Outdoor Human Motion Capture using Inverse Kinematics and von Mises-Fisher Sampling

Gerard Pons-Moll[1,*] Andreas Baak[2] Juergen Gall[3] Laura Leal-Taixé[1]
Meinard Müller[2] Hans-Peter Seidel[2] Bodo Rosenhahn[1]
[1]Leibniz University Hannover, Germany [2]Saarland University & MPI Informatik, Germany [3]BIWI, ETH Zurich

Abstract

Human motion capturing (HMC) from multiview image sequences is an extremely difficult problem due to depth and orientation ambiguities and the high dimensionality of the state space. In this paper, we introduce a novel hybrid HMC system that combines video input with sparse inertial sensor input. Employing an annealing particle-based optimization scheme, our idea is to use orientation cues derived from the inertial input to sample particles from the manifold of valid poses. Then, visual cues derived from the video input are used to weight these particles and to iteratively derive the final pose. As our main contribution, we propose an efficient sampling procedure where the particles are derived analytically using inverse kinematics on the orientation cues. Additionally, we introduce a novel sensor noise model to account for uncertainties based on the von Mises-Fisher distribution. Doing so, orientation constraints are naturally fulfilled and the number of needed particles can be kept very small. More generally, our method can be used to sample poses that fulfill arbitrary orientation or positional kinematic constraints. In the experiments, we show that our system can track even highly dynamic motions in an outdoor environment with changing illumination, background clutter, and shadows.

1. Introduction

Recovering 3D human motion from 2D video footage is an active field of research [19, 3, 6, 9, 28, 32]. Although extensive work on human motion capturing (HMC) from multiview image sequences has been pursued for decades, there are only few works, *e.g.* [13], that handle challenging motions in outdoor scenes.

To make tracking feasible in complex scenarios, motion priors are often learned to constrain the search space [16, 25, 26, 27, 32]. On the downside, such priors impose certain assumptions on the motions to be tracked, thus limiting the applicability of the tracker to general human motions. While approaches exist to account for transitions between different types of motion [2, 5, 10], general human motion is highly unpredictable and difficult to be modeled by prespecified action classes.

Even under the use of strong priors, video HMC is limited by current technology: depth ambiguities, occlusions, changes in illumination, as well as shadows and background clutter are frequent in outdoor scenes and make state-of-the-art algorithms break down. Using many cameras does not resolve the main difficulty in outdoor scenes, namely extracting reliable image features. Strong lighting conditions also rule out the use of depth cameras. Inertial sensors (IMU) do not suffer from such limitations but they are intrusive by nature: at least 17 units must be attached to the body which poses a problem from biomechanical studies and sports sciences. Additionally, IMU's alone fail to measure accurately translational motion and suffer from drift. Therefore, similar to [22, 30], we argue for a hybrid approach where visual cues are supplemented by orientation cues obtained by a small number of additional inertial sensors. While in [30] only arm motions are considered, the focus in [22] is on indoor motions in a studio environment where the cameras and sensors can be very accurately calibrated and the images are nearly noise- and clutter-free. By contrast, we consider full-body tracking in an outdoor setting where difficult lighting conditions, background clutter, and calibration issues pose additional challenges.

In this paper, we introduce a novel hybrid tracker that combines video input from four consumer cameras with orientation data from five inertial sensors, see Fig. 1. Within a probabilistic optimization framework, we present several contributions that enable robust tracking in challenging outdoor scenarios. Firstly, we show how the high-dimensional space of all poses can be projected to a lower-dimensional manifold that accounts for kinematic constraints induced by the orientation cues. To this end, we introduce an explicit analytic procedure based on Inverse Kinematics (IK).

[1]*Corresponding author: pons@tnt.uni-hannover.de

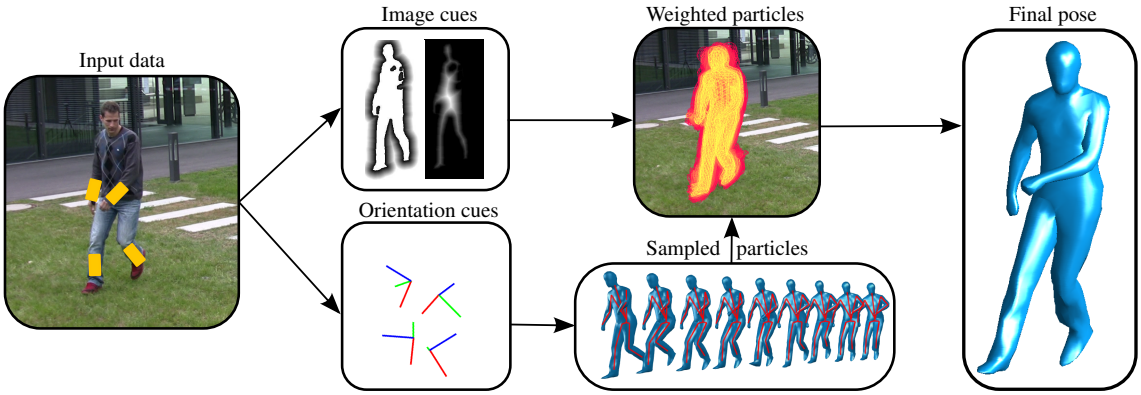

Figure 1: Orientation cues extracted from inertial sensors are used to efficiently sample valid poses using inverse kinematics. The generated samples are evaluated against image cues in a particle filter framework to yield the final pose.

Secondly, by sampling particles from this low-dimensional manifold the constraints imposed by the orientation cues are naturally fulfilled. Therefore, only a small number of particles is needed, leading to a significant improvement in efficiency. Thirdly, we show how to integrate a sensor noise model based on the von Mises-Fisher distribution in the optimization scheme to account for uncertainties in the orientation data. In the experiments, we demonstrate that our approach can track even highly dynamic motions in complex outdoor settings with changing illumination, background clutter, and shadows. We can resolve typical tracking errors such as miss-estimated orientations of limbs and swapped legs that often occur in pure video-based trackers. Moreover, we compare it with three different alternative methods to integrate orientation data. Finally, we make the challenging dataset and sample code used in this paper available for scientific use[1].

2. Related Work

For solving the high-dimensional pose optimization problem, many approaches rely on local optimization techniques [4, 13, 23], where recovery from false local minima is a major issue. Under challenging conditions, global optimization techniques based on particle filters [6, 9, 33] have proved to be more robust against ambiguities in the data. Thus, we build upon the particle-based annealing optimization scheme described in [9]. Here, one drawback is the computational complexity which constitutes a bottleneck when optimizing in high-dimensional pose spaces.

Several approaches show that constraining particles using external pose information sources can reduce ambiguities [1, 11, 12, 14, 15, 18, 29]. For example, [15] uses the known position of an object a human actor is interacting with and [1, 18] use hand detectors to constrain the pose hypothesis. To integrate such constraints into a particle-based framework, several solutions are possible. Firstly, the cost function that weights the particles can be augmented by additional terms that account for the constraints. Although robustness is added, no benefits in efficiency are achieved, since the dimensionality of the search space is not reduced. Secondly, rejection sampling, as used in [15], discards invalid particles that do not fulfill the constraints. Unfortunately, random sampling can be very inefficient and does not scale well with the number of constraints as we will show. Thirdly, approaches such as [8, 11, 17, 29] suggest to explicitly generate valid particles by solving an IK problem on detected body parts. While the proposals in [17, 29] are tailored to deal with depth ambiguities in monocular imagery, [11] relies on local optimization which is not suited for outdoor scenes as we will show. In the context of particle filters, the von Mises-Fisher distribution has been used as prior distribution for extracting white matter fiber pathways from MRI data [35].

In contrast to previous work, our method can be used to sample particles that fulfill arbitrary kinematic constraints by reducing the dimension of the state space. Furthermore, none of the existing approaches perform a probabilistic optimization in a constrained low-dimensional manifold. To the best of our knowledge, this is the first work in HMC to use IK based on the *Paden-Kahan* subproblems and to model rotation noise with the von Mises-Fisher distribution.

3. Global Optimization with Sensors

To temporally align and calibrate the input data obtained from a set of uncalibrated and unsynchronized cameras and from a set of orientation sensors, we apply preprocessing steps as explained in Sect. 3.1. Then, we define orientation data within a human motion model (Sect. 3.2) and explain the probabilistic integration of image and orientation cues into a particle-based optimization framework (Sect. 3.3).

[1] http://www.tnt.uni-hannover.de/staff/pons/

3.1. Calibration and Synchronization

We recorded several motion sequences of subjects wearing 10 inertial sensors (we used XSens [31]) which we split in two groups of 5: the *tracking sensors* which we use for tracking and the *validation sensors* which we use for evaluation. The tracking sensors are placed in the back and the lower limbs and the validation sensors are placed on the chest and the upper limbs. An inertial sensor s measures the orientation of its local coordinate system F_s^S w.r.t. a fixed global frame of reference F^T. In this paper, we refer to the sensor orientations by $\mathbf{R}^{TS}$ and, where appropriate, by using the corresponding quaternion representation $\mathbf{q}^{TS}$. The video sequences recorded with four off-the-shelf consumer cameras are synchronized by cross correlating the audio signals as proposed in [13]. Finally, we synchronize the IMU's with the cameras using a clapping motion, which can be detected in the audio data as well as in the acceleration data measured by IMU's.

3.2. Human Motion Model

We model the motion of a human by a skeletal kinematic chain containing $N = 25$ joints that are connected by rigid bones. The global position and orientation of the kinematic chain are parameterized by a twist $\xi_0 \in \mathbb{R}^6$ [20]. Together with the joint angles $\Theta := (\theta_1 \dots \theta_N)$, the configuration of the kinematic chain is fully defined by a D=6+N-dimensional vector of pose parameters $\mathbf{x} = (\xi_0, \Theta)$. We now describe the relative rigid motion matrix $\mathbf{G}_i$ that expresses the relative transformation introduced by the rotation in the i^{th} joint. A joint in the chain is modeled by a location $\mathbf{m}_i$ and a rotation axis ω_i. The exponential map of the corresponding twist $\xi_i = (-\omega_i \times \mathbf{m}_i, \omega_i)$ yields $\mathbf{G}_i$ by

$$\mathbf{G}_i = \exp(\theta_i \widehat{\xi_i}). \tag{1}$$

Let $\mathcal{J}_i \subseteq \{1, \dots, n\}$ be the ordered set of parent joint indices of the i^{th} bone. The total rigid motion $\mathbf{G}_i^{TB}$ of the bone is given by concatenating the global transformation matrix $\mathbf{G}_0 = \exp(\widehat{\xi_0})$ and the relative rigid motions matrices $\mathbf{G}_i$ along the chain by

$$\mathbf{G}_i^{TB} = \mathbf{G}_0 \prod_{j \in \mathcal{J}_i} \exp(\theta_j \widehat{\xi_j}). \tag{2}$$

The rotation part of $\mathbf{G}_i^{TB}$ is referred to as *tracking bone orientation* of the i^{th} bone. In the standard configuration of the kinematic chain, *i.e.*, the zero pose, we choose the local frames of each bone to be coincident with the global frame of reference F^T. Thus, $\mathbf{G}_i^{TB}$ also determines the orientation of the bone relative to F^T. A surface mesh of the actor is attached to the kinematic chain by assigning every vertex of the mesh to one of the bones. Let $\bar{\mathbf{p}}$ be the homogeneous coordinate of a mesh vertex $\mathbf{p}$ in the zero pose associated to the i^{th} bone. For a configuration $\mathbf{x}$ of the kinematic chain, the vertex is transformed to $\bar{\mathbf{p}}'$ using $\bar{\mathbf{p}}' = \mathbf{G}_i^{TB} \bar{\mathbf{p}}$.

3.3. Optimization Procedure

If several cues are available, *e.g.* image silhouettes and sensor orientation $\mathbf{z} = (\mathbf{z}^{im}, \mathbf{z}^{sens})$, the human pose $\mathbf{x}$ can be found by minimizing a weighted combination of cost functions for both terms as in [22]. Since in outdoor scenarios the sensors are not perfectly calibrated and the observations are noisy, fine tuning of the weighting parameters would be necessary to achieve good performance. Furthermore, the orientation information is not used to reduce the state space, and thus the optimization cost. Hence, we propose a probabilistic formulation of the optimization problem that can be solved globally and efficiently:

$$\arg\max_{\mathbf{x}} \; p(\mathbf{x}|\mathbf{z}^{im}, \mathbf{z}^{sens}). \tag{3}$$

Assuming independence between sensors and a uniform prior $p(\mathbf{x})$, the posterior can be factored into

$$p(\mathbf{x}|\mathbf{z}^{im}, \mathbf{z}^{sens}) \propto p(\mathbf{z}^{im}|\mathbf{x}) p(\mathbf{x}|\mathbf{z}^{sens}). \tag{4}$$

The weighting function $p(\mathbf{z}^{im}|\mathbf{x})$ can be modeled by any image-based likelihood function. Our proposed model of $p(\mathbf{x}|\mathbf{z}^{sens})$, as introduced in Sect. 4, integrates uncertainties in the sensor data and constrains the poses to be evaluated to a lower dimensional manifold. For optimization, we use the method proposed in [9]; the implementation details are given in Sect. 4.3.

4. Manifold Sampling

Assuming that the orientation data $\mathbf{z}^{sens}$ of the N_s orientation sensors is accurate and that each sensor has 3 DoF that are not redundant, the D dimensional pose $\mathbf{x}$ can be reconstructed from a lower dimensional vector $\mathbf{x}_a \in \mathbb{R}^d$ where $d = D - 3N_s$. In our experiments, a 31 DoF model can be represented by a 16 dimensional manifold using 5 inertial sensors as shown in Fig. 2 (a). The mapping is denoted by $\mathbf{x} = g^{-1}(\mathbf{x}_a, \mathbf{z}^{sens})$ and is described in Sect. 4.1. In this setting, Eq. (3) can be rewritten as

$$\arg\max_{\mathbf{x}_a} \; p\left(\mathbf{z}^{im}|g^{-1}(\mathbf{x}_a, \mathbf{z}^{sens})\right). \tag{5}$$

Since the orientation data $\mathbf{z}^{sens}$ is not always accurate due to sensor noise and calibration errors, we introduce a term $p(\mathbf{z}_{gt}^{sens}|\mathbf{z}^{sens})$ that models the sensor certainty, *i.e.*, the probability of the true orientation $\mathbf{z}_{gt}^{sens}$ given the sensor data $\mathbf{z}^{sens}$. The probability is described in Sect. 4.2. Hence, we get the final objective function:

$$\arg\max_{\mathbf{x}_a} \int p\left(\mathbf{z}^{im}|g^{-1}(\mathbf{x}_a, \mathbf{z}_{gt}^{sens})\right) p\left(\mathbf{z}_{gt}^{sens}|\mathbf{z}^{sens}\right) d\mathbf{z}_{gt}^{sens}. \tag{6}$$

The integral can be approximated by importance sampling, *i.e.*, drawing particles from $p(\mathbf{z}_{gt}^{sens}|\mathbf{z}^{sens})$ and weighting them by $p(\mathbf{z}^{im}|\mathbf{x})$.

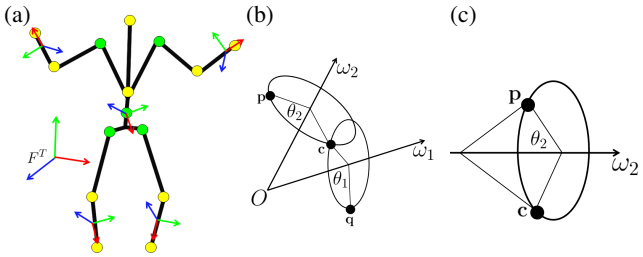

Figure 2: Inverse Kinematics: **(a)** decomposition into active (yellow) and passive (green) parameters. Paden-Kahan sub-problem 2 **(b)** and sub-problem 1 **(c)**.

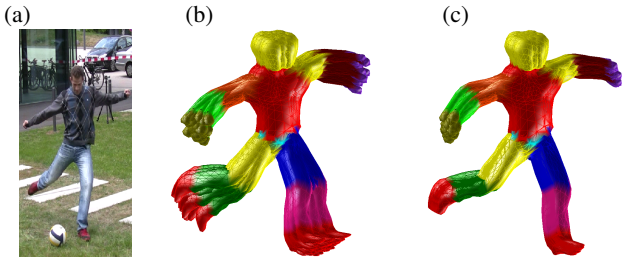

Figure 3: Manifold Sampling: **(a)** Original image. **(b)** Full space sampling. **(c)** Manifold sampling.

4.1. Inverse Kinematics using Inertial Sensors

For solving Eq. (6), we derive an analytical solution for the map $g : \mathbb{R}^D \mapsto \mathbb{R}^{D-3N_s}$ and its inverse g^{-1}. Here, g projects $\mathbf{x} \in \mathbb{R}^D$ to a lower dimensional space and its inverse function g^{-1} uses the sensor orientations and the coordinates in the lower dimensional space $\mathbf{x}_a \in \mathbb{R}^{D-3N_s}$ to reconstruct the parameters of the full pose, *i.e.*,

$$g(\mathbf{x}) = \mathbf{x}_a \qquad g^{-1}(\mathbf{x}_a, \mathbf{z}^{sens}) = \mathbf{x}. \tag{7}$$

To derive a set of minimal coordinates, we observe that given the full set of parameters $\mathbf{x}$ and the kinematic constraints placed by the sensor orientations, a subset of these parameters can be written as a function of the others. Specifically, the full set of parameters is decomposed into a set of *active parameters* $\mathbf{x}_a$ which we want to optimize according to Eq. (6) and a set of *passive parameters* $\mathbf{x}_p$ that can be derived from the constraint equations and the active set. In this way, the state can be written as $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_p)$ with $\mathbf{x}_a \in \mathbb{R}^d$ and $\mathbf{x}_p \in \mathbb{R}^{D-d}$. Thereby, the direct mapping g is trivial since from the full set only the active parameters are retained. The inverse mapping g^{-1} can be found by solving *inverse kinematics* (IK) sub-problems.

Several choices for the decomposition into active and passive set are possible. To guarantee the existence of solution for all cases, we choose the passive parameters to be the set of 3 DoF joints that lie on the kinematic branches where a sensor is placed. In our experiments using 5 sensors, we choose the passive parameters to be the two shoulder joints, the two hips and the root joint adding up to a total of 15 parameters which corresponds to $3N_s$ constraint equations, see Fig. 2 (a). Since each sensor $s \in \{1 \dots 5\}$ is rigidly attached to a bone, there exists a constant rotational offset $\mathbf{R}_s^{SB}$ between the i-th bone and the local coordinate system F_s^S of the sensor attached to it. This offset can be computed from the tracking bone orientation $\mathbf{R}_{i,0}^{TB}$ in the first frame and the sensor orientation $\mathbf{R}_{s,0}^{TS}$

$$\mathbf{R}_s^{SB} = (\mathbf{R}_{s,0}^{TS})^T \mathbf{R}_{i,0}^{TB}. \tag{8}$$

At each frame t, we obtain *sensor bone orientations* $\mathbf{R}_{s,t}^{TS}\mathbf{R}_s^{SB}$ by applying the rotational offset. In the absence of sensor noise, it is desired to enforce that the tracking bone orientation and the sensor bone orientation are equal:

$$\mathbf{R}_{i,t}^{TB} = \mathbf{R}_{s,t}^{TS}\mathbf{R}_s^{SB} \tag{9}$$

In Sect. 4.2 we show how to deal with noise in the measurements. Let $\mathbf{R}_j$ be the relative rotation of the j-th joint given by the rotational part of Eq. (1). The relative rotation $\mathbf{R}_j$ associated with the passive parameters can be isolated from Eq. (9). To this end, we expand the tracking bone orientation $\mathbf{R}_{i,t}^{TB}$ to the product of 3 relative rotations[2] $\mathbf{R}_j^p$, the total rotation motion of parent joints in the chain, $\mathbf{R}_j$, the unknown rotation of the joint associated with the passive parameters, and $\mathbf{R}_j^c$, the relative motion between the j-th joint and the i-th joint where the sensor is placed:

$$\mathbf{R}_j^p \mathbf{R}_j \mathbf{R}_j^c = \mathbf{R}_s^{TS}\mathbf{R}_s^{SB} \tag{10}$$

Note that $\mathbf{R}_j^p$ and $\mathbf{R}_j^c$ are constructed from the active set of parameters $\mathbf{x}_a$ using the product of exponentials formula (2). From Eq. (10), we obtain the relative rotation matrix

$$\mathbf{R}_j = (\mathbf{R}_j^p)^T \mathbf{R}_s^{TS}\mathbf{R}_s^{SB}(\mathbf{R}_j^c)^T. \tag{11}$$

Having $\mathbf{R}_j$ and the known fixed rotation axes $\omega_1, \omega_2, \omega_3$ of the j-th joint, the rotation angles $\theta_1, \theta_2, \theta_3$, *i.e.*, the passive parameters, must be determined such that

$$\exp(\theta_1\widehat{\omega}_1)\exp(\theta_2\widehat{\omega}_2)\exp(\theta_3\widehat{\omega}_3) = \mathbf{R}_j. \tag{12}$$

This problem can be solved by decomposing it into sub-problems [21]. The basic technique for simplification is to apply the kinematic equations to specific points. By using the property that the rotation of a point on the rotation axis is the point itself, we can pick a point $\mathbf{p}$ on the third axis ω_3 and apply it to both sides of Eq. (12) to obtain

$$\exp(\theta_1\widehat{\omega}_1)\exp(\theta_2\widehat{\omega}_2)\mathbf{p} = \mathbf{R}_j\mathbf{p} = \mathbf{q} \tag{13}$$

which is known as the *Paden-Kahan sub-problem 2*.

[2]The temporal index t is omitted for the sake of clarity

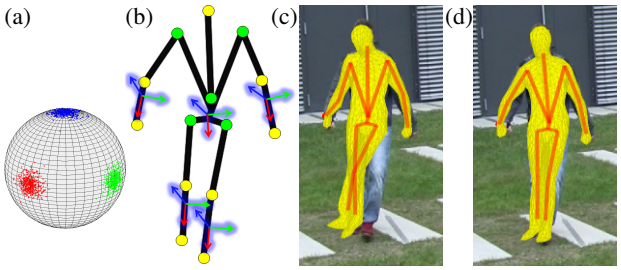

Figure 4: Sensor noise model. **(a)** Points disturbed with rotations sampled from a von Mises-Fisher distribution. **(b)** The orientation of the particles can deviate from the sensor measurements. Tracking without **(c)** and with **(d)** sensor noise model.

Eq. (13) is further decomposed into two problems

$$\exp(\theta_2\widehat{\omega}_2)\mathbf{p} = \mathbf{c} \quad \text{and} \quad \exp(-\theta_1\widehat{\omega}_1)\mathbf{q} = \mathbf{c}, \tag{14}$$

where $\mathbf{c}$ is the intersection point between the circles created by the rotating point $\mathbf{p}$ around axis ω_2 and the point $\mathbf{q}$ rotating around axis ω_1 as shown in Fig. 2 (b). Once the intersection point $\mathbf{c}$ has been calculated, the problem simplifies to finding the rotation angle about a fixed axis that brings a point $\mathbf{p}$ to a second one $\mathbf{c}$, which is known as *Paden-Kahan sub-problem 1*. Hence, the angles θ_1 and θ_2 can be easily computed from Eq. (14) using *Paden-Kahan sub-problem 1*, see Fig. 2 (c). Finally, θ_3 is obtained from Eq. (12) after substituting θ_1 and θ_2. By solving these sub-problems for every sensor, we are able to reconstruct the full state $\mathbf{x}$ using only a subset of the parameters $\mathbf{x}_a$ and the sensor measurements $\mathbf{z}^{sens}$.[3] In this way, the inverse mapping $g^{-1}(\mathbf{x}_a, \mathbf{z}_{sens}) = \mathbf{x}$ is fully defined and we can sample from the manifold, see Fig. 3.

4.2. Sensor Noise Model

In practice, perfect alignment and synchronization of inertial and video data is not possible. In fact, there are at least four sources of uncertainty in the inertial sensor measurements, namely inherent sensor noise from the device, temporal unsynchronization with the images, small alignment errors between the tracking coordinate frame F^T and the inertial frame F^I, and errors in the estimation of $\mathbf{R}_s^{SB}$. Hence, we introduce a noise model $p(\mathbf{z}_{gt}^{sens}|\mathbf{z}^{sens})$ in our objective function (6). Rotation errors are typically modeled by assuming that the measured rotations are distributed according to a Gaussian in the tangent spaces which is implemented by adding Gaussian noise v^i on the parameter components, *i.e.*, $\tilde{\mathbf{x}}_j = \mathbf{x}_j + v^i$. The topological structure of the elements, a 3-sphere S^3 in case of quaternions, is therefore ignored. The *von Mises-Fisher* distribution models errors of elements that lie on a unit sphere S^{p-1} [7] and

[3]For more details on the computation of the inverse kinematics, we refer the reader to the appendix included as supplemental material

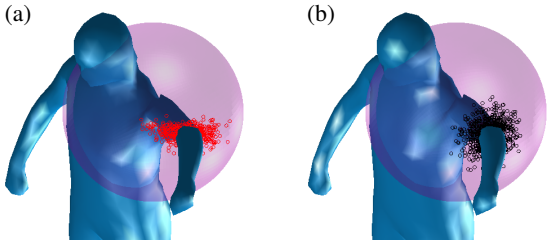

Figure 5: Sensor noise model. 500 samples of the IK elbow location are shown as points using: **(a)** added Gaussian noise and **(b)** noise from the von Mises-Fisher distribution.

is defined as

$$f_p(\mathbf{x}; \mu, \kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{d/2-1}(\kappa)} \exp(\kappa \mu^T \mathbf{x}) \tag{15}$$

where I_v denotes the modified Bessel function of the first kind, μ is the mean direction, and κ is a concentration parameter that determines the dispersion form the true position. The distribution is illustrated in Fig. 4. In order to approximate the integral in Eq. (6) by importance sampling, we use the method proposed in [34] to draw samples $\mathbf{q}_w$ from the von Mises-Fisher distribution with $p = 4$ and $\mu = (1, 0, 0, 0)^T$, which is the quaternion representation of the identity. We use a fixed dispersion parameter of $\kappa = 1000$. The sensor quaternions are then rotated by the random samples $\mathbf{q}_w$:

$$\tilde{\mathbf{q}}_s^{TS} = \mathbf{q}_s^{TS} \circ \mathbf{q}_w \tag{16}$$

where $\circ$ dennotes quaternion multiplication. In this way, for every particle, samples $\tilde{\mathbf{q}}_s^{TS}$ are drawn from $p(\mathbf{z}_{gt}^{sens}|\mathbf{z}^{sens})$ using Eq. (16) obtaining a set of distributed measurements $\tilde{\mathbf{z}}^{sens} = \left(\tilde{\mathbf{q}}_1^{TS} \dots \tilde{\mathbf{q}}_{N_s}^{TS}\right)$. Thereafter, the full pose is reconstructed from the newly computed orientations with $g^{-1}(\mathbf{x}_a, \tilde{\mathbf{z}}^{sens})$ as explained in Sect. 4.1 and weighted by $p(\mathbf{z}^{im}|\mathbf{x})$.

In Fig. 5, we compare the inverse kinematic solutions of 500 samples $i \in \{1 \dots 500\}$ by simply adding Gaussian noise *only* on the passive parameters $\{g^{-1}(\mathbf{x}_a, \mathbf{z}^{sens}) + \mathbf{v}^i\}_i$ and by modeling sensor noise with the von Mises-Fisher distribution $\{g^{-1}(\mathbf{x}_a, \tilde{\mathbf{z}}^{sens,i})\}_i$. For the generated samples, we fixed the vector of manifold coordinates $\mathbf{x}_a$ and we used equivalent dispersion parameters for both methods. To visualize the reconstructed poses we only show, for each sample, the elbow location represented as a point in the sphere. This example shows that simply adding Gaussian noise on the parameters is biased towards one direction that depends on the current pose $\mathbf{x}$. By contrast, the samples using von Mises-Fisher are uniformly distributed in all directions and the concentration decays with the angular error from the mean. Note, however, that Fig. 5 is a 3D visualization, in reality the bone orientations of the reconstructed poses should be visualized as points in a 3-sphere S^3.

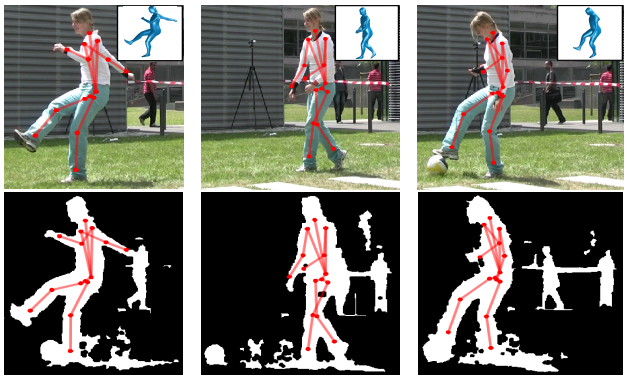

Figure 6: Tracking with background clutter.

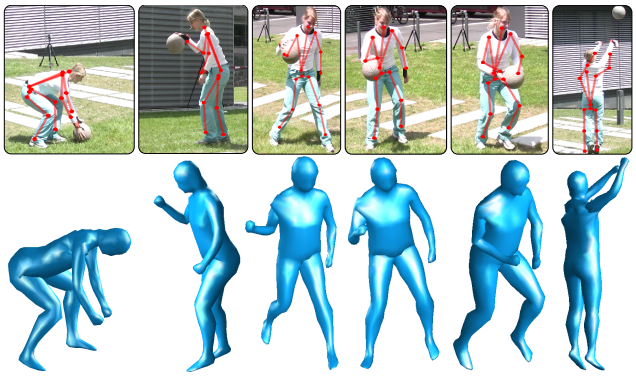

Figure 7: Tracking with strong illumination

4.3. Implementation Details

To optimize Eq. (6), we have implemented the global optimization approach that has been proposed in [9] and use only the first layer of the algorithm. As cost function, we use the silhouette and color terms

$$V(\mathbf{x}) = \lambda_1 V_{silh}(\mathbf{x}) + \lambda_2 V_{app}(\mathbf{x}) \tag{17}$$

with the setting $\lambda_1 = 2$ and $\lambda_2 = 40$. During tracking, the initial particles $\{\mathbf{x}_a^i\}_i$ are predicted from the particles in the previous frame using a 3rd order autoregression and projected to the low-dimensional manifold using the mapping g; see Sect. 4.1. The optimization is performed only over the active parameters $\mathbf{x}_a \in \mathbb{R}^{D-3N_s}$, *i.e.*, the mutation step is performed in $\mathbb{R}^{D-3N_s}$. For the weighting step, we use the approach described in Sect. 4.2 to generate a sample $\tilde{\mathbf{z}}^{sens,i}$ from $p(\mathbf{z}_{gt}^{sens}|\mathbf{z}^{sens})$ for each particle $\mathbf{x}_a^i$. Consequently, we can map each particle back to the full space using $\mathbf{x}^i = g^{-1}(\mathbf{x}_a^i, \tilde{\mathbf{z}}^{sens,i})$ and weight it by $\pi_k^i = \exp(-\beta_k V(\mathbf{x}^i))$, where β_k is the inverse temperature of the annealing scheme at iteration k. In our experiments, we used 15 iterations for optimization. Finally, the pose estimate is obtained from the remaining particle set at the last iteration as

$$\hat{\mathbf{x}}_t = \sum_i \pi_k^{(i)} \, g^{-1}(\mathbf{x}_{a,k}^{(i)}, \tilde{\mathbf{z}}^{sens,i}). \tag{18}$$

5. Experiments

The standard benchmark for human motion capture is *HumanEva* that consists of indoor sequences. However, no outdoor benchmark data comprising video as well as inertial data exists for free use yet. Therefore, we recorded eight sequences of two subjects performing four different activities, namely walking, karate, basketball and soccer. Multiview image sequences are recorded using four unsynchronized off-the-shelf video cameras. To record orientation data, we used an Xsens Xbus Kit [31] with 10 sensors. Five of the sensors, placed at the lower limbs and the back, were used for tracking, and five of the sensors, placed at the upper limbs and at the chest, were used for validation. As for any comparison measurements taken from sensors or marker-based systems, the accuracy of the validation data is not perfect but good enough to evaluate the performance of a given approach. The eight sequences in the data set comprise over 3 minutes of footage sampled at 25 Hz. Note that the sequences are significantly more difficult than the sequences of *HumanEva* since they include fast motions, illumination changes, shadows, reflections and background clutter. For the validation of the proposed method, we additionally implemented five baseline trackers: two video-based trackers based on local (L) and global optimization (G) respectively and three hybrid trackers that also integrate orientation data: local optimization (LS), global optimization (GS) and rejection sampling (RS), see [24] for more details. Let the *validation set* be the set of quaternions representing the sensor bone orientations *not* used for tracking as $\mathbf{v}^{sens} = \{\mathbf{q}_1^{val}, \ldots, \mathbf{q}_5^{val}\}$. Let $i_s, s \in \{1 \ldots t\}$ be the corresponding bone index, and $\mathbf{q}_{i_s}^{TB}$ the quaternions of the tracking bone orientation (Sect. 3.2). We define the *error measure* as the average geodesic angle between the sensor bone orientation and the tracking orientation for a sequence of T frames as

$$d_{quat} = \frac{1}{5\,T} \sum_{s=1}^{5} \sum_{t=1}^{T} \frac{180^\circ}{\pi} 2 \arccos \left| \langle \mathbf{q}_s^{val}(t), \mathbf{q}_{i_s}^{TB}(t) \rangle \right|. \tag{19}$$

We compare the performance of four different tracking algorithms using the distance measure, namely (L), (G), (LS) and our proposed approach (P). We show d_{quat} for the eight sequences and each of the four trackers in Fig. 8. For (G) and (P) we used the same number of particles $N = 200$. As it is apparent from the results, local optimization is not suitable for outdoor scenes as it gets trapped in local minima almost immediately. Our experiments show that LS as proposed in [22] works well until there is a tracking failure in

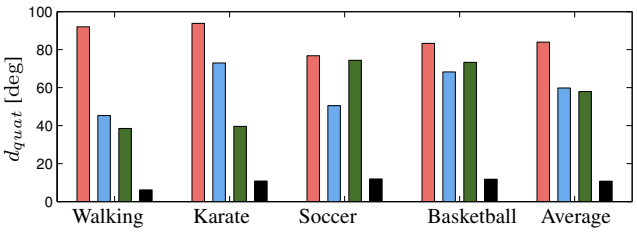

Figure 8: Mean orientation error of our 8 sequences (2 subjects) for methods (bars left to right) L (local optimization), LS (local+sensors), GL (global optimization), and ours P.

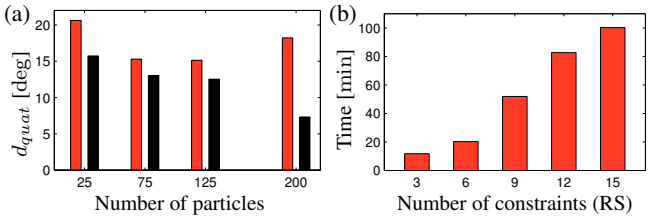

Figure 9: **(a)**: Orientation error with respect to number of particles with (red) the GS method and (black) our algorithm. **(b)**: Running time of *rejection sampling* (RS) with respect to number of constraints. By contrast our proposed method takes 0.016 seconds for 15 DoF constraints. The time to evaluate the image likelihood is excluded as it is independent of the algorithm.

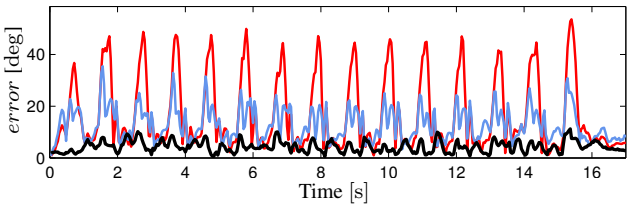

Figure 10: Angular error for the left hip of a walking motion with (red) no sensor noise model (NN), (blue) Gaussian noise model (GN) and (black) our proposed (MFN).

Figure 11: Tracking results of a soccer sequence

which case the tracker recovers only by chance. Even using (G), the results are unstable since the video-based cues are too ambiguous and the motions too fast to obtain reliable pose estimates. By contrast, our proposed tracker achieves an average error of $10.78\,^\circ \pm 8.5^\circ$ and clearly outperforms the pure video-based trackers and (LS).

In Fig. 9 (a), we show d_{quat} for a varying number of particles using the (GS) and our proposed algorithm (P) for a walking sequence. For (GS) we optimize a cost function $V(\mathbf{x}) = \mu_1 V^{im}(\mathbf{x}) + \mu_2 V^{sens}(\mathbf{x})$ where the image term $V^{im}(\mathbf{x})$ is the one defined in Eq. (17) and $V^{sens}(\mathbf{x})$ is chosen to be to be an increasing linear function of the angular error between the tracking and the sensor bone orientations. We hand tuned the influence weights μ_1, μ_2 to obtain the best possible performance. The error values show that optimizing a combined cost function leads to bigger errors for the same number of particles when compared to our method. This was an expected result since we reduce the dimension of the search space by sampling from the manifold and consequently less particles are needed for equal accuracy. Most importantly, the visual quality of the 3D animation deteriorates more rapidly with (GS) as the number of particles are reduced[4]. This is partly due to the fact that the constraints are not always satisfied when additional error terms guide the optimization. Another option for combining inertial data with video images is to draw particles directly from $p(\mathbf{x}_t|\mathbf{z}^{sens})$ using a simple rejection sampling scheme. In our implementation of (RS), we reject a particle when the angular error is bigger than 10 degrees. Unfortunately, this approach can be very inefficient especially if the manifold of poses that fulfill the constraints lies in a narrow region of the parameter space. This is illustrated in Fig. 9 (b) where we show the processing time per frame (excluding image likelihood evaluation) using 200 particles as a function of the number of constraints. Unsurprisingly, rejection sampling does not scale well with the number of constraints taking as much as 100 minutes for 15 DoF constraints imposed by the 5 sensors. By contrast, our proposed sampling method takes in the worst case (using 5 sensors) 0.016 seconds per frame. These findings show that sampling directly from the manifold of valid poses is a much more efficient alternative. To evaluate the influence of the sensor noise model, we tracked one of the walking sequences in our dataset using no noise (NN), additive Gaussian noise (GN) in the passive parameters and noise from the von Mises-Fisher (MFN) distribution as proposed in Sect. 4.2. In Fig. 10 we show the angular error of the left hip using each of the three methods. With (NN) error peaks occur when the left leg is matched with the right leg during walking, see Fig. 4. This typical example shows that slight missalignments (as little as $5^\circ - 10^\circ$) between video and sensor data can miss-guide the tracker if no noise model

[4] see the video for a comparison of the estimated motions

is used. The error measure was 26.8° with no noise model, 13° using Gaussian noise and 7.3° with the proposed model. The error is reduced by 43% with (MFN) compared to (GN) which shows that the von Mises-Fisher is a more suited distribution to explore orientation spaces than the commonly used Gaussian. This last result might be of relevance not only to model sensor noise but to any particle-based HMC approach. Finally, pose estimation results for typical sequences of our dataset are shown in Fig. 6, 7 and 11.

6. Conclusions

By combining video with IMU input, we introduced a novel particle-based hybrid tracker that enables robust 3D pose estimation of arbitrary human motions in outdoor scenarios. As the two main contributions, we first presented an analytic procedure based on inverse kinematics for efficiently sampling from the manifold of poses that fulfill orientation constraints. Secondly, robustness to uncertainties in the orientation data was achieved by introducing a sensor noise model based on the von Mises-Fisher distribution instead of the commonly used Gaussian distribution. Our experiments on diverse complex outdoor video sequences reveal major improvements in the stability and time performance compared to other state-of-the art trackers. Although in this work we focused on the integration of constraints derived from IMU, the proposed sampling scheme can be used to integrate general kinematic constraints. In future work, we plan to extend our algorithm to integrate additional constraints derived directly from the the video data such as body part detections, scene geometry or object interaction.

Acknowledgments. We give special thanks to Thomas Helten for his kind help with the recordings. This work has been supported by the German Research Foundation (DFG CL 64/5-1 and DFG MU 2686/3-1). Meinard Müller is funded by the Cluster of Excellence on Multimodal Computing and Interaction.

References

[1] P. Azad, T. Asfour, and R. Dillmann. Robust real-time stereo-based markerless human motion capture. In *Proc. 8th IEEE-RAS Int. Conf. Humanoid Robots*, 2008. 2

[2] A. Baak, B. Rosenhahn, M. Müller, and H.-P. Seidel. Stabilizing motion tracking using retrieved motion priors. In *ICCV*, 2009. 1

[3] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *CVPR*, 2007. 1

[4] C. Bregler, J. Malik, and K. Pullen. Twist based acquisition and tracking of animal and human kinematics. *IJCV*, 56(3):179–194, 2004. 2

[5] J. Chen, M. Kim, Y. Wang, and Q. Ji. Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition. In *CVPR*, pages 2655–2662. IEEE, 2009. 1

[6] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2005. 1, 2

[7] R. Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Mathematical and Physical Sciences*, 1953. 5

[8] M. Fontmarty, F. Lerasle, and P. Danes. Data fusion within a modified annealed particle filter dedicated to human motion capture. In *IRS*, 2007. 2

[9] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *IJCV*, 87:75–92, 2010. 1, 2, 3, 6

[10] J. Gall, A. Yao, and L. Van Gool. 2D action recognition serves 3D human pose estimation. In *ECCV*, pages 425–438, 2010. 1

[11] V. Ganapathi, C. Plagemann, S. Thrun, and D. Koller. Real time motion capture using a time-of-flight camera. In *CVPR*, 2010. 2

[12] D. Gavrila and L. Davis. 3D model based tracking of humans in action: a multiview approach. In *CVPR*, 1996. 2

[13] N. Hasler, B. Rosenhahn, T. Thormählen, M. Wand, J. Gall, and H.-P. Seidel. Markerless motion capture with unsynchronized moving cameras. In *CVPR*, pages 224–231, 2009. 1, 2, 3

[14] S. Hauberg, J. Lapuyade, M. Engell-Norregard, K. Erleben, and K. Steenstrup Pedersen. Three dimensional monocular human motion analysis in end-effector space. In *EMMCVPR*, 2009. 2

[15] H. Kjellstromm, D. Kragic, and M. J. Black. Tracking people interacting with objects. In *CVPR*, pages 747–754, 2010. 2

[16] C. Lee and A. Elgammal. Coupled visual and kinematic manifold models for tracking. *IJCV*, 2010. 1

[17] M. W. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *CVPR*, volume 2, 2004. 2

[18] N. Lehment, D. Arsic, M. Kaiser, and G. Rigoll. Automated pose estimation in 3D point clouds applying annealing particle filters and inverse kinematics on a gpu. In *CVPR Workshop*, 2010. 2

[19] T. Moeslund, A. Hilton, V. Krueger, and L. Sigal, editors. *Visual Analysis of Humans: Looking at People*. Springer, 2011. 1

[20] R. Murray, Z. Li, and S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Baton Rouge, 1994. 3

[21] B. Paden. *Kinematics and control of robot manipulators*. PhD thesis, 1985. 4

[22] G. Pons-Moll, A. Baak, T. Helten, M. Müller, H.-P. Seidel, and B. Rosenhahn. Multisensor-fusion for 3D full-body human motion capture. In *CVPR*, pages 663–670, 2010. 1, 3, 6

[23] G. Pons-Moll, L. Leal-Taixé, T. Truong, and B. Rosenhahn. Efficient and robust shape matching for model based human motion capture. In *DAGM*, 2011. 2

[24] G. Pons-Moll and B. Rosenhahn. Model-based pose estimation. *Visual Analysis of Humans*, pages 139–170, 2011. 6

[25] M. Salzmann and R. Urtasun. Combining discriminative and generative methods for 3d deformable surface and articulated pose reconstruction. In *CVPR*, June 2010. 1

[26] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV*, pages 750–757, 2003. 1

[27] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *ECCV*, 2000. 1

[28] L. Sigal, L. Balan, and M. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, pages 1337–1344, 2008. 1

[29] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *CVPR*, 2003. 2

[30] Y. Tao, H. Hu, and H. Zhou. Integration of vision and inertial sensors for 3D arm motion tracking in home-based rehabilitation. *IJRR*, 26(6):607, 2007. 1

[31] X. M. Technologies. http://www.xsens.com/. 3, 6

[32] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with gaussian process dynamical models. In *CVPR*, 2006. 1

[33] P. Wang and J. M. Rehg. A modular approach to the analysis and evaluation of particle filters for figure tracking. In *CVPR*, 2006. 2

[34] A. Wood. Simulation of the von mises-fisher distribution. *Communications in Statistics - Simulation and Computation*, 1994. 5

[35] F. Zhang, E. R. Hancock, C. Goodlett, and G. Gerig. Probabilistic white matter fiber tracking using particle filtering and von mises-fisher sampling. *Medical Image Analysis*, 13(1):5–18, 2009. 2