Joint 3D Human Motion Capture and Physical Analysis from Monocular Videos

Petrissa Zell, Bastian Wandt, Bodo Rosenhahn
Leibniz University Hannover
{zell,wandt,rosenhahn}@tnt.uni-hannover.de

Abstract

Motion analysis is often restricted to a laboratory setup with multiple cameras and force sensors which requires expensive equipment and knowledgeable operators. Therefore it lacks in simplicity and flexibility. We propose an algorithm combining monocular 3D pose estimation with physics-based modeling to introduce a statistical framework for fast and robust 3D motion analysis from 2D videodata. We use a factorization approach to learn 3D motion coefficients and join them with physical parameters, that describe the dynamic of a mass-spring-model. Our approach does neither require additional force measurement nor torque optimization and only uses a single camera while allowing to estimate unobservable torques in the human body. We show that our algorithm improves the monocular 3D reconstruction by enforcing plausible human motion and resolving the ambiguity of camera and object motion.

The performance is evaluated on different motions and multiple test data sets as well as on challenging outdoor sequences.

1. Introduction

Nowadays, a vast amount of labeled 2D video data of moving persons in various scenarios, e.g. sports, acting, health etc. is easily accessible [8, 17, 27]. The crucial problem is, how to effectively assess hidden information, such as force and torque characteristics, that can teach us about the efficiency and healthiness of human motion. While the two topics of 3D motion reconstruction from 2D landmarks (known as *nonrigid structure from motion*) and physical human modeling are heavily investigated separately there exists only small amount of research on the combination of both. In this paper we will show how a joint model can enhance both the 3D reconstruction and the physical modeling of human motion.

The recovery of 3D human poses in monocular image sequences is an inherently ill-posed problem, since the observed projection on a 2D plane can be explained by various 3D poses and camera positions. Most recent works reconstruct human poses from learned subspaces and define priors on the reconstructed 3D poses based on knowledge about human motion or simple physical priors [1, 23, 33, 34]. These methods achieve acceptable results but are too restrictive as they limit the solution to a predefined skeleton or violate biomechanical constraints by ignoring knowledge about the kinematics of human motion.

Figure 1. Analysis of lifting motions: Based on 2D motion data we estimate the torque in the lower back for different kinds of lifting motions. In contrast to other approaches we use the laboratory setup only for evaluation. The orange arrows represent modeled outer forces and the red areas the maximum joint torque. Lengths and radii are proportional to the force and torque magnitudes, respectively.

In this paper we use a factorization approach similar to [13, 21, 23, 34]. We assume a set of labeled joints throughout the sequence. Our goal is to decompose it into three factors, namely camera motion, base poses and mixing coefficients. We iteratively estimate the camera parameters and mixing coefficients to obtain a 3D reconstruction of the human motion. Since this 3D reconstruction violates biomechanical constraints we project it to a joint model of mixing coefficients and physical parameters to obtain a plausible human motion. Due to a physical simulation we are able to eliminate the ambiguity between object and camera motion which is inherent in all structure from motion problems.

In addition to the 3D pose estimation we analyze the reconstructed 3D motion regarding inner forces, referred to as joint torques. They are the sum of all moments effecting a joint, caused by muscles, ligaments and neighboring bone segments. Joint torques are of special interest in biomechanical studies, since they can act as a measure for the

strain at a joint. For example, the alignment of prosthetics can be characterized by investigating joint torques during gait [25]. The method of choice, used in gait analysis laboratories is to inversely calculate joint torques from ground reaction forces (GRF) [36] and joint trajectories. This poses a disadvantage, since the trajectory data and the related accelerations have a high uncertainty, in general.

Alternatively, the estimation of joint torques can be achieved via forward dynamics optimization. Here torques are implemented in the equations of motion (EOM) and estimated by solving an optimization problem, that includes the integration of EOM. This method has the advantage of directly accessible joint torques, but the required integration entails high computational cost [37]. Furthermore the choice of objective (usually some form of energy function) to be minimized is crucial for the quality of results and the convergence of the optimization algorithm requires sufficient initialization. Therefore this method lacks in stability and robustness.

In order to circumvent these issues, we apply a data-driven statistical approach. Physical model parameters are learned together with associated motion mixing coefficients resulting in a combined statistical model, that enables us to directly infer 3D information from a monocular image sequence, without the need for expensive optimization.

We will test the performance of our joint model for two motion types (walking and lifting). We focus on the analysis of gait patterns, because this basic form of movement is essential in biomechanical research [12, 25]. Additionally, an exemplary measure for healthiness of lifting motions is defined to demonstrate practicability of our proposed method for health applications (cf. Fig. 1). We will show 3D reconstructions of motion capture sequences as well as reconstructions of the challenging outdoor sequences of the KTH data base [17] and evaluate the stability of 3D reconstructions with respect to noisy and occluded input joint trajectories. Our algorithm is stable up until an occlusion of 25 % of the joints, while other physics-based methods [6, 16] fail when confronted with incomplete input information.

To the best of our knowledge this is the first approach to combine 3D reconstructions from moving uncalibrated monocular cameras with a 3D physical model in a joint framework.

Summarizing, our contributions are:

- A joint model for 3D motion reconstruction and physical analysis from moving monocular cameras. The joint model naturally solves the ambiguity of camera and object motion.
- Estimation of formerly non-observable inner and exterior forces from a set of monocular images without tedious optimization.
- Analyzing *healthiness* of lifting motions.

2. Related Work

The factorization of a set of tracked 2D features of a nonrigid object in two sets of variables describing camera motion and object motion was first proposed by Bregler et al. [4]. They model a pose in a single image as a linear combination of rigid base shapes. Since these base shapes are ambiguous [38] there are multiple works constraining the formulation of [4] with additional priors [2, 9, 30, 31].

The mentioned solutions to the nonrigid structure from motion problem create good results for benchmark data sets. However, they fail for most real world sequences where there is insufficient camera motion as shown by [2]. Therefore a number of authors propose the use of different priors on the reconstructed shapes. Common approaches for reconstruction of human motion from single images or image sequences use trained base poses and anthropometric constraints [1, 23, 33, 34]. These works show the efficiency of using information about human anthropometry. However, they do not model any forces or torques acting on the human body.

In this work we include a physical model for human motion that considers inner (joint torques) and outer forces (ground reaction forces/interaction with objects). Physics-based modeling of human motion is an established approach in the field of computer graphics for the synthesis of physically-valid movement [11, 24, 28, 35, 40]. In computer vision physical models are used to facilitate object and person tracking [5, 20, 32], since they address frequently occurring inaccuracies, such as unrealistic or instable movement. Furthermore the integration of a physical model allows for video-based motion analysis, i.e. the estimation of forces and moments, acting on the observed body [3]. Regarding a complex dynamical system, such as the human body, this problem has been sparsely examined.

Brubaker et al. [6] use an articulated mass-spring model to estimate joint torques and contact dynamics based on 3D motion capture data. The authors decouple the EOM at different frames by introducing additional root forces in order to avoid the high computational cost induced by the integration step in forward dynamics optimization. While our physical model is inspired by [6], we do not use unphysical root forces and bypass the expensive optimization by learning model parameters on a training set of motion sequences. Furthermore [6] requires knowledge about the global position and orientation of the root joint and is therefore inapplicable for moving camera scenarios.

Other works focus on the synthesis of realistic motions using physics-based modeling, e.g. to learn a subject specific description of motion styles [18] or to enforce physical constraints on statistical motion priors [35]. The authors concentrate on the generation of natural-looking movement and do not discuss the soundness of resulting force and torque profiles.

Recently researchers attempt to learn a mapping from a motion to the corresponding joint torques: Johnson et al. [16] investigate sparse coding for inverse dynamics regression and find that the resulting torque errors are unacceptably large. A related problem is treated in [39]. The authors introduce a joint statistical model for the physical analysis of gait patterns in 2D. In [19] a data-driven prior model for contact information and joint torques is constructed to reduce the ambiguity of inverse dynamics. In contrast to these approaches, we propose a 3D statistical model, consisting of motion mixing coefficients and physical parameters. On this basis, we demonstrate monocular torque estimation, that is very robust with respect to noisy and occluded input joint trajectories and requires small computation times, compared to optimization-based methods.

3. Pose Estimation

The pose estimation from 2D joint labels is based on a nonrigid structure from motion formulation, such as [4]. We assume that the input data can be decomposed in camera matrices, base poses and coefficients for these base poses. The 3D base poses are learned from 3D sequences of the same motion category as the motion we want to reconstruct. The algorithm iteratively minimizes a reprojection error to solve for the camera matrices and coefficients.

Let $\boldsymbol{W}_{2d} \in \mathbb{R}^{2f \times j}$ be the input data consisting of stacked 2D poses $\boldsymbol{P}_{1,\dots,f} \in \mathbb{R}^{2\times j}$ for f frames and j joints. Each row of $\boldsymbol{P}_i$ for $i = 1, \dots, f$ contains the 2D point coordinates of the j joints.

To achieve translational invariance each $\boldsymbol{P}_i$ is subtracted by the mean of all $\boldsymbol{P}_i$ which centers it at the origin. We assume that the input data can be factored in two terms representing camera matrices $\boldsymbol{K} \in \mathbb{R}^{2f\times 3f}$ and 3D poses $\boldsymbol{S} \in \mathbb{R}^{3f\times j}$

$$\boldsymbol{W}_{2d} = \boldsymbol{K}\boldsymbol{S}. \tag{1}$$

$\boldsymbol{S}$ is constructed in the same way as $\boldsymbol{W}_{2d}$ by stacking 3D poses $\boldsymbol{S}_i \in \mathbb{R}^{3\times j}$. We construct $\boldsymbol{S}_i$ from a linear combination of k base poses $\boldsymbol{Q}_{0,1,\dots,k} \in \mathbb{R}^{3\times j}$ similar to [4].

$$\boldsymbol{S}_i = \boldsymbol{Q}_0 + \sum_{l=1}^{k} \alpha_l \boldsymbol{Q}_l \tag{2}$$

$\boldsymbol{S}$ can now be written as

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{Q}_0 + \sum_{l=1}^{k} \alpha_{l,1}\boldsymbol{Q}_l \\ \vdots \\ \boldsymbol{Q}_0 + \sum_{l=1}^{k} \alpha_{l,f}\boldsymbol{Q}_l \end{pmatrix} = \boldsymbol{A} \begin{pmatrix} \boldsymbol{Q}_0 \\ \vdots \\ \boldsymbol{Q}_k \end{pmatrix} = \boldsymbol{A}\boldsymbol{Q}, \tag{3}$$

where

$$\boldsymbol{A} = \begin{pmatrix} 1 & \alpha_{1,1} & \cdots & \alpha_{l,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_{1,f} & \cdots & \alpha_{l,f} \end{pmatrix} \otimes \boldsymbol{I}_3 \tag{4}$$

and $\otimes$ denotes the Kronecker product. While the base poses $\boldsymbol{Q}_l$ can be learned by a principal component analysis on similar motion sequences, the only unknowns are the camera matrices $\boldsymbol{K}$ and the mixing coefficients $\boldsymbol{A}$. Approaches that iteratively estimate the camera matrices and the coefficients have proven to give good results, even on single images [23, 34, 33]. However, the number of coefficients is large. Our approach is motivated by [2] as we are representing the coefficients in a trajectory basis. In contrast to [2] we use our previously learned shape basis and iteratively solve for the coefficients and camera parameters instead of solving for the shape basis. This is done by decomposing the coefficient matrix $\boldsymbol{A}$ in the trajectory basis matrix $\boldsymbol{B} \in \mathbb{R}^{f\times b}$ and a weighting matrix $\boldsymbol{D} \in \mathbb{R}^{b\times k}$ with b as the number of trajectory bases which reduces the number of unknowns to $b \cdot k$.

$$\boldsymbol{A} = (\boldsymbol{B}\boldsymbol{D}) \otimes \boldsymbol{I}_3 \tag{5}$$

While an arbitrary set of basis functions is possible, base functions of a discrete cosine transform (DCT) have proven to give good results due to their excellent energy compaction for highly correlated data [2]. Using basis functions also enforces smoothness of the weights in $\boldsymbol{A}$, which appears to be a valid assumption considering human motion (cf. [33]).

With Eqs. (1),(3) and (5) a reprojection error can be minimized by

$$\min_{\boldsymbol{K},\boldsymbol{D}} \|\boldsymbol{W}_{2d} - \boldsymbol{K}(\boldsymbol{B}\boldsymbol{D} \otimes \boldsymbol{I}_3)\boldsymbol{Q}\|_F. \tag{6}$$

Eq. (6) is solved by iteratively optimizing for $\boldsymbol{K}$ and $\boldsymbol{D}$. Note, that both sub problems are convex and can be efficiently solved by modeling tools for specifying and solving convex programs such as [14, 15].

4. Physical Model

The physical simulation is based on a 3D mass-spring-model, consisting of 13 segments. The kinematic chain is parametrized according to Denavit-Hartenberg [29] and has 29 degrees of freedom (DOF), that constitute the mutually independent model coordinates $\boldsymbol{q}$. Six of them describe the global position and orientation of the root link, leaving 23 DOF for internal joint angles. The state of the model at a time t is defined by its link configuration and the corresponding linear and angular velocities $[\boldsymbol{q}(t), \dot{\boldsymbol{q}}(t)]^T$.

Each of the internal links is associated with a torsional spring that exerts a torque τ_j on adjacent segments, according to

$$\tau_j = -\kappa_j(q_j - q_j^{(0)}) - d_j\dot{q}_j\,, \tag{7}$$

with joint link angle q_j, stiffness κ_j, resting angle $q_j^{(0)}$ and damping constant d_j. In order to generate realistic human

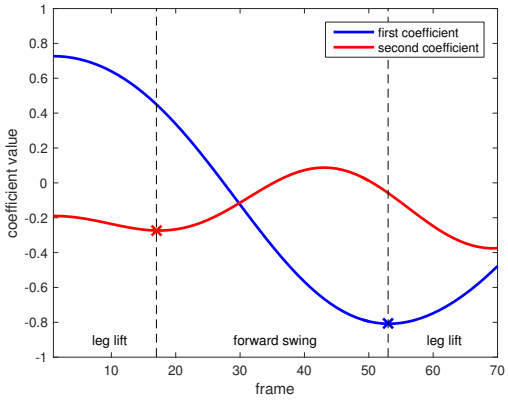

Figure 2. Detection of the phases of a walking motion using the extrema of the first and second base pose coefficient.

movement, we divide motion sequences into several phases with separate sets of spring parameters. Since one phase corresponds to a uniform motion the phase transitions can be easily estimated by detecting the extrema in the base shape coefficients as shown in Fig. 2.

The GRFs are simulated by incorporating a contact model which is inspired by [6]. Vertical reaction forces and frictional forces are implemented as very stiff damped spring forces, modulated by sigmoid functions and act on a set of contact points at the sole of the foot, when they approach the ground. One sigmoid causes the contact forces to be zero when the contact points are distant from the ground and the other prevents an acceleration towards the ground. The total contact force is defined as

$$\begin{aligned} \boldsymbol{F}_c(\boldsymbol{q},\dot{\boldsymbol{q}}) =& F_v(y_c(\boldsymbol{q}))\,\sigma(s_1 F_v(\boldsymbol{q}))\boldsymbol{e}_y \\ & - d_h \boldsymbol{J}_c\dot{\boldsymbol{q}}\,\sigma(s_2(y_0 - y_c(\boldsymbol{q}))) \\ F_v(y_c) =& (-\kappa_v(y_c - y_0) - d_v\dot{y}_c)\,\sigma(s_2(y_0 - y_c))\,, \end{aligned} \tag{8}$$

where $\sigma = (1+\exp(-x))^{-1}$. The two terms in Eq. (8) represent vertical contact force and friction, respectively. The ground plane is described by its normal vector $\boldsymbol{e}_y$ and its offset y_0. The contact force is depending on the contact point height $y_c(\boldsymbol{q})$ and the linear and rotational horizontal velocity components $\boldsymbol{J}_c\dot{\boldsymbol{q}}$, with contact Jacobian $\boldsymbol{J}_c$. The spring stiffness κ_v and attenuation constants d_v and d_h are optimization parameters, i.e. variable for every motion. The residual factors are empirically set to $[s_1, s_2] = [30, 3\cdot 10^3]$.

To simulate a motion of the physical model, equations of motion (EOM) have to be formulated and solved numerically. We derive the EOM by means of the TMT-method [26]:

$$\boldsymbol{\mathcal{M}}\ddot{\boldsymbol{q}} = \boldsymbol{\tau} + \boldsymbol{J}^T(\boldsymbol{M}(\boldsymbol{a}_g - \boldsymbol{G}) + \boldsymbol{F}_c)\,. \tag{9}$$

Here $\boldsymbol{\mathcal{M}} = \boldsymbol{J}^T\boldsymbol{M}\boldsymbol{J}$ and $\boldsymbol{M}$ denote the mass matrix in generalized and mutually dependent coordinates, respectively, with the Jacobian $\boldsymbol{J}$, describing the corresponding transformation. Relative segment masses, moments of inertia and center of mass positions are set according to anthropometric data [7]. In addition to joint torques $\boldsymbol{\tau}$ and contact force $\boldsymbol{F}_c$, we incorporate gravitational acceleration $\boldsymbol{a}_g$ and convective acceleration $\boldsymbol{G}$.

The estimation of active joint torques is done via forward dynamics optimization. In other words, we simulate a motion, using our physical model and optimize parameters $\boldsymbol{\Theta} = [\boldsymbol{q}_0^T, \dot{\boldsymbol{q}}_0^T, \boldsymbol{\kappa}^T, \boldsymbol{q}^{(0)T}, \boldsymbol{d}^T]$ to minimize the distance between states and additional regularization terms. Here $[\boldsymbol{q}_0^T, \dot{\boldsymbol{q}}_0^T]^T$ denotes the initial state of the model and the vectors $\boldsymbol{\kappa}$, $\boldsymbol{q}^{(0)}$ and $\boldsymbol{d}$ include spring constants, resting angles and damping constants for all modeled springs. The corresponding optimization problem reads as follows,

$$\begin{aligned} \boldsymbol{\Theta} = \arg\min_{\boldsymbol{\Theta}} \Bigg\{ & \frac{w_0}{T}\sum_{t=1}^{T} \left|[\boldsymbol{q},\dot{\boldsymbol{q}}]^T_{\mathrm{mod},t} - [\boldsymbol{q},\dot{\boldsymbol{q}}]^T_{\mathrm{targ},t}\right|^2 \\ & + \frac{w_1}{T}\sum_{t=1}^{T}\left(\boldsymbol{\tau}_t^T\boldsymbol{\tau}_t + \dot{\boldsymbol{\tau}}_t^T\dot{\boldsymbol{\tau}}_t\right) \\ & + \frac{w_2}{T}\sum_{t=1}^{T}(\boldsymbol{J}_c\dot{\boldsymbol{q}}_{\mathrm{mod},t})^T(\boldsymbol{J}_c\dot{\boldsymbol{q}}_{\mathrm{mod},t})\Bigg\}\,, \end{aligned} \tag{10}$$

with $[\boldsymbol{q},\dot{\boldsymbol{q}}]^T_{\mathrm{mod},t} = \boldsymbol{\mathcal{D}}(\boldsymbol{\Theta}, t)$ and $\boldsymbol{\tau}_t = \boldsymbol{\tau}(\boldsymbol{\Theta}, [\boldsymbol{q},\dot{\boldsymbol{q}}]^T_{\mathrm{mod},t})$. The function $\boldsymbol{\mathcal{D}}$ represents the dynamical state development, i.e. the numerical integration of EOM using the Runge-Kutta Dormand-Prince 5 method [10] and starting from the initial state $[\boldsymbol{q}_0^T, \dot{\boldsymbol{q}}_0^T]^T$. The vector $\boldsymbol{\tau}_t$ is composed of all joint torques, which are active at time t and is calculated with Eq. (7).

In the first term of Eq. (10), we evaluate the squared distance between simulated motion and target motion. The second term represents dynamic effort and jerk with torque components calculated according to Eq. (7). This term penalizes highly energetic, jittery and quickly oscillating movement. The last term of the objective function describes the mean velocity of feet contact points on the ground, which is supposed to be near zero. The weights (w_0, w_1, w_2) are empirically set to $(1, 10^{-2}, 10^{-2})$ for walking and to $(1, 5\cdot 10^{-4}, 10^{-2})$ for lifting. This adjustment is necessary, since the required torques to lift weights like 17 kg (as it was performed by the subjects in the test sequences) are essentially higher than active torques during locomotion. We solve the optimization problem using a standard Sequential Quadratic Programming (SQP) approach [22] starting from multiple points.

5. Joint Pose and Torque Estimation

In this section the algorithm which combines the 3D reconstruction of Sec. 3 with the physical model of Sec. 4 is presented. Our proposed algorithm first performs the 3D pose reconstruction (cf. Sec. 3) followed by the physical simulation (cf. Sec. 4) to eliminate any ambiguities between

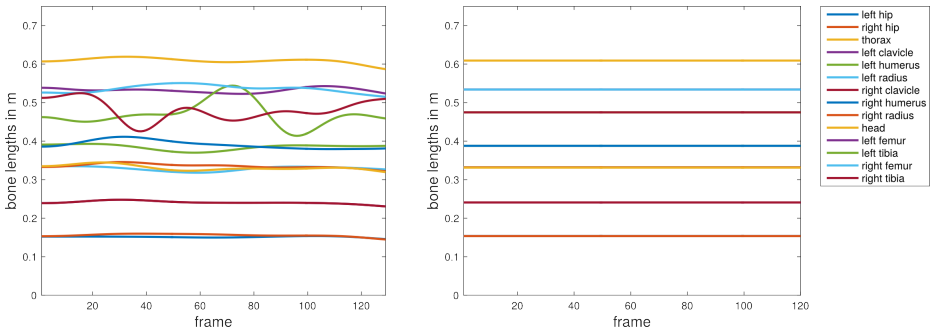

Figure 3. Comparison of the temporal behavior of bone lengths after 3D reconstruction (left) and after applying the physical model (right). Obviously, the desired bone length constancy is assured.

camera and object motion as well as enforcing a physically plausible reconstruction.

Before starting the reconstruction we build the joint parameter space consisting of the weighting matrix $\boldsymbol{D}$ and the physical parameters $\boldsymbol{\Theta}$. While the parameters $\boldsymbol{\Theta}$ are the results of the optimization problem in Eq. (10), the weighting matrix $\boldsymbol{D}$ can be calculated with Eq. (3) and Eq. (5). To eliminate the Kronecker product the matrices $\boldsymbol{S}$ and $\boldsymbol{Q}$ are reshaped so that each column represents a single pose, respectively. By combining Eq. (3) and Eq. (5) the 3D shape can be written as $\tilde{\boldsymbol{S}} = \boldsymbol{B}\boldsymbol{D}\tilde{\boldsymbol{Q}}$, where $\tilde{\boldsymbol{S}}$ and $\tilde{\boldsymbol{Q}}$ correspond to the reshaped matrices of $\boldsymbol{S}$ and $\boldsymbol{Q}$. For known 3D shapes $\boldsymbol{S}$ from the training data, $\boldsymbol{D}$ can be directly calculated via

$$\boldsymbol{D} = \boldsymbol{B}^{+}\tilde{\boldsymbol{S}}\tilde{\boldsymbol{Q}}^{+}, \tag{11}$$

where $\boldsymbol{B}^{+}$ and $\tilde{\boldsymbol{Q}}^{+}$ denote the Moore-Penrose pseudoinverses of $\boldsymbol{B}$ and $\tilde{\boldsymbol{Q}}$, respectively. Finally, a vector $\boldsymbol{v}_k$ composed of the vectorized weighting matrix $\boldsymbol{D}$ and the physical parameters $\boldsymbol{\Theta}$ is assigned to each sequence k in the training set:

$$\boldsymbol{v}_k = \begin{pmatrix} vec(\boldsymbol{D}) \\ \boldsymbol{\Theta} \end{pmatrix}. \tag{12}$$

Here, $vec(\cdot)$ is the vectorization operator which stacks the columns of the matrix into a vector. We assume that a newly observed motion lies in the space spanned by the vectors $\boldsymbol{v}_k$ for each sequence k.

In the first step, the proposed algorithm performs a 3D reconstruction of an observation represented by the matrix $\boldsymbol{W}_{2d}$ as described in Sec. 3. The result is a camera matrix $\boldsymbol{M}$ and the weighting matrix $\boldsymbol{D}$. The consequent 3D shapes are calculated applying Eq. (3) and Eq. (5). Since we use a linear model to represent nonlinear deformations the estimated 3D motion is expected to differ from the real motion. This can be easily seen when analyzing the temporal behavior of the bone lengths, shown on the left in Fig. 3. After the 3D reconstruction the bone lengths fluctuate heavily due to the above mentioned linear model. We address this issue by limiting our parameter space to physically valid motions, in other words, we infer weighting coefficients and physical parameters by means of a k-nearest-neighbor (k-NN) regression in the space spanned by the vectors $\boldsymbol{v}_k$. As suggested by [39] we use the local k-NN regression, which outperforms global approaches like PCA or asymmetric PCA for this particular problem. The recovered physical parameters $\boldsymbol{\Theta}$ are now employed to simulate a 3D motion by integrating the corresponding set of EOM, as described in Sec. 4.

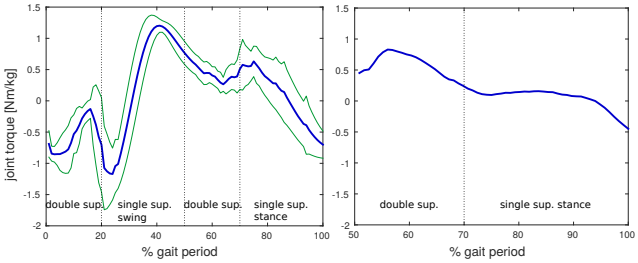

Figure 4. Consistency of torques: Estimated knee torques for all reconstructed walking sequences (left). The torques are shown for a full gait cycle with mean value in blue and standard deviation in green. On the right hand side an example of the knee torque for the stance phase, calculated via inverse dynamics is displayed.

This step converts the rough 3D pose estimation to a physically feasible 3D reconstruction of the observed motion. Comparing the bone lengths variation before (cf. Fig. 3 left) and after physical simulation (cf. Fig. 3 right) indicates an improvement regarding plausibility. Further evaluation is done in Sec. 6.

Additionally, the use of the physical model allows us to resolve the ambiguity between camera and object motion. Based on the knowledge we gain from the physical simulation of the observed object, e.g. the forward movement during walking, we can use a standard camera calibration technique with known 2D-3D point correspondences to reconstruct the camera parameters.

6. Experiments

We evaluate our joint model concerning 3D motion and torque estimation using a training set of MoCap sequences,

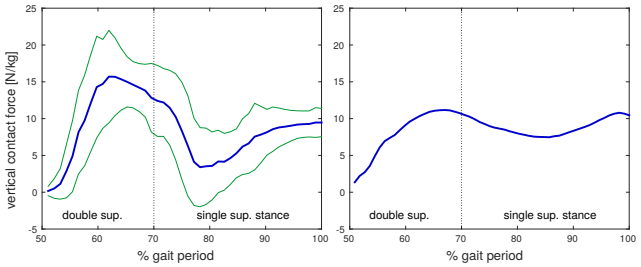

Figure 5. Consistency of forces: Modeled vertical contact forces (left) with mean value in blue and standard deviation in green for the whole set of gait sequences. The curve on the right hand side displays the vertical component of a measured GRF vector.

consisting of 45 walking and 31 lifting motions. The data recording was performed with a Vicon T-series MoCap-system and the corresponding GRF was measured by synchronized AMTI force plates. For each reconstruction, following the method described in Sec. 5, the considered sequence is excluded from the training set.

6.1. Torque and Force Estimation

First of all, we evaluate the estimation of knee torques from 2D walking data. The results are generated as described in Sec. 5, i.e. we reconstruct a 3D motion and infer physical parameters $\boldsymbol{\Theta}$ from the resulting motion coefficients, applying a k-NN regression in our joint parameter space. Based on $\boldsymbol{\Theta}$ we simulate a gait and determine the corresponding model torques using Eq. (7). The resulting mean value of knee torques for all reconstructed walking sequences is shown in Fig. 4 together with the related standard deviation.

For comparison, we calculate the knee torque of an example sequence via inverse dynamics, utilizing force plate data. The model results cover a full gait cycle, while the inverse dynamics solution is only determined for the stance phase. This is due to the inapplicability of inverse dynamics for joints in the swing leg, since the kinematic chain from contact point to joint becomes too long.

The estimated torques are consistent for all reconstructed 3D motions and the shape of the curves and absolute values are similar to inverse dynamics torques. Although, reconstructed extension knee torques tend to be too high from mid-stance to heal-off, compared to values found in biomechanics literature. This might be due to model inaccuracies, concerning mass distribution or imprecision of the skeleton, fitted to the MoCap data.

To analyze the adequacy of our physical model, we depict vertical contact spring forces on the left in Fig. 5. On the right hand side the vertical component of the GRF, measured by force plates is shown for comparison. Minimum and maximum values are slightly high and low, respectively but the overall curve progression resembles ground truth data.

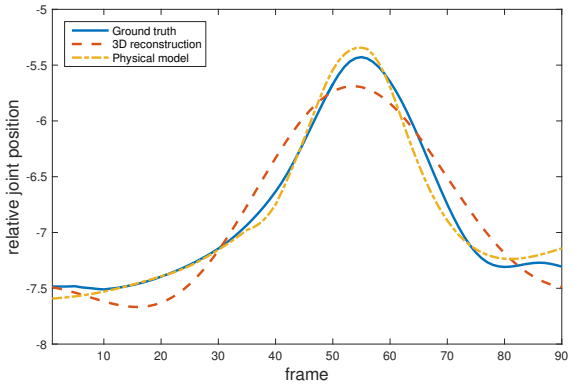

Figure 6. Distance from heel to root joint normalized by leg length. While the 3D reconstruction penetrates the ground in the frames 10 to 20, the physical simulation eliminates all implausible motion.

This experiment shows, that our joint model provides a sound estimation of unobservable 3D torques from monocular videos. The mean computation time for the 3D reconstruction and torque estimation amounts to 21 s, including the optimization for camera and coefficient matrices (Eq. (1)), which accounts for over 99 %. The pure regression of torques from a 3D motion requires computation times in the order of 0.1 s. To put this value into perspective, we implement the method introduced in [6] and optimize using SQP. In doing so we receive computation times that surpass the afore stated result by approximately two orders. All calculations were performed on an 8-core processor and based on unoptimized Matlab code.

6.2. 3D Reconstruction

We performed 3D reconstructions of multiple data sets including walking and lifting motions, both with additional force measurement (cf. Sec. 6). For each sequence multiple random weak perspective cameras are created to obtain monocular input data. Our method achieves a mean 3d error of $0.219m \pm 0.032m$ for walking motion and $0.257m \pm 0.026m$ for lifting motion. Since it is not possible to find an objective measure for physical plausibility we performed multiple experiments observing bone lengths and joint trajectories to show the plausibility of our results.

6.2.1 Physical Plausibility

For human motion analysis bone lengths have to be constant over time. The 3D reconstruction represents the mixing coefficients as weighted DCT bases (cf. Sec. 3) which results in heavily fluctuating bone lengths as shown on the left in Fig. 3. After projecting the obtained weights in the joint parameter space the physical model fixes the bone lengths in the sequence. The bone lengths for the physical simulation are obtained by calculating the mean of the bone lengths in the 3D reconstruction. Fig. 3 shows the bone lengths be-

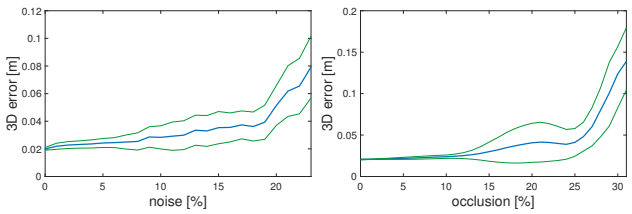

Figure 7. Influence of additional noise (top) and occlusions (bottom) on the reconstruction results: The figure shows the mean (blue) and standard deviation (green) of the 3D error. The noise is given by percent of the subjects body height. The reconstruction stays stable up to 18 % noise on the input data and up to 25 % of occluded data points.

fore and after projecting into the joint parameter space for an example sequence of our test data set.

Since feet motion is one of the most important factors in human gait we evaluate for feet motion separately. Fig. 6 shows the distance from heel to root joint normalized by the leg length for the same sequence as in Fig. 3. In this case the DCT bases cause the heel trajectory to overshoot which causes the heel to penetrate the ground. Applying the physical model adjusts the heels trajectory by eliminating implausible motion from the 3D reconstruction.

6.2.2 Stability

To evaluate the noise stability of our method we add Gaussian noise to the input data. According to Eq. 6 we define a 3D error by $e = \|\boldsymbol{W}_{3d} - \boldsymbol{P}\|_F$, where $\boldsymbol{W}_{3d}$ describes the ground truth poses and $\boldsymbol{P}$ describes the reconstructed poses. The left of Fig. 7 shows the 3D error as a function of the noise relative to the percentage of body height. With noise as large as 18 % of the body height, the 3D error is still close to the 3D error with noiseless reconstruction.

In realistic scenarios (i.e. not using motion capture equipment) one or multiple body parts can be occluded, either by an object or by other body parts. To evaluate the robustness against occlusions we randomly occlude points in the joint trajectories. This can easily be done by setting the values corresponding to the occluded points to zero in the objective function in Eq. (6), which equals to canceling equations. The right of Fig. 7 shows the 3D error as a function of percentage of occluded points in the joint trajectory. The proposed algorithm appears to be stable up to 25 % occlusion. Since the reconstructions in the stable regions are almost identical, the joint torques are consequently very similar. Therefore, we pass an extensive evaluation here and refer to Sec. 6.1.

6.2.3 Real World Data

The proposed algorithm is not restricted to a laboratory setup and can also be used in real world scenarios. Here, we use the KTH football data set [17] as it contains multi-view sequences of a challenging noisy outdoor scene, which shows a football player walking over a playfield. Fig. 8 shows the 3D reconstruction and estimated torques from camera 1. As expected the reconstructions from the other two cameras are very similar and have a maximal reconstruction error of $0.05\,m$. The estimated torques appear to be plausible compared to the torques in Sec. 6.1. Note, that compared to Fig. 7 the torques are larger due to the more dynamic gait pattern of the subject in the KTH data set.

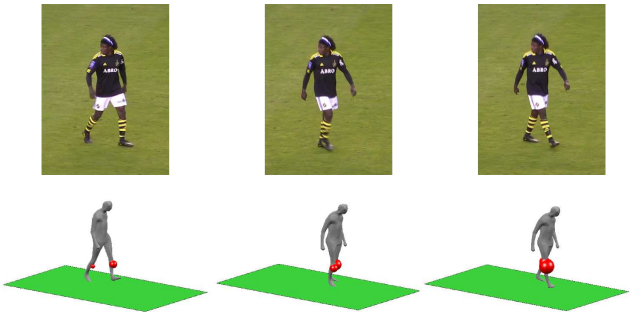

Figure 8. 3D Reconstruction and estimated torques of the KTH football data set. The reconstructions and torques (red spheres) appear to be plausible compared to the corresponding images and torques in Sec. 6.1.

6.2.4 Camera Path

One of the largest benefits of using a physical model for pose estimation is the known object translation which allows for a calibration of the cameras. Every 3D reconstruction technique which does not know the object translation suffers from the ambiguity of camera and object motion. Due to our proposed physical model it is possible to recover the translation of the person which allows us to perform a camera calibration which recovers the camera parameters. For evaluation we created multiple artificial camera paths of weak perspective cameras in a distance up to $10\,m$ from the ground truth 3D data and use the projections as input data for our algorithm. For 50 randomly created camera paths on different sequences we achieve a mean distance of $0.6\,m$ from the ground truth path and a mean angle error of $9.4°$.

6.3. Motion Analysis

In order to demonstrate the benefit of our joint model for health applications, we examine torques during lifting exercises. For this purpose we recorded MoCap sequences of subjects lifting a box that weighed 17 kg. The subjects were asked to lift the box in two different ways: first with bent knees and straight back and then with straight knees and bent back as it is shown in Fig. 1. The data was used to

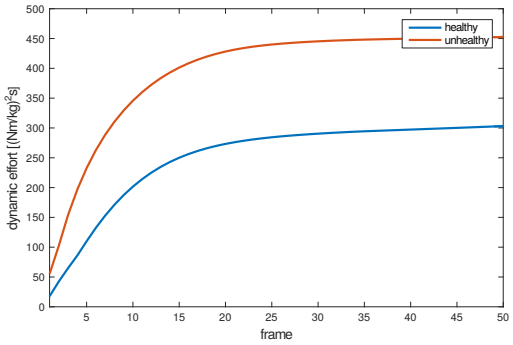

Figure 9. Comparison of different lifting motions: Accumulated quadratic torques (dynamic effort) for the forward-backward flexion of the lumbar vertebrae joint.

construct a joint model space for lifting motions that allows us to analyze the acting joint torques.

We determine the dynamic effort $E_{\mathrm{lv}} = \sum_t \tau_{\mathrm{lv},t}^2$ resulting from the extension torque τ_{lv} in the lower back joint (lumbar vertebrae) to define a health measure for the considered lifting motions. Fig. 9 shows health measure values for a healthy and an unhealthy lifting style, respectively. The motions are clearly distinguishable according to E_{lv}, since the accumulated sum of quadratic torques is about 50 % higher in the case of the unhealthy lifting motion with straight knees and bent back. A visualization of the comparison including estimated torques and modeled contact forces can be found in Fig. 1. This example shows, that our joint model opens the possibility for a direct health rating of 2D motion data supported by profound knowledge about the underlying physics.

7. Conclusion

This paper proposes a joint statistical model for human motion reconstruction and joint torque estimation from monocular image sequences. We combine 3D pose estimation by means of a factorization approach with a physical model of the human body to enforce physical validity on estimated 3D motions and to allow the estimation of unobservable inner moments. We tested the performance of the proposed method in terms of plausibility, accuracy and subjective quality on a dataset of 45 walking and 31 lifting motions as well as on the real world example of the KTH database [17]. Some of the reconstructions including torque estimation are shown in Fig. 8 and Fig. 10.

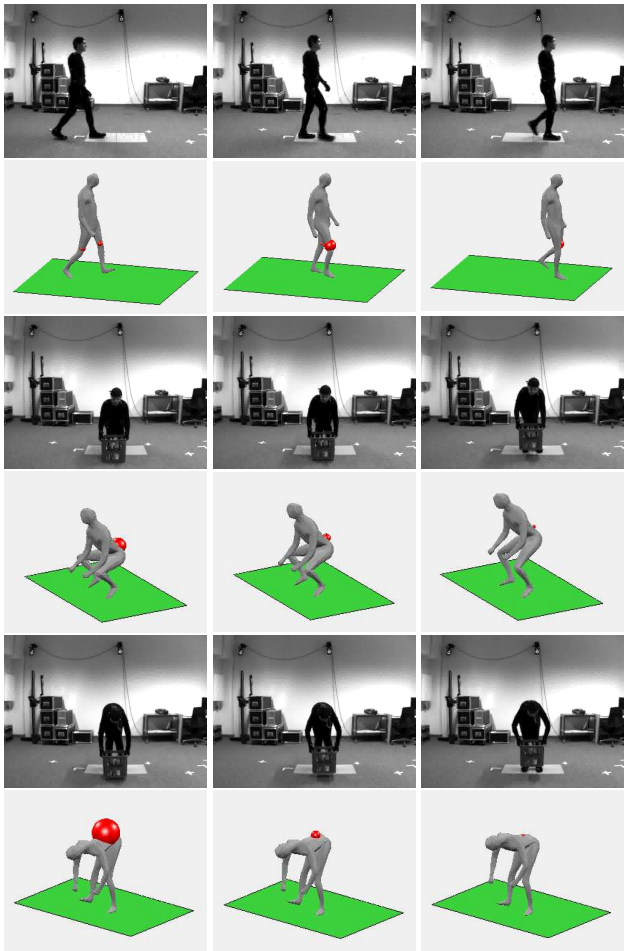

Figure 10. 3D reconstruction and estimated torques of a walking, a healthy lifting and an unhealthy lifting motion (from top to bottom).

References

[1] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 1446–1455, June 2015. 1, 2

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1442–1456, July 2011. 2, 3

[3] K. S. Bhat, S. M. Seitz, J. Popović, and P. K. Khosla. *Computer Vision — ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part I*, chapter Computing the Physical Parameters of Rigid-Body Motion from Video, pages 551–565. Springer Berlin Heidelberg, 2002. 2

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 690–696, 2000. 2, 3

[5] M. A. Brubaker and D. J. Fleet. The kneed walker for human pose tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008. 2

[6] M. A. Brubaker, L. Sigal, and D. J. Fleet. Estimating contact dynamics. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2389–2396, Sept 2009. 2, 4, 6

[7] R. F. Chandler, C. E. Clauser, J. T. McConville, H. M. Reynolds, and J. W. Young. Investigation of inertial properties of the human body. Technical report, Department of Transportation, Report No DOT HS-801 430, Mar 1975. 4

[8] CMU. Human motion capture database, 2014. 1

[9] Y. Dai and H. Li. A simple prior-free method for non-rigid structure-from-motion factorization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 2018–2025, Washington, DC, USA, 2012. IEEE Computer Society. 2

[10] P. Deuflhard, W. Rheinboldt, and F. Bornemann. *Scientific Computing with Ordinary Differential Equations*. Texts in Applied Mathematics. Springer New York, 2012. 4

[11] A. C. Fang and N. S. Pollard. Efficient synthesis of physically valid human motion. *ACM Trans. Graph.*, 22(3):417–426, July 2003. 2

[12] B. J. Fregly, J. A. Reinbolt, K. L. Rooney, K. H. Mitchell, and T. L. Chmielewski. Design of patient-specific gait modifications for knee osteoarthritis rehabilitation. *IEEE Transactions on Biomedical Engineering*, 54(9):1687–1695, Sept 2007. 2

[13] P. Gotardo and A. Martinez. Non-rigid structure from motion with complementary rank-3 spaces. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1

[14] M. Grant and S. Boyd. Graph implementations for non-smooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. 3

[15] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, March 2014. 3

[16] L. Johnson and D. H. Ballard. Efficient codes for inverse dynamics during walking. In *AAAI*, pages 343–349, 2014. 2, 3

[17] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan. Multi-view body part recognition with random forests. In *British Machine Vision Conference (BMVC)*, 2013. 1, 2, 7, 8

[18] C. K. Liu, A. Hertzmann, and Z. Popović. Learning physics-based motion style with nonlinear inverse optimization. *ACM Trans. Graph.*, 24(3):1071–1081, July 2005. 2

[19] X. Lv, J. Chai, and S. Xia. Data-driven inverse dynamics for human motion. *ACM Trans. Graph.*, 35(6):163:1–163:12, 2016. 3

[20] R. Mann and A. Jepson. Towards the computational perception of action. In *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pages 794–799, Jun 1998. 2

[21] H. S. Park, T. Shiratori, I. Matthews, and Y. Sheikh. 3d reconstruction of a moving point from a series of 2d projections. *European Conference on Computer Vision (ECCV)*, September 2010. 1

[22] M. J. D. Powell. *Numerical Analysis: Proceedings of the Biennial Conference held at Dundee, June 28–July 1, 1977*, chapter A fast algorithm for nonlinearly constrained optimization calculations, pages 144–157. Springer Berlin Heidelberg, 1978. 4

[23] V. Ramakrishna, T. Kanade, and Y. A. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision (ECCV)*, October 2012. 1, 2, 3

[24] A. Safonova, J. K. Hodgins, and N. S. Pollard. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Trans. Graph.*, 23(3):514–521, August 2004. 2

[25] T. Schmalz, S. Blumentritt, and R. Jarasch. Energy expenditure and biomechanical characteristics of lower limb amputee gait: The influence of prosthetic alignment and different prosthetic components. *Gait & Posture*, 16(3):255 – 263, 2002. 2

[26] A. L. Schwab and G. M. J. Delhaes. Lecture Notes Multibody Dynamics B, wb1413. 2009. 4

[27] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1-2):4–27, 2010. 1

[28] K. W. Sok, M. Kim, and J. Lee. Simulating biped behaviors from human motion data. *ACM Trans. Graph.*, 26(3), July 2007. 2

[29] F. Steinparz. Co-ordinate transformation and robot control with denavit-hartenberg matrices. *Journal of Microcomputer Applications*, 8(4):303 – 316, 1985. 3

[30] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3d shape from 2d motion. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Neural Information Processing Systems (NIPS)*. MIT Press, 2003. 2

[31] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 2008. 2

[32] M. Vondrak, L. Sigal, and O. C. Jenkins. Physical simulation for probabilistic motion tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008. 2

[33] B. Wandt, H. Ackermann, and B. Rosenhahn. 3d reconstruction of human motion from monocular image sequences. *Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1505–1516, 2016. 1, 2, 3

[34] C. Wang, Y. Wang, Z. Lin, A. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2, 3

[35] X. Wei, J. Min, and J. Chai. Physically valid statistical models for human motion generation. *ACM Trans. Graph.*, 30(3):19:1–19:10, May 2011. 2

[36] M. W. Whittle. Clinical gait analysis: A review. *Human Movement Science*, 15(3):369–387, 1996. 2

[37] Y. Xiang, J. S. Arora, and K. Abdel-Malek. Physics-based modeling and simulation of human walking: a review of optimization-based and other approaches. *Structural and Multidisciplinary Optimization*, 42(1):1–23, 2010. 2

[38] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. In *European Conference on Computer Vision (ECCV)*, May 2004. 2

[39] P. Zell and B. Rosenhahn. *Pattern Recognition: 37th German Conference, GCPR 2015, Aachen, Germany, October 7-10, 2015, Proceedings*, chapter A Physics-Based Statistical Model for Human Gait Analysis, pages 169–180. Springer International Publishing, 2015. 3, 5

[40] V. B. Zordan, A. Majkowska, B. Chiu, and M. Fast. Dynamic response for motion capture animation. *ACM Trans. Graph.*, 24(3):697–701, July 2005. 2