)$$

where $\mathbf{G}(\delta\mathbf{x})$ is the forward kinematic map computed at the current pose $\mathbf{x}$. Using the Taylor expansion (Eq. (3)) of the exponential map of SE(3) up to the first order we get

$$\bar{\mathbf{p}}(\mathbf{x}, \delta\mathbf{x}) \approx \bar{\mathbf{p}}(\mathbf{x}) + \xi_{\delta\mathbf{x}}^{\times}\bar{\mathbf{p}}(\mathbf{x}). \quad (26)$$

The second term of the previous equation can be rewritten as

$$\xi_{\delta\mathbf{x}}^{\times}\mathbf{G}^{GB}(\mathbf{x})\bar{\mathbf{p}}(0) = \mathbf{J}_{p(\mathbf{x})}\delta\mathbf{x} \quad (27)$$

where $\mathbf{J}_{p(\mathbf{x})} : \mathbb{R}^d \to \mathbb{R}^3$ is the positional Jacobian at point $p(\mathbf{x})$, mapping an incremental variation of the pose vector to positional increments in Cartesian coordinates, see also [PMR11]. By combining the position estimates of three successive time steps we get the linearized acceleration error according to

$$\mathbf{e}_{acc}(t, \delta\mathbf{x}) \approx \mathbf{e}_{acc}(t) + \begin{bmatrix}\mathbf{J}_{p(\mathbf{x}_{t-1})} & -2\mathbf{J}_{p(\mathbf{x}_t)} & \mathbf{J}_{p(\mathbf{x}_{t+1})}\end{bmatrix} \begin{bmatrix}\delta\mathbf{x}_{t-1} \\ \delta\mathbf{x}_t \\ \delta\mathbf{x}_{t+1}\end{bmatrix}. \quad (28)$$

The residual terms related to anthropomorphic consistency defined in Eq. (18) and Eq. (19) are already linear in the pose $\mathbf{x}$. For the Mahalanobis prior we compute the Cholesky factorization of the inverse covariance matrix

$$\Sigma_{\mathbf{x}}^{-1} = \mathbf{L}^T\mathbf{L} \quad (29)$$

and rewrite the squared Mahalanobis distance as

$$d_{\text{mahal}}^2 = (\mathbf{x} - \mu_{\mathbf{x}})^T\mathbf{L}^T\mathbf{L}(\mathbf{x} - \mu_{\mathbf{x}}) = \mathbf{e}_{\text{mahal}}^T\mathbf{e}_{\text{mahal}}. \quad (30)$$

Then it becomes obvious that $\mathbf{e}_{\text{mahal}} : \mathbf{x} \to L(\mathbf{x} - \mu_{\mathbf{x}})$ is a linear mapping with $\mathbf{J}_{\text{mahal}} = \mathbf{L}$.

In order to compute a descent update step to minimize E_{motion}, we can now simply stack the linearized residual terms for all frames. For orientation and anthropometric terms this leads to sparse equations with the following block-diagonal structure

$$\begin{bmatrix}\ddots & & & & \\ & \mathbf{J}_{t-1} & & & \\ & & \mathbf{J}_t & & \\ & & & \mathbf{J}_{t+1} & \\ & & & & \ddots\end{bmatrix}\begin{bmatrix}\vdots \\ \delta\mathbf{x}_{t-1} \\ \delta\mathbf{x}_t \\ \delta\mathbf{x}_{t+1} \\ \vdots\end{bmatrix} = \begin{bmatrix}\vdots \\ \mathbf{e}(t-1) \\ \mathbf{e}(t) \\ \mathbf{e}(t+1) \\ \vdots\end{bmatrix}, \quad (31)$$

where $\mathbf{J}_t$ denotes the respective Jacobian of the residual term $\mathbf{e}(t)$ at time step t. Similarly, the linearized residual terms of the acceleration residuals can be combined to obtain

$$\begin{bmatrix}\ddots & \ddots & & & \\ \ddots & -2\mathbf{J}_{t-1} & \mathbf{J}_t & & \\ & \mathbf{J}_{t-1} & -2\mathbf{J}_t & \mathbf{J}_{t+1} & \\ & & \mathbf{J}_t & -2\mathbf{J}_{t+1} & \ddots \\ & & & \ddots & \ddots\end{bmatrix}\begin{bmatrix}\vdots \\ \delta\mathbf{x}_{t-1} \\ \delta\mathbf{x}_t \\ \delta\mathbf{x}_{t+1} \\ \vdots\end{bmatrix} = \begin{bmatrix}\vdots \\ \mathbf{e}_{\text{acc}}(t-1) \\ \mathbf{e}_{\text{acc}}(t) \\ \mathbf{e}_{\text{acc}}(t+1) \\ \vdots\end{bmatrix}. \quad (32)$$

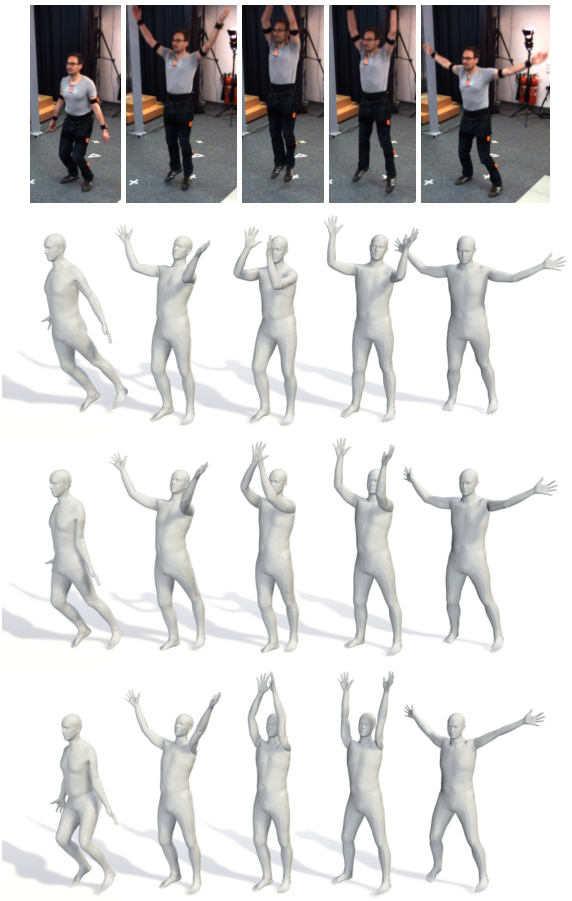


Figure 7: *We show three iterations of the optimization of E_{motion} for a jumping jack sequence. First row: images of the scene, second row: pose initialization obtained by minimizing orientation and anthropometric consistency, third row: intermediate iteration, forth row: result of SIP, i.e. final pose estimates after convergence.*

By stacking the respective linearized multi-frame residual terms, we can now simply solve for the parameter updates and iterate until convergence. Iteration results for a jumping jack sequence are illustrated in Figure 7.

4.5. IMU placement

Our proposed Sparse Inertial Poser is capable of recovering human motion from only 6 IMUs strapped to the lower legs, the lower arms, waist and head, see Figure 3(b). We found that this sensor configuration constrains a large number of pose parameters and produces good quantitative and qualitative results (see the supplemental video). An alternative sensor configuration would be to move the lower-leg and lower-arm IMUs to the end-effectors, i.e. feet and hands. Theoretically, this would constraint all joint parameters of the human body. However, we found that this adds too much uncertainty along the kinematic chain structure and results in worse performance than the proposed sensor placement.

5. Experiments

We evaluate here the performance of SIP. In Section 5.1 we present details on the general tracking procedure and computation times. Section 5.2 introduces two baseline trackers which we use to compare and evaluate the tracking performance. We provide a quantitative assessment on a publicly available data set in Section 5.3 and present qualitative results on additional recordings in Section 5.4. We refer to the video for more results.

5.1. Tracker Setup

In order to reconstruct the full-body motion with our proposed SIP we require

- A SMPL body model of the actor,
- The initial pose at the beginning of the sequence
- IMU sensor locations on the body.

Initial pose and sensor locations are required to determine the sensor to bone offsets $\mathbf{G}^{BS}$, see Section 3.3. Since IMUs are attached to different locations on the body, we manually selected the SMPL vertices once, and use them as sensor locations for all actors and experiments. Initial poses for the quantitative assessment were provided by the TNT15 data set. For the outdoor recordings we simply asked the actor to pose upright with straight arms and legs at the beginning of each sequence. We obtained SMPL body models by fitting the SMPL template to laser scans. If laser scans are not available we can also run SIP with approximate body models estimated with the method of "bodies from words" [SQRH*16]. In this case shape is estimated from only height, weight and 15 user ratings of the actor body shape.

The general tracking procedure then works as follows. Starting with the initial pose we optimize pose for every frame sequentially using the orientation and anthropometric terms. We call this method Sparse Orientation Poser (SOP) and we use it as a baseline later. The resultant pose trajectory from SOP serves as initialization for optimizing the full cost function defined in Eq. (9). As can be seen in Figure 7, optimizing orientation and anthropometric consistency terms already recovers the pose reasonably well. This step is important, since Eq. (9) is highly non-linear and we apply a local, gradient-based optimization approach. After initialization, we use a standard Levenberg-Marquardt algorithm to optimize the full cost cost function and iterate until convergence.

For all experiments, we use the same energy weighting parameters listed in Table 1, which have been determined empirically. The overall processing time for a 1000 frame sequence and 20 cost function evaluations on a quad-core Intel Core i7 3.5GHz CPU is 7.5 minutes using single-core, non-optimized MATLAB code. For each iteration the majority of time is spent on updating the body model (14.4s) and setting up the Jacobians (3.3s), while solving the sparse equations for a Levenberg-Marquardt update step takes approximately 1.5s. Parallelization of model updates and Jacobian entries on the GPU would drastically reduce computation time and we leave it as future work.

5.2. Baseline Trackers

We compare our tracking results to two baseline methods:

w_{ori}	w_{acc}	w_{anthro}	w_{mahal}	w_{limits}
1	0.05	1	0.003	0.1

Table 1: *Weighting parameters of E_{motion}, which have been used for all experiments.*

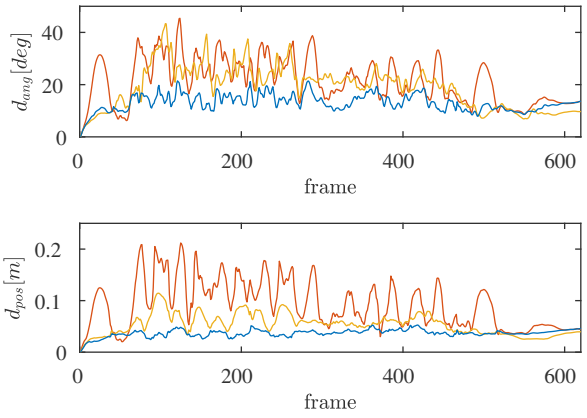


Figure 8: *Mean orientation and position error of a jumping jack sequence of the TNT15 data set. Our proposed SIP (blue) clearly outperforms both baseline trackers SOP (red) and SIP-M (yellow).*

- *Sparse Orientation Poser* (SOP): Minimizes orientation and anthropomorphic consistency terms but disregards acceleration.
- *SIP using an alternative body model* (SIP-M): Identical to SIP, but uses a manually rigged body model.

The estimated pose trajectory obtained by SOP is used as the initialization of our proposed SIP. The second baseline, the SIP-M, uses a body model provided along the TNT15 data set as depicted in Figure 2(b). It is a body model with manually placed joints and fewer pose parameters. Anatomical constraints are imposed by using hinge joints, e.g. for the knee. In total, the body model has 31 pose parameters and the manual rigging procedure is representative for models that have been used for tracking so far (e.g. [VBMP08,PMBG*11,MPMR16,GSDA*09]). In contrast, the SMPL model of SIP uses a statistical model to estimate joint positions. Every joint has 3 DoFs and anatomical constraints are imposed with the covariance of joint parameters. By comparing SIP and SIP-M we want to asses the significance of using a statistically learned body model in contrast to a typical hand-rigged one.

We also experimented with a single-frame acceleration tracker, which combines the SOP approach with acceleration data using a Kalman filter (similarly as in [VAV*07, RLS07] but with only 6 sensors). Unfortunately, only 6 IMUs do not provide sufficient constraints on the poses to prevent drift caused by acceleration. In all cases, the tracker got unstable and failed after a few frames.

5.3. Quantitative Results

For a quantitative analysis we evaluate the tracking performance of our proposed Sparse Inertial Poser (SIP) against the baseline trackers on the publicly available TNT15 data set published along [MPMR16]. This data set contains recordings of four subjects performing five activities each and provides inertial sensor data of 10 IMUs attached to lower legs, thighs, lower arms, upper arms, waist and chest. Additionally, multi-view video is provided which we only use for visualization purposes. Similar to [MPMR16] we split the 10 IMUs into tracking and validation sets. IMUs attached to lower legs, lower arms, waist and chest are used for tracking and the other IMUs serve as validation sensors.

In order to evaluate the tracking performance we define two error metrics. On the one hand we use the mean orientation error d_{ori} of the $N_v = 4$ validation IMUs

$$d_{ori} = \frac{1}{TN_v} \sum_{t=1}^{T} \sum_{n=1}^{N_v} ||\mathbf{e}_{ori,n}(t)||^2, \tag{33}$$

where $\mathbf{e}_{ori,n}$ is defined in Eq. (11) and T is the number of frames of the respective sequence. Second we compare the mean position error d_{pos} of $N_m = 13$ virtual markers on the body model

$$d_{pos} = \frac{1}{TN_m} \sum_{t=1}^{T} \sum_{n=1}^{N_m} ||\mathbf{p}_n(t) - \hat{\mathbf{p}}_n(t)||^2 \tag{34}$$

where $\mathbf{p}$ is considered as ground-truth marker position obtained by tracking with all 10 IMUs and $\hat{\mathbf{p}}$ is the estimated marker position based on the estimated poses. The virtual marker positions comprise the SMPL-model joint locations of hips, knees, ankles, shoulders, elbows, wrists and neck. Since we cannot obtain stable ground-truth global translation from 10 IMUs alone, we set it to zero for calculating d_{pos}.

The mean position error is a common metric in video-based human motion tracking benchmarks (e.g. HumanEva [SBB10], Human3.6M [IPOS14]) and is partially complementary to the mean orientation error. While the joint locations might be perfect, a rotation about a bone's axis does not alter the position error. This is only visible in the orientation error. On the other hand, a vanishing orientation error of the 4 validation IMUs does not necessarily imply correct joint positions as the spine or end-effectors might be incorrectly oriented. Hence, tracking performance is considered good if both error metrics are small.

Figure 8 shows the tracking errors for a jumping jack sequence of the TNT15 data set. This sequence contains extended arm and leg motions, also visible in Figure 7, as well as two foot stamps around frames 25 and 500. The SOP fails to accurately reconstruct these motions as orientation measurements of 6 IMUs are too ambiguous. This is easily illustrated for the case of a foot stamp, which can be seen in the second column of Figure 12. During this motion the lower leg is tilted, but without acceleration data it is impossible to infer whether the thigh was lifted at the same time. The SIP-M can resolve this ambiguity but the limited body model is not sufficiently expressive to accurately reconstruct the jumping jacks and skiing exercises. In contrast our proposed SIP shows low orientation and position errors for the whole sequence and clearly outperforms both baseline trackers.

The tracking result of the jumping jack sequence is exemplary for the overall tracking performances on the TNT15 data set. In Figure 9 we show the average orientation error for all actors, separated by activities. Similarly, Figure 10 shows the mean position

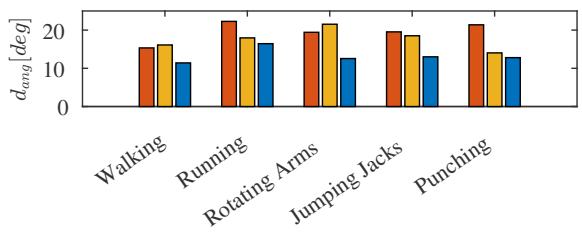


Figure 9: *Mean orientation error on the TNT15 data set: comparison of SOP(red), SIP-M(yellow) against our proposed SIP (blue).*

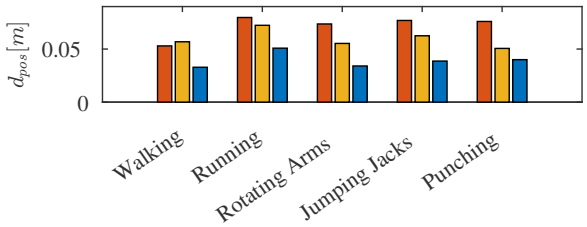


Figure 10: *Mean position error on the TNT15 data set: comparison of SOP(red), SIP-M(yellow) against our proposed SIP (blue).*

Approach	$\mu_{ang}[deg]$	$\sigma_{ang}[deg]$	$\mu_{pos}[m]$	$\sigma_{pos}[m]$
SOP	19.64	17.35	0.072	0.089
SIP-M	18.24	15.82	0.06	0.053
SIP	**13.32**	**10.13**	**0.039**	**0.04**
SIP-BW	13.45	9.94	0.042	0.04
SIP-110	13.67	10.38	0.046	0.045
SIP-120	14.27	10.6	0.056	0.053

Table 2: *Tracking errors on TNT15.*

Figure 11: *SIP is capable of recovering the handwriting on a whiteboard. Left figure: image of the writing scene, middle figure: recovered pose at the end of the handwriting, right figure: recovered wrist motion projected on the whiteboard plane.*

error. Additionally, Table 2 shows the overall tracking errors on the TNT15 data set. We have added additional rows for SIP-BW, SIP-110 and SIP-120. SIP-BW is identical to SIP but uses a SMPL model estimated with the "bodies from words" approach. The tracking error difference is insignificant, which further improves applicability of SIP. Thus, we do not need the accuracy of a laser scan, making the proposed solution very easy to use. SIP-110 and SIP-120 use a scaled version of the SIP body model, where body size was increased by 10% and 20% respectively. Again, the the tracking error remains comparably small and it further demonstrates that SIP is very robust to moderate variations in body shape.

It is remarkable, that SIP-M and SIP achieve a mean orientation error of 18.24° and 13.32°, respectively. [MPMR16] reported an average orientation error of 15.71°, using 5 IMUs and 8 cameras minimizing single-frame orientation and silhouette consistency terms. SIP-M uses the same body model and is just slightly worse. Using the SMPL body model in SIP results in an even smaller orientation error. Thus, without relying on visual cues of 8 cameras we achieve competitive orientation errors by simply taking IMU accelerations into account and optimizing over all frames simultaneously.

Quantitative results demonstrate that accurate full-body motion tracking with sparse IMU data becomes feasible by incorporating acceleration data. In comparison to the SOP which uses only orientation data, our proposed SIP reduces the mean orientation error on the TNT15 data set from 19.64° to 13.32° and the mean position error decreases from 7.2*cm* to 3.9*cm*. We have also shown that for our tracking approach, the statistically learned body model SMPL leads to more accurate tracking results than using a representative manually rigged body model. Further, the SMPL model can be even created using only linguistic ratings, which obviates the need for a laser scan of the person. In Figure 12 we show several example frames of the tracking results obtained on the TNT15 data set.

5.4. Qualitative Results

In order to further demonstrate the capabilities of our proposed SIP we recorded additional motions. For all recordings we have used 6 Xsens MTw IMUs [Xse] attached to the lower legs, wrists, head and back. The sensor placement is illustrated in Figure 2(b). Orientation and acceleration data were recorded at 60Hz and transmitted wirelessly to a laptop. Additionally, we have captured the motions with a smartphone camera to qualitatively assess the tracking accuracy.

In Figure 13 we show several tracking results for challenging outdoor motions, such as jumping over a wall, warming exercises, biking and climbing. For all cases, our proposed SIP approach is able to successfully track the overall motion. For most of the cases, the recovered poses are visually accurate using only 6 IMUs. Finally, in Figure 11 we demonstrate that SIP is capable of reconstructing the handwriting on a whiteboard. For this experiment, we attached IMUs to the lower legs, wrists, back and chest and recorded IMU data while the actor was writing “Eurographics” on a white board. The resulting wrist motion clearly resembles the hand writing.

6. Conclusions and Future Work

SIP provides a new method for estimating the pose from sparse inertial sensors. SIP makes this possible by exploiting a statistical body model and jointly optimizing pose over multiple frames to fit both orientation and acceleration data. We further demonstrate that the approach works even with approximate body models obtained from a few body word ratings. Quantitative evaluation shows that SIP can accurately reconstruct human pose accurately, with orientation errors of 13.32 degrees and positional errors of 3.9 cm.

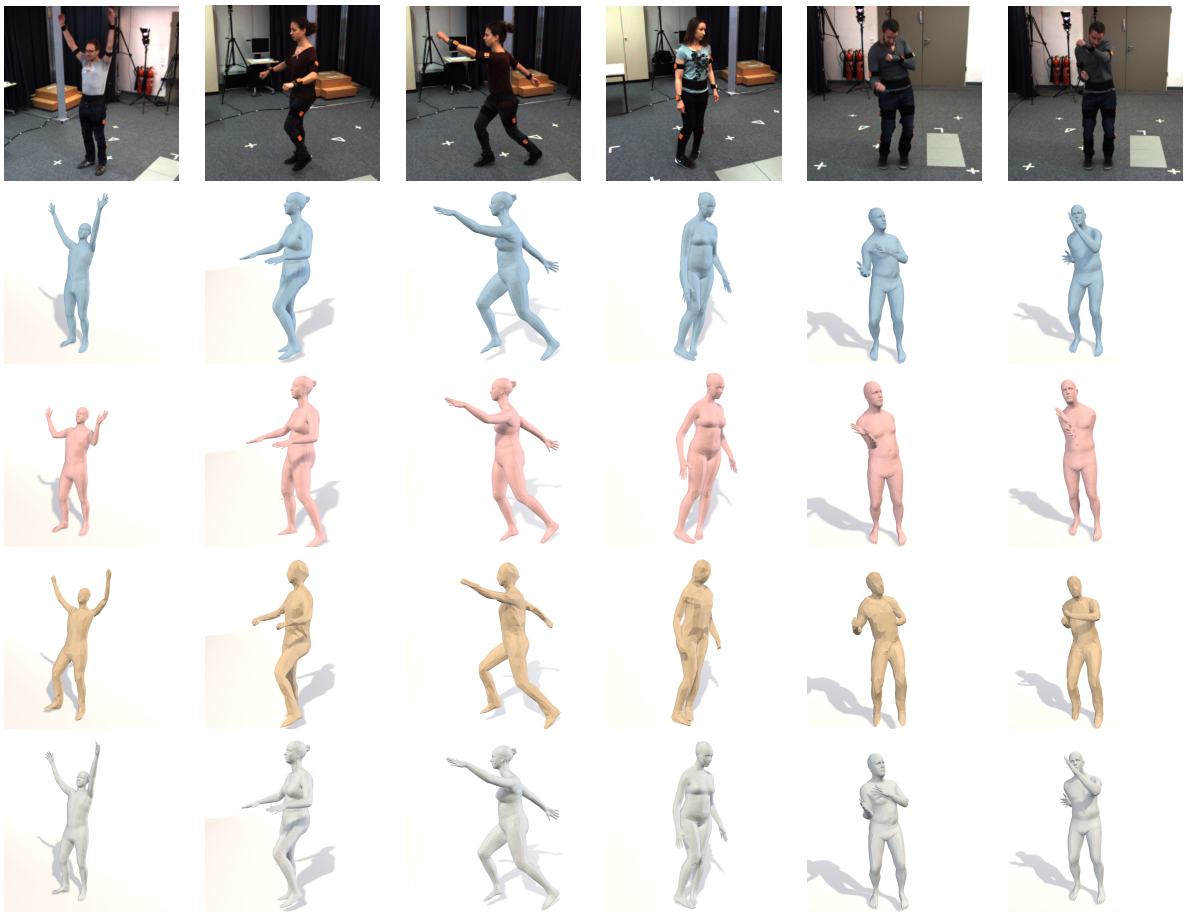


Figure 12: *We compare our proposed SIP to ground truth and two baselines, the Sparse Orientation Poser (SOP), and our SIP with a manually rigged body model (SIP-M). Top row: images from the TNT dataset sequences, second row: ground truth poses obtained by tracking with 10 IMUs (for reference), third row: results obtained with SOP, fourth row: results obtained with SIP-M and fifth row: results obtained with SIP. Best results are obtained with SIP. Without acceleration the pose remains ambiguous for the orientation poser (SOP) and leads to incorrect estimates, the SIP-M can disambiguate the poses by incorporating acceleration data but suffers from a limited skeletal model, which prevents the pose from appropriately fitting to the sensor data. Differences are best seen in the supplemental video.*

This technology opens up many directions for future research. While SIP is able to track the full-body pose without drift, global position estimates still suffer from drift over time. To that end, we plan to integrate simple physical constraints into the optimisation such as centre of mass preservation and ground contacts. Exploiting laws of conservation of energies is very involved whereas modeling ground contacts is comparably easier: ground contacts produce high peaks in the accelerometer signal which are easy to detect. Temporally fixing the position of body model points is straightforward to integrate in the proposed cost function and will compensate drift. However, modeling ground contacts depends on the motion to be tracked and assumes static friction [AHK*16]. Other options we will explore to compensate drift are integrating GPS measurements (e.g. from a cell carried phone on the body), or visual data from a body mounted camera [RRC*16, SPS*11].

Our current solution can not accurately capture wrist and ankle joint parameters due to the IMU placement on the body, see Figure 3(b) and Section 4.5. While these unobserved parameters are also optimized within the anthropometric prior, we plan to incorporate constraints derived from the 3D world geometry. Also, instead of using static joint limits in the anthropometric term one could also incorporate pose-conditioned joint angle limits [AB15] to obtain physically plausible poses. We further plan to learn human motion models from captured data in every day situations.

Finally, we would like to extend SIP to capture not only articulated motion but also soft-tissue motion by leveraging models of human shape in motion such as [PMRMB15]. SIP provides the technology to capture human motion with as few as 6 IMUs which is much less intrusive than existing technologies. There are many potential applications for this such as virtual reality, sports analysis, monitoring for health assessment, or recording of movement for psychological and social studies.

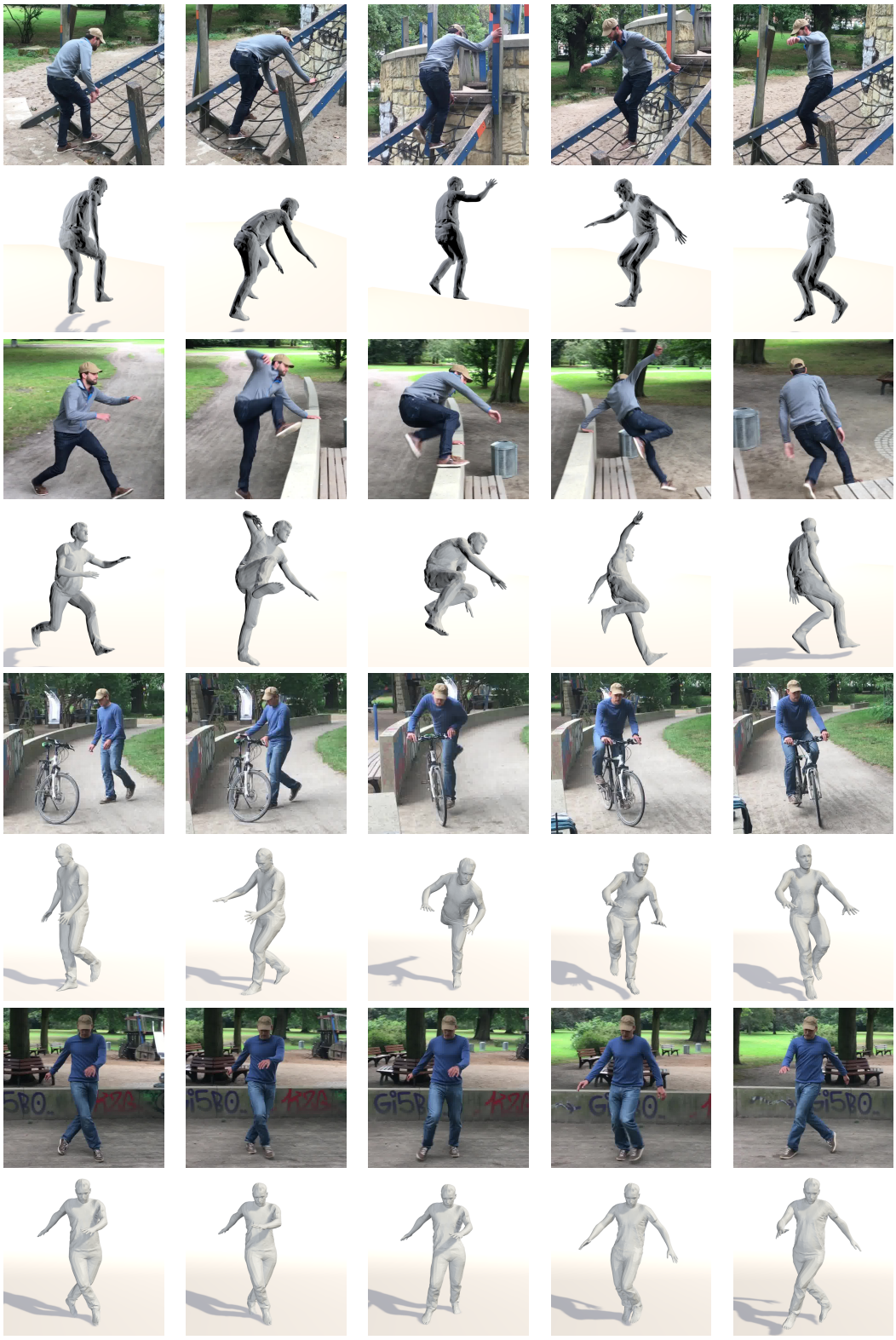


Figure 13: *We show several results obtained using SIP: For most of the cases SIP successfully recovers the full human pose. This will enable to capture people performing everyday activities in a minimally intrusive way. Results are best seen in the supplemental video.*

Acknowledgments. This work is partly funded by the DFG-Project RO 2497/11-1. Authors gratefully acknowledge the support. We thank Timo Bolkart, Laura Sevilla, Sergi Pujades, Naureen Mahmood, Melanie Feldhofer and Osman Ulusoy for proofreading, Bastian Wandt and Aron Sommer for help with motion recordings, Talha Zaman for voice recordings, Alejandra Quiros for providing the bodies from words and Senya Polikovsky, Andrea Keller and Jorge Marquez for technical support.

References

[AB15] AKHTER I., BLACK M. J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition* (2015), pp. 1446–1455. 10

[AHK*16] ANDREWS S., HUERTA I., KOMURA T., SIGAL L., MITCHELL K.: Real-time physics-based motion capture with sparse sensors. In *Proceedings of the 13th European Conference on Visual Media Production (CVMP 2016)* (2016), ACM, p. 5. 3, 10

[BHM*10] BAAK A., HELTEN T., MÜLLER M., PONS-MOLL G., ROSENHAHN B., SEIDEL H.-P.: Analyzing and evaluating markerless motion tracking using inertial sensors. In *European Conference on Computer Vision* (2010), Springer, pp. 139–152. 4

[BKL*16] 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 *Computer Vision – ECCV 2016* (Oct. 2016), Lecture Notes in Computer Science, Springer International Publishing. 2

[CH05] CHAI J., HODGINS J. K.: Performance animation from low-dimensional control signals. In *ACM Transactions on Graphics (TOG)* (2005), vol. 24, ACM, pp. 686–696. 2

[CMU] CMU motion capture database. http://mocap.cs.cmu.edu/. 1

[GSDA*09] GALL J., STOLL C., DE AGUIAR E., THEOBALT C., ROSENHAHN B., SEIDEL H.-P.: Motion capture using joint skeleton tracking and surface estimation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009* (2009), pp. 1746–1753. 8

[HKP*16] HANNINK J., KAUTZ T., PASLUOSTA C., GASSMANN K.-G., KLUCKEN J., ESKOFIER B.: Sensor-based gait parameter extraction with deep convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics* (2016). 2

[HMST13] HELTEN T., MULLER M., SEIDEL H.-P., THEOBALT C.: Real-time body tracking with one depth camera and inertial sensors. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 1105–1112. 3

[HTDL13] HARTLEY R., TRUMPF J., DAI Y., LI H.: Rotation averaging. *International Journal of Computer Vision 103*, 3 (2013), 267–305. 5

[IPOS14] 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 36*, 7 (jul 2014), 1325–1339. 8

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 34*, 6 (Oct. 2015), 248:1–248:16. 2, 3

[LWC*11] LIU H., WEI X., CHAI J., HA I., RHEE T.: Realtime human motion control with a small number of inertial sensors. In *Symposium on Interactive 3D Graphics and Games* (2011), ACM, pp. 133–140. 2

[Mix] Mixamo. http://www.mixamo.com/. 1

[MLSS94] MURRAY R. M., LI Z., SASTRY S. S., SASTRY S. S.: *A mathematical introduction to robotic manipulation*. CRC press, 1994. 3

[Mov] House of moves. http://moves.com/. 1

[MPMR16] MARCARD T. V., PONS-MOLL G., ROSENHAHN B.: Human pose estimation from video and IMUs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 38*, 8 (aug 2016), 1533–1547. 2, 3, 5, 8, 9

[PMBG*11] PONS-MOLL G., BAAK A., GALL J., LEAL-TAIXE L., MULLER M., SEIDEL H., ROSENHAHN B.: Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. pp. 1243–1250. 3, 8

[PMR09] PONS-MOLL G., ROSENHAHN B.: Ball joints for marker-less human motion capture. In *Applications of Computer Vision (WACV), 2009 Workshop on* (2009), IEEE, pp. 1–8. 3

[PMR11] PONS-MOLL G., ROSENHAHN B.: *Model-Based Pose Estimation*. Springer, 2011, ch. 9, pp. 139–170. 3, 6

[PMRMB15] PONS-MOLL G., ROMERO J., MAHMOOD N., BLACK M. J.: Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH) 34*, 4 (2015), 120. 10

[RLS07] ROETENBERG D., LUINGE H., SLYCKE P.: Moven: Full 6dof human motion tracking using miniature inertial sensors. *Xsen Technologies, December* (2007). 2, 3, 8

[RRC*16] RHODIN H., RICHARDT C., CASAS D., INSAFUTDINOV E., SHAFIEI M., SEIDEL H.-P., SCHIELE B., THEOBALT C.: EgoCap: egocentric marker-less motion capture with two fisheye cameras. 162. 10

[SBB10] SIGAL L., BALAN A., BLACK M.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal on Computer Vision (IJCV) 87*, 1 (2010), 4–27. 8

[SH08] SLYPER R., HODGINS J.: Action capture with accelerometers. In *ACM SIGGRAPH/Eurographics, SCA* (2008). 2

[Sim] Simi Reality Motion Systems. http://www.simi.com. 1

[SMN09] SCHWARZ L., MATEUS D., NAVAB N.: Discriminative human full-body pose estimation from wearable inertial sensor data. *Modelling the Physiological Human* (2009), 159–172. 2

[SPS*11] SHIRATORI T., PARK H. S., SIGAL L., SHEIKH Y., HODGINS J. K.: Motion capture from body-mounted cameras. In *ACM Transactions on Graphics (TOG)* (2011), vol. 30, ACM, p. 31. 10

[SQRH*16] STREUBER S., QUIROS-RAMIREZ M. A., HILL M. Q., HAHN C. A., ZUFFI S., O'TOOLE A., BLACK M. J.: Body Talk: Crowdshaping realistic 3D avatars with words. *ACM Trans. Graph. (Proc. SIGGRAPH) 35*, 4 (July 2016), 54:1–54:14. 7

[TBC*16] TAYLOR J., BORDEAUX L., CASHMAN T., CORISH B., KESKIN C., SHARP T., SOTO E., SWEENEY D., VALENTIN J., LUFF B., ET AL.: Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics (TOG) 35*, 4 (2016), 143. 2

[TST*15] TAGLIASACCHI A., SCHRÖDER M., TKACH A., BOUAZIZ S., BOTSCH M., PAULY M.: Robust articulated-ICP for real-time hand tracking. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library, pp. 101–114. 2

[TZK*11] TAUTGES J., ZINKE A., KRÜGER B., BAUMANN J., WEBER A., HELTEN T., MÜLLER M., SEIDEL H.-P., EBERHARDT B.: Motion reconstruction using sparse accelerometer data. *ACM Transactions on Graphics (TOG) 30*, 3 (2011), 18. 2

[VAV*07] VLASIC D., ADELSBERGER R., VANNUCCI G., BARNWELL J., GROSS M., MATUSIK W., POPOVIĆ J.: Practical motion capture in everyday surroundings. vol. 26, ACM, p. 35. 2, 3, 8

[VBMP08] VLASIC D., BARAN I., MATUSIK W., POPOVIĆ J.: Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics (TOG)* (2008), vol. 27, ACM, p. 97. 8

[Vic] Vicon. http://www.vicon.com. 1

[Xse] XSens. https://www.xsens.com/products/. 2, 9