PhysCap: Physically Plausible Monocular 3D Motion Capture in Real Time

SOSHI SHIMADA, Max Planck Institute for Informatics, Saarland Informatics Campus
VLADISLAV GOLYANIK, Max Planck Institute for Informatics, Saarland Informatics Campus
WEIPENG XU, Facebook Reality Labs
CHRISTIAN THEOBALT, Max Planck Institute for Informatics, Saarland Informatics Campus

Marker-less 3D human motion capture from a single colour camera has seen significant progress. However, it is a very challenging and severely ill-posed problem. In consequence, even the most accurate state-of-the-art approaches have significant limitations. Purely kinematic formulations on the basis of individual joints or skeletons, and the frequent frame-wise reconstruction in state-of-the-art methods greatly limit 3D accuracy and temporal stability compared to multi-view or marker-based motion capture. Further, captured 3D poses are often physically incorrect and biomechanically implausible, or exhibit implausible environment interactions (floor penetration, foot skating, unnatural body leaning and strong shifting in depth), which is problematic for any use case in computer graphics.

We, therefore, present PhysCap, the first algorithm for physically plausible, real-time and marker-less human 3D motion capture with a single colour camera at 25 fps. Our algorithm first captures 3D human poses purely kinematically. To this end, a CNN infers 2D and 3D joint positions, and subsequently, an inverse kinematics step finds space-time coherent joint angles and global 3D pose. Next, these kinematic reconstructions are used as constraints in a real-time physics-based pose optimiser that accounts for environment constraints (e.g., collision handling and floor placement), gravity, and biophysical plausibility of human postures. Our approach employs a combination of ground reaction force and residual force for plausible root control, and uses a trained neural network to detect foot contact events in images. Our method captures physically plausible and temporally stable global 3D human motion, without physically implausible postures, floor penetrations or foot skating, from video in real time and in general scenes. PhysCap achieves state-of-the-art accuracy on established pose benchmarks, and we propose new metrics to demonstrate the improved physical plausibility and temporal stability.

Fig. 1. PhysCap captures global 3D human motion in a physically plausible way from monocular videos in real time, automatically and without the use of markers. (Left:) Video of a standing long jump [Peng et al. 2018] and our 3D reconstructions. Thanks to its formulation on the basis of physics-based dynamics, our algorithm recovers challenging 3D human motion observed in 2D while significantly mitigating artefacts such as foot sliding, foot-floor penetration, unnatural body leaning and jitter along the depth channel that troubled earlier monocular pose estimation methods. (Right) Since the output of PhysCap is environment-aware and the returned root position is global, it is directly suitable for virtual character animation, without any further post-processing. The 3D characters are taken from [Adobe 2020]. See our supplementary video for further results and visualisations.
1 INTRODUCTION

3D human pose estimation from monocular RGB images is a very active area of research. Progress is fueled by many applications with an increasing need for reliable, real time and simple-to-use pose estimation. Here, applications in character animation, VR and AR, telepresence, or human-computer interaction, are only a few examples of high importance for graphics.

Monocular and markerless 3D capture of the human skeleton is a highly challenging and severely underconstrained problem [Kovalenko et al. 2019; Martinez et al. 2017; Mehta et al. 2017b; Pavlakos et al. 2018; Wandt and Rosenhahn 2019]. Even the best state-of-the-art algorithms, therefore, exhibit notable limitations. Most methods capture pose kinematically using individually predicted joints but do not produce smooth joint angles of a coherent kinematic skeleton. Many approaches perform per-frame pose estimates with notable temporal jitter, and reconstructions are often in root-relative but not global 3D space. Even if a global pose is predicted, depth prediction from the camera is often unstable. Also, interaction with the environment is usually entirely ignored, which leads to poses with severe collision violations, e.g., floor penetration or the implausible foot sliding and incorrect foot placement. Established kinematic formulations also do not explicitly consider biomechanical plausibility of reconstructed poses, yielding reconstructed poses with improper balance, inaccurate body leaning, or temporal instability.

We note that all these artefacts are particularly problematic in the aforementioned computer graphics applications, in which temporally stable and visually plausible motion control of characters from all virtual viewpoints, in global 3D, and with respect to the physical environment, are critical. Further on, we note that established metrics in widely-used 3D pose estimation benchmarks [Ionescu et al. 2013; Mehta et al. 2017a], such as mean per joint position error (MPJPE) or 3D percentage of correct keypoints (3D-PCK), which are often even evaluated after a 3D rescaling or Procrustes alignment, do not adequately measure these artefacts. In fact, we show (see Sec. 4, and supplemental video) that even some top-performing methods on these benchmarks produce results with substantial temporal noise and unstable depth prediction, with frequent violation of environment constraints, and with frequent disregard of physical and anatomical pose plausibility. In consequence, there is still a notable gap between monocular 3D pose human estimation approaches and the gold standard accuracy and motion quality of suit-based or marker-based motion capture systems, which are unfortunately expensive, complex to use and not suited for many of the aforementioned applications requiring in-the-wild capture.

We, therefore, present PhysCap – a new approach for easy-to-use monocular global 3D human motion capture that significantly narrows this gap and substantially reduces the aforementioned artefacts, see Fig. 1 for an overview. PhysCap is, to our knowledge, the first method that jointly possesses all the following properties: it is fully-automatic, markerless, works in general scenes, runs in real time, captures a space-time coherent skeleton pose and global 3D pose sequence of state-of-the-art temporal stability and smoothness. It exhibits state-of-the-art posture and position accuracy, and captures physically and anatomically plausible poses that correctly adhere to physics and environment constraints. To this end, we rethink and bring together in new way ideas from kinematics-based monocular pose estimation and physics-based human character animation.

The first stage of our algorithm is similar to [Mehta et al. 2017b] and estimates 3D body poses in a purely kinematic, physics-agnostic way. A convolutional neural network (CNN) infers combined 2D and 3D joint positions from an input video, which are then refined in a space-time inverse kinematics to yield the first estimate of skeletal joint angles and global 3D poses. In the second stage, the foot contact and the motion states are predicted for every frame. Therefore, we employ a new CNN that detects heel and forefoot placement on the ground from estimated 2D keypoints in images, and classifies the observed poses into stationary or non-stationary. In the third stage, the finally physically plausible 3D skeletal joint angle and pose sequence is computed in real time. This stage regularises human motion with a torque-controlled physics-based character represented by a kinematic chain with a floating base. To this end, the optimal control forces for each degree of freedom (DoF) of the kinematic chain are computed, such that the kinematic pose estimates from the first stage – in both 2D and 3D – are reproduced as closely as possible. The optimisation ensures that physics constraints like gravity, collisions, foot placement, as well as physical pose plausibility (e.g., balancing), are fulfilled. To summarise, our contributions in this article are:

- The first, to the best of our knowledge, marker-less monocular 3D human motion capture approach on the basis of an explicit physics-based dynamics model which runs in real time and captures global, physically plausible skeletal motion (Sec. 4).
- A CNN to detect foot contact and motion states from images (Sec. 4.2).
- A new pose optimisation framework with a human parametrised by a torque-controlled simulated character with a floating base and PD joint controllers; it reproduces kinematically captured 2D/3D poses and simultaneously accounts for physics constraints like ground reaction forces, foot contact states and collision response (Sec. 4.3).
- Quantitative metrics to assess frame-to-frame jitter and floor penetration in captured motions (Sec. 5.3.1).
- Physically-justified results with significantly fewer artefacts, such as frame-to-frame jitter, incorrect leaning, foot sliding and floor penetration than related methods (confirmed by a user study and metrics), as well as state-of-the-art 2D and 3D accuracy and temporal stability (Sec. 5).

We demonstrate the benefits of our approach through experimental evaluation on several datasets (including newly recorded videos) against multiple state-of-the-art methods for monocular 3D human motion capture and pose estimation.

2 RELATED WORK

Our method mainly relates to two different categories of approaches – (markerless) 3D human motion capture from colour imagery, and physics-based character animation. In the following, we review related types of methods, focusing on the most closely related works.
Multi-View Methods for 3D Human Motion Capture from RGB. Reconstructing humans from multi-view images is well studied. Multi-view motion capture methods track the articulated skeletal motion, usually by fitting an articulated template to imagery [Bo and Sminchisescu 2010; Brox et al. 2010; Elhayek et al. 2016, 2014; Gall et al. 2010; Stoll et al. 2011; Wang et al. 2018; Zhang et al. 2020]. Other methods, sometimes termed performance capture methods, additionally capture the non-rigid surface deformation, e.g., of clothing [Cagniart et al. 2010; Starck and Hilton 2007; Vlasic et al. 2009; Waschbüsch et al. 2005]. They usually fit some form of a template model to multi-view imagery [Bradley et al. 2008; De Aguiar et al. 2008; Martin-Brualla et al. 2018] that often also has an underlying kinematic skeleton [Gall et al. 2009; Liu et al. 2011; Vlasic et al. 2008; Wu et al. 2012]. Multi-view methods have demonstrated compelling results and some enable free-viewpoint video. However, they require expensive multi-camera setups and often controlled studio environments.

Monocular 3D Human Motion Capture and Pose Estimation from RGB. Marker-less 3D human pose estimation (reconstruction of 3D joint positions only) and motion capture (reconstruction of global 3D body motion and joint angles of a coherent skeleton) from a single colour or greyscale image are highly ill-posed problems. The state of the art on monocular 3D human pose estimation has greatly progressed in recent years, mostly fueled by the power of trained CNNs [Habibie et al. 2019; Mehta et al. 2017a]. Some methods estimate 3D pose by combining 2D keypoints prediction with body depth regression [Dabral et al. 2018; Newell et al. 2016; Yang et al. 2018; Zhou et al. 2017] or with regression of 3D joint location probabilities [Mehta et al. 2017b; Pav lakos et al. 2017] in a trained CNN. Lifting methods predict joint depths from detected 2D keypoints [Chen and Ramanan 2017; Martinez et al. 2017; Pavlakos et al. 2018; Tomè et al. 2017]. Other CNNs regress 3D joint locations directly [Mehta et al. 2017a; Rhodin et al. 2018; Tekin et al. 2016]. Another category of methods combines CNN-based keypoint detection with constraints from a parametric body model, e.g., by using reprojection losses during training [Bogo et al. 2016; Braun and Jiang 2016; Habibie et al. 2019]. Some works approach monocular multi-person 3D pose estimation [Rogez et al. 2019] and motion capture [Mehta et al. 2020], or estimate non-rigidly deforming human surface geometry from monocular video on top of skeletal motion [Habermann et al. 2020, 2019; Xu et al. 2020]. In addition to greyscale images, [Xu et al. 2020] use an asynchronous event stream from an event camera as input. Both these latter directions are complementary but orthogonal to our work.

The majority of methods in this domain estimates 3D pose as a root-relative 3D position of the body joints [Kovalenko et al. 2019; Martinez et al. 2017; Moreno-Noguer 2017; Pavlakos et al. 2018; Wandt and Rosenhahn 2019]. This is problematic for applications in graphics, as temporal jitter, varying bone lengths and the often not recovered global 3D pose make animating virtual characters hard. Other monocular methods are trained to estimate parameters or joint angles of a skeleton [Zhou et al. 2016] or parametric model [Kanazawa et al. 2018]. [Mehta et al. 2020, 2017b] employ inverse kinematics on top of CNN-based 2D/3D inference to obtain joint angles of a coherent skeleton in global 3D and in real-time.

Results of all aforementioned methods frequently violate laws of physics, and exhibit foot-floor penetrations, foot sliding, and unbalanced or implausible poses floating in the air, as well as notable jitter. Some methods try to reduce jitter by exploiting temporal information [Kanazawa et al. 2019; Kocabas et al. 2020], e.g., by estimating smooth multi-frame scene trajectories [Peng et al. 2018]. [Zou et al. 2020] try to reduce foot sliding by ground contact constraints. [Zanfir et al. 2018] jointly reason about ground planes and volumetric occupancy for multi-person pose estimation. [Monszpart et al. 2019] jointly infer coarse scene layout and human pose from monocular interaction video, and [Hassan et al. 2019] use a pre-scanned 3D model of scene geometry to constrain kinematic pose optimisation. To overcome the aforementioned limitations, no prior work formulates monocular motion capture on the basis of an explicit physics-based dynamics model and in real-time, as we do.

Physics-Based Character Animation. Character animation on the basis of physics-based controllers has been investigated for many years [Barzel et al. 1996; Sharon and van de Panne 2005; Wrotek et al. 2006], and remains an active area of research, [Andrews et al. 2016; Bergamin et al. 2019; Levine and Popović 2012; Zheng and Yamane 2013]. [Levine and Popovic 2012] employ a quasi-physical simulation that approximates a reference motion trajectory in real-time. They can follow non-physical reference motion by applying a direct actuation at the root. By using proportional derivative (PD) controllers and computing optimal torques and contact forces, [Zheng and Yamane 2013] make a character follow a reference motion captured while keeping balance. [Liu et al. 2010] proposed a probabilistic algorithm for physics-based character animation. Due to the stochastic property and inherent randomness, their results evince variations, but the method requires multiple minutes of runtime per sequence. [Andrews et al. 2016] employ rigid dynamics to drive a virtual character from a combination of marker-based motion capture and body-mounted sensors. This animation setting is related to motion transfer onto robots. [Nakaoka et al. 2007] transferred human motion captured by a multi-camera marker-based system onto a robot, with an emphasis on leg motion. [Zhang et al. 2014] leverage depth cameras and wearable pressure sensors and apply physics-based motion optimisation. We take inspiration from these works for our setting, where we have to capture in a physically correct way and in real time global 3D human motion from images, using intermediate pose reconstruction results that exhibit notable artefacts and violations of physics laws. PhysCap, therefore, combines an initial kinematics-based pose reconstruction with PD controller based physical pose optimisation.

Several recent methods apply deep reinforcement learning to virtual character animation control [Bergamin et al. 2019; Lee et al. 2019; Peng et al. 2018]. Peng et al. [2018] propose a reinforcement learning approach for transferring dynamic human performances observed in monocular videos. They first estimate smooth motion trajectories with recent monocular human pose estimation techniques, and then train an imitating control policy for a virtual character. [Bergamin et al. 2019] train a controller for a virtual character from several minutes of motion capture data which covers the expected variety of motions and poses. Once trained, the virtual character can follow directional commands of the user in real-time.
Physically Plausible Monocular 3D Human Motion Capture. Only a few works on monocular 3D human motion capture using explicit physics-based constraints exist [Li et al. 2019; Vondrak et al. 2012; Wei and Chai 2010; Zell et al. 2017]. [Wei and Chai 2010] capture 3D human poses from uncalibrated monocular video using physics constraints. Their approach requires manual user input for each frame of a video. In contrast, our approach is automatic, runs in real time, and uses a different formulation for physics-based pose optimisation geared to our setting. [Vondrak et al. 2012] capture bipedal controllers from a video. Their controllers are robust to perturbations and generalise well for a variety of motions. However, unlike our PhysCap, the generated motion often looks unnatural and their method does not run in real time. [Zell et al. 2017] capture poses and internal body forces from images only for certain classes of motion (e.g., lifting and walking) by using a data-driven approach, but not an explicit forward dynamics approach handling a wide range of motions, like ours.

Our PhysCap bears most similarities with the rigid body dynamics based monocular human pose estimation by Li et al. [2019]. Li et al. estimate 3D poses, contact states and forces from input videos with physics-based constraints. However, their method and our approach are substantially different. While Li et al. focus on object-person interactions, we target a variety of general motions, including complex acrobatic motions such as backflipping without objects. Their method does not run in real time and requires manual annotations on images to train the contact state estimation networks. In contrast, we leverage the PD controller based inverse dynamics tracking, which results in physically plausible, smooth and natural skeletal pose and root motion capture in real time. Moreover, our contact state estimation network relies on annotations generated in a semi-automatic way. This enables our architecture to be trained on large datasets, which results in the improved generalisability. No previous method of the reviewed category “physically plausible monocular 3D human motion capture” combines the ability of our algorithm to capture global 3D human pose of similar quality and physical plausibility in real time.

3 BODY MODEL AND PRELIMINARIES

The input to PhysCap is a 2D image sequence \( I_t, t \in \{1, \ldots, T \} \), where \( T \) is the total number of frames and \( t \) is the frame index. We assume a perspective camera model and calibrate the camera and floor location before tracking starts. Our approach outputs a physically plausible real-time 3D motion capture result \( \mathbf{q}_t^{\text{phys}} \in \mathbb{R}^m \) (where \( m \) is the number of degrees of freedom) that adheres to the image observation, as well as physics-based posture and environment constraints. For our human model, \( m = 43 \). Joint angles are parametrised by Euler angles. The mass distribution of our character is computed following [Liu et al. 2016]. Our character model has a skeleton composed of 37 joints and links. A link defines the volumetric extent of a body part via a collision proxy. The foot and heel links, centred at the respective joints of our character (see Fig. 2), are involved in the mesh collision checks with the floor plane in the physics engine [Coumans and Bai 2016].
4 METHOD

Our PhysCap approach includes three stages, see Fig. 3 for an overview. The first stage performs kinematic pose estimation. This encompasses 2D heatmap and 3D location map regression for each body joint with a CNN, followed by a model-based space-time pose optimisation step (Sec. 4.1). This stage returns 3D skeleton pose in joint angles \( q_{kin}^{t} \in \mathbb{R}^{m} \) along with the 2D joint keypoints \( K_t \in \mathbb{R}^{s \times 2} \) for every image; \( s \) denotes the number of 2D joint keypoints. As explained earlier, this initial kinematic reconstruction \( q_{kin}^{t} \) is prone to physically implausible effects such as foot-floor penetration, foot skating, anatomically implausible body leaning and temporal jitter, especially notable along the depth dimension.

The second stage performs foot contact and motion state detection, which uses 2D joint detections \( K_t \) to classify the poses reconstructed so far into stationary and non-stationary – this is stored in one binary flag. It also estimates binary foot-floor contact flags, i.e., for the toes and heels of both feet, resulting in four binary flags (Sec. 4.2). This stage outputs the combined state vector \( b_t \in \mathbb{R}^{5} \).

The third and final stage of PhysCap is the physically plausible global 3D pose estimation (Sec. 4.3). It combines the estimates from the first two stages with physics-based constraints to yield a physically plausible real-time 3D motion capture result that adheres to physics-based posture and environment constraints \( q_{phys}^{t} \in \mathbb{R}^{m} \). In the following, we describe each of the stages in detail.

4.1 Stage I: Kinematic Pose Estimation

Our kinematic pose estimation stage follows the real-time VNect algorithm [Mehta et al. 2017b], see Fig. 3, stage I. We first predict heatmaps of 2D joints and root-relative location maps of joint positions in 3D with a specially tailored fully convolutional neural network using a ResNet [He et al. 2016] core. The ground truth joint locations for training are taken from the MPII [Andriluka et al. 2014] and MPI-INF-3DHP [Mehta et al. 2017a] and Human3.6m [Ionescu et al. 2013] datasets in the 2D case, and MPI-INF-3DHP [Mehta et al. 2017a] and Human3.6m [Ionescu et al. 2013] datasets in the 3D case.

Next, the estimated 2D and 3D joint locations are temporally filtered and used as constraints in a kinematic skeleton fitting step.
The interplay of internal body forces and the ground reaction force (GRF) – applied when the feet touch the ground – allows humans to walk and control their posture. The ground reaction force (GRF) enables humans to walk and control their posture.

The result of stage I is a temporally-consistent joint angle sequence $q_{kin}^t$, the projection term $E_{proj}$, the projection term $E_{proj}$, the temporal stability term $E_{smooth}$, and the depth uncertainty correction term $E_{depth}$. The data term which constrains the 3D pose to be close to the 3D joint predictions from the CNN. The projection term $E_{proj}$ enforces the pose $q_{kin}^t$ to project it to the 2D keypoints (joints) detected by the CNN. Note that this reprojection constraint, together with calibrated camera and calibrated bone lengths, enables computation of the global 3D root (pelvis) position in the camera space. Temporal stability is further imposed by penalising the root’s acceleration and variations along the depth channel by $E_{smooth}$ and $E_{depth}$, respectively. The energy (3) is optimised by non-linear least squares (Levenberg-Marquardt algorithm [Levenberg 1944; Marquardt 1963]), and the obtained vector of joint angles and the root position $q_{kin}^t$ of a skeleton with fixed bone lengths are smoothed by an adaptive first-order low-pass filter [Casiez et al. 2012].

Skelton bone lengths of a human can be computed, up to a global scale, from averaged 3D joint detections of a few initial frames. Knowing the metric height of the human determines the scale factor to compute metrically correct global 3D poses.

The result of stage I is a temporally-consistent joint angle sequence but, as noted earlier, captured poses can exhibit artefacts and contradict physical plausibility (e.g., evince floor penetration, incorrect body leaning, temporal jitter, etc.).

The supervisory labels for training this network are automatically computed on a subset of the 3D motion sequences of the Human3.6M [Ionescu et al. 2013] and DeepCap [Habermann et al. 2020] datasets using the following criteria: the forefoot and heel joint contact labels are computed based on the assumption that a joint in contact is not sliding, i.e., the velocity is lower than 5 cm/sec. In addition, we use a height criterion, i.e., the forefoot/heel, when in contact with the floor, has to be at a 3D height that is lower than a threshold $h_{thres}$. To determine this threshold for each sequence, we calculate the average heel $h_{avg}^{heel}$ and forefoot $h_{avg}^{foot}$ heights for each subject using the first ten frames (when both feet touch the ground). Thresholds are then computed as $h_{thres}^{heel} = h_{avg}^{heel} + 5$ cm for heels and $h_{thres}^{foot} = h_{avg}^{foot} + 5$ cm for feet.
$h_{\text{avg}} + 5\text{cm}$ for the forefeet. This second criterion is needed since, otherwise, a foot in the air that is kept static could also be labelled as being in contact.

We also automatically label stationary and non-stationary poses on the same sequences. When standing and walking, the CoG of the human body typically lies close to the pelvis in 3D, which corresponds to the skeletal root position in both the Human3.6M and DeepCap datasets. Therefore, when the velocity of 3D root is lower than a threshold $\varphi_v$, we classify the pose as stationary, and non-stationary otherwise. In total, around 600k sets of contact and motion state labels for the human images are generated.

### 4.3 Stage III: Physically Plausible Global 3D Pose Estimation

Stage III uses the results of stages I and II as inputs, i.e., $q_{\text{kin}}^k$ and $b_i$. It transforms the kinematic motion estimate into a physically plausible global 3D pose sequence that corresponds to the images and adheres to anatomy and environmental constraints imposed by the laws of physics. To this end, we represent the human as a torque-controlled simulated character with a floating base and PD joint controllers [A. Salem and Aly 2015]. The core is to solve an energy-based optimisation problem to find the vector of forces $\tau$ and accelerations $\ddot{q}$ of the character such that the equations of motion with constraints are fulfilled (Sec. 4.3.5). This optimisation is preceded by several preprocessing steps applied to each frame.

First i), we correct $\ddot{q}_{\text{kin}}^k$ if it is strongly implausible based on several easy-to-test criteria (Sec. 4.3.1). Second ii), we estimate the desired acceleration $\ddot{q}_{\text{des}} \in \mathbb{R}^n$ necessary to reproduce $\ddot{q}_{\text{kin}}^k$ based on the PD control rule (Secs. 4.3.2). Third iii), in input frames in which a foot is in contact with the floor (Sec. 3.3.3), we estimate the ground reaction force (GRF) $\lambda$ (Sec. 4.3.4). Fourth iv), we solve the optimisation problem (10) to estimate $\tau$ and accelerations $\ddot{q}$ where the equation of motion with the estimated GRF $\lambda$ and the contact constraint to avoid foot-floor penetration (Sec. 4.3.5) are integrated as constraints. Note that the contact constraint is integrated only when the foot is in contact with the floor. Otherwise, only the equation of motion without GRF is introduced as a constraint in (10). v) Lastly, the pose is updated using the finite-difference method (Eq. (1)) with the estimated acceleration $\ddot{q}$. The steps ii) - v) are iterated $n = 4$ times for each frame of video.

As also observed by [Andrews et al. 2016], this two-step optimisation iii) and iv) reduces direct actuation of the character’s root as much as possible (which could otherwise lead to slightly unnatural locomotion), and explains the kinematically estimated root position and orientation by torques applied to other joints as much as possible when there is a foot-floor contact. Moreover, this two-step optimisation is computationally less expensive than directly computing $\ddot{q}$, $\tau$ and $\lambda$ simultaneously [Zheng and Yamane 2013].

Our algorithm thus finds a plausible balance between pose accuracy, physical accuracy, the naturalness of captured motion and real-time performance.

#### 4.3.1 Pose Correction

Due to the error accumulation in stage I (e.g., as a result of the deviation of 3D annotations from the joint rotation centres in the skeleton model, see Fig. 5-(a), as well as inaccuracies in the neural network predictions and skeleton fitting), the estimated 3D pose $\ddot{q}_{\text{kin}}^k$ is often not physically plausible. Therefore, prior to torque-based optimisation, we pre-correct a pose $\ddot{q}_{\text{kin}}^k$ from stage I if it is 1) stationary and 2) unbalanced, i.e., the CoG projects outside the BoS. If both correction criteria are fulfilled, we compute the angle $\theta_t$ between the ground plane normal $v_p$ and the vector $v_b$ that defines the direction of the spine relative to the root in the local character’s coordinate system (see Fig. 5-(b) for the schematic visualisation). We then correct the orientation of the virtual character towards a posture, for which CoG projects inside BoS. Correcting $\theta_t$ in one large step could lead to instabilities in physics-based pose optimisation. Instead, we reduce $\theta_t$ by a small rotation of the virtual character around its horizontal axis (i.e., the axis passing through the transverse plane of a human body) starting with the corrective angle $\xi_t = \frac{\theta_t}{10}$ for the first frame. Thereby, we accumulate the degree of correction in $\xi$ for the subsequent frames, i.e., $\xi_{t+1} = \xi_t + \frac{\theta_t}{10}$. Note that $\theta_t$ is decreasing for every frame and the correction step is performed for all subsequent frames until 1) the pose becomes non-stationary or 2) CoG projects inside BoS.

However, simply correcting the spine orientation by the skeleton rotation around the horizontal axis can lead to implausible standing poses, since the knees can still be unnaturally bent for the obtained upright posture (see Fig. 5-(c) for an example). To account for that, we adjust the respective DoFs of the knees and hips such that the relative orientation between upper legs and spine, as well as upper and lower legs, are more straight. The hip and knee correction starts if both correction criteria are still fulfilled and $\theta_t$ is already very small. Similarly to the $\theta$ correction, we introduce accumulator variables for every knee and every hip. The correction step for knees and hips is likewise performed until 1) the pose becomes non-stationary or 2) CoG projects inside BoS.

#### 4.3.2 Computing the Desired Accelerations

To control the physics-based virtual character such that it reproduces the kinematic estimate $\ddot{q}_{\text{kin}}^k$, we set the desired joint acceleration $\ddot{q}_{\text{des}}$ following the PD controller rule:

$$\ddot{q}_{\text{des}} = \ddot{q}_{\text{kin}}^k + k_p (q_{\text{kin}}^k - q) + k_d (q_{\text{kin}}^k - \ddot{q}).$$

The desired acceleration $\ddot{q}_{\text{des}}$ is later used in the GRF estimation step (Sec. 4.3.4) and the final pose optimisation (Sec. 4.3.5). Controlling the character motion on the basis of a PD controller in the system enables the character to exert torques $\tau$ which reproduce the kinematic estimate $\ddot{q}_{\text{kin}}^k$, while significantly mitigating undesired effects such as joint and base position jitter.

#### 4.3.3 Foot-Floor Collision Detection

To avoid foot-floor penetration in the final pose sequence and to mitigate contact position sliding, we integrate hard constraints in the physics-based pose optimisation to enforce zero velocity of forefront and heel links in Sec. 4.3.5. However, these constraints can lead to unnatural motion in rare cases when the state prediction network may fail to estimate the correct foot contact states (e.g., when the foot suddenly stops in the air while walking). We thus update the contact state output of the state prediction network $b_{t,j \in \{1,\ldots,4\}}$, to yield $b_{t,j \in \{1,\ldots,4\}}$ as

1either after the correction or already in $\ddot{q}_{\text{kin}}^k$ provided by stage I
follows:

\[
k'_{\ell, j / \{1, \ldots, 4\}} = \begin{cases} 
1, & \text{if } (b^j = 1 \text{ and } h^j < \psi ) \text{ or } \\
\text{the } j\text{-th link collides with the floor plane,} & \\
0, & \text{otherwise.}
\end{cases}
\] (5)

This means we consider a foot or heel link to be in contact only if its height \(h^j\) is less than a threshold \(\psi = 0.1\)m above the calibrated ground plane.

In addition, we employ the Pybullet \cite{Coumans2016} physics engine to detect foot-floor collision for the left and right foot links. Note that combining the mesh collision information with the predictions from the state prediction network is necessary because 1) the foot may not touch the floor plane in the simulation when the subject’s foot is actually in contact with the floor due to the inaccuracy of \(q_{\text{kin}}^j\), and 2) the foot can penetrate into the mesh floor plane if the network misdetects the contact state when there is actually a foot contact in \(I_1\).

4.3.4 **Ground Reaction Force (GRF) Estimation.** We first compute the GRF \(\lambda\) – when there is a contact between a foot and floor – which best explains the motion of the root as coming from stage I. However, the target trajectory from stage I can be physically implausible, and we will thus eventually also require a residual force directly applied on the root to explain the target trajectory; this force will be computed in the final optimisation. To compute the GRF, we solve the following minimisation problem:

\[
\min_{\lambda} \| M_1 \tilde{q}_{\text{des}} - J^T_{\lambda} G \lambda \|,
\] (6)

s.t. \(\lambda \in F\),

where \(\|\cdot\|\) denotes \(L_2\)-norm, and \(M_1 \in \mathbb{R}^{6 \times 6}\) together with \(J^T_{\lambda} \in \mathbb{R}^{6 \times 6}\) are the first six rows of \(M\) and \(J^T\) that correspond to the root joint, respectively. Since we do not consider sliding contact, the contact force \(\lambda\) has to satisfy friction cone constraints. Thus, we formulate a linearised friction cone constraint \(F\). That is,

\[
F^\lambda = \{ \lambda \in \mathbb{R}^6 | |\lambda^j_n| > 0, |\lambda^j_n| \leq \bar{\mu} |\tilde{\lambda}^j_{\text{des}}|, |\lambda^j_i| \leq \bar{\mu} |\tilde{\lambda}^j_{\text{des}}| \},
\] (7)

where \(\tilde{\lambda}^j_{\text{des}}\) is a normal component, \(\lambda^j_n\) and \(\lambda^j_i\) are the tangential components of a contact force at the \(j\)-th contact position; \(\mu\) is a friction coefficient which we set to 0.8 and the friction coefficient of inner linear cone approximation reads \(\bar{\mu} = \mu / \sqrt{2}\).

The GRF \(\lambda\) is then integrated into the subsequent optimisation step (10) to estimate torques and accelerations of all joints in the body, including an additional residual direct root actuation component that is needed to explain the difference between the global 3D root trajectory of the kinematic estimate and the final physically correct result. The aim is to keep this direct root actuation as small as possible, which is best achieved by a two-stage strategy that first estimates the GRF separately. Moreover, we observed this two-step optimisation enables faster computation than estimating \(\lambda, \tilde{q}\) and \(\tau\) all at once. It is hence more suitable for our approach which aims at real-time operation.

4.3.5 **Physics-Based Pose Optimisation.** In this step, we solve an optimisation problem to estimate \(\tau\) and \(\tilde{q}\) to track \(q_{\text{kin}}^j\) using the equation of motion (2) as a constraint. When contact is detected (Sec. 4.3.3), we integrate the estimated ground reaction force \(\lambda\) (Sec. 4.3.4) in the equation of motion. In addition, we introduce contact constraints to prevent foot-floor penetration and foot sliding when contacts are detected.

Let \(\ell_j\) be the velocity of the \(j\)-th contact link. Then, using the relationship between \(\ell_j\) and \(q\) \cite{Featherstone2014}, we can write:

\[
J^T \tilde{q} = \ell_j.
\] (8)

When the link is in contact with the floor, the velocity perpendicular to the floor has to be zero or positive to prevent penetration. Also, we allow the contact links to have a small tangential velocity \(\sigma\) to prevent an immediate foot motion stop which creates visually unnatural motion. Our contact constraint inequalities read:

\[
0 \leq r^n_j, |r^n_j| \leq \sigma, \text{ and } |r^o_j| \leq \sigma,
\] (9)

where \(r^n_j\) is the normal component of \(\ell_j\), and \(r^o_j\) along with \(r^f_j\) are the tangential elements of \(\ell_j\).

Using the desired acceleration \(\tilde{q}_{\text{des}}\) (Eq. (4)), the equation of motion (2), optimal GRF \(\lambda\) estimated in (6) and contact constraints (9), we formulate the optimisation problem for finding the physics-based motion capture result as:

\[
\min_{\tilde{q}} \| \tilde{q} - \tilde{q}_{\text{des}} \| + \| \tau \|,
\]

s.t. \(M \bar{\epsilon} - \tau = J^T \bar{\epsilon} + c(q, \tilde{q}),\) and

\[
0 \leq r^n_j, |r^n_j| \leq \sigma, |r^f_j| \leq \sigma, \forall j.
\] (10)

The first energy term forces the character to reproduce \(q_{\text{kin}}^j\). The second energy term is the regulariser that minimises \(\tau\) to prevent the overshooting, thus modelling natural human-like motion.

After solving (10), the character pose is updated by Eq. (1). We iterate the steps ii) - v) (see stage III in Fig. 3) \(n = 4\) times, and stage III returns the \(n\)-th output from v) as the final character pose \(q_{\text{phys}}^j\).

The final output of stage III is a sequence of joint angles and global root translations and rotations that explains the image observations, followed the purely kinematic reconstruction from stage I, yet is physically and anatomically plausible and temporally stable.

5 **RESULTS**

We first provide implementation details of PhysCap (Sec. 5.1) and then demonstrate its qualitative state-of-the-art results (Sec. 5.2). We next evaluate PhysCap’s performance quantitatively (Sec. 5.3) and conduct a user study to assess the visual physical plausibility of the results (Sec. 5.4).

We test PhysCap on widely-used benchmarks \cite{Habermann2020, Ionescu2013, Mehta2017a} as well as on backflip and jump sequences provided by \cite{Peng2018}. We also collect

<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Sequence Name</th>
<th>Duration [sec]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>building 1</td>
<td>132</td>
</tr>
<tr>
<td>2</td>
<td>building 2</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>forest</td>
<td>105</td>
</tr>
<tr>
<td>4</td>
<td>backyard</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>balance beam 1</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>balance beam 2</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 1. Names and duration of our six newly recorded outdoor sequences captured using SONY DSC-RX0 at 25 fps.
PhysCap: Physically Plausible Monocular 3D Motion Capture in Real Time

Fig. 6. Two examples of reprojected 3D keypoints obtained by our approach (light blue colour) and Vnect [Mehta et al. 2017b] (yellow colour) together with the corresponding 3D visualisations from different view angles. PhysCap produces much more natural and physically plausible postures whereas Vnect suffers from unnatural body leaning (see also the supplementary video).

5.1 Implementation

Our method runs in real time (25 fps on average) on a PC with a Ryzen7 2700 8-Core Processor, 32 GB RAM and GeForce RTX 2070 graphics card. In stage I, we proceed from a freely available demo version of Vnect [Mehta et al. 2017b]. Stages II and III are implemented in python. In stage II, the network is implemented with PyTorch [Paszke et al. 2019]. In stage III, we use the Rigid Body Dynamics Library [Felis 2017] to compute dynamic quantities. We employ the Pybullet [Coumans and Bai 2016] as a physics engine for the character motion visualisation and collision detection. In this paper, we set the proportional gain value \( k_p \) and derivative gain value \( k_d \) for all joints to 300 and 20, respectively. For the root angular acceleration, \( k_p \) and \( k_d \) are set to 340 and 30, respectively. \( k_p \) and \( k_d \) of the root linear acceleration are set to 1000 and 80, respectively. These settings are used in all experiments.

5.2 Qualitative Evaluation

The supplementary video and result figures in this paper, in particular Figs. 1 and 11 show that PhysCap captures global 3D human poses in real time, even of fast and difficult motions, such as a backflip and a jump, which are of significantly improved quality compared to previous monocular methods. In particular, captured motions are much more temporally stable, and adhere to laws of physics with respect to the naturalness of body postures and fulfillment of environmental constraints, see Figs. 6–8 and 10 for the examples of more natural 3D reconstructions. These properties are essential for many applications in graphics, in particular for stable real-time character animation, which is feasible by directly applying our method’s output (see Fig. 1 and the supplementary video).
of testing a method on multiple sequences and reporting the accuracy of 3D joint positions as well as the accuracy of the reprojection into the input views. The accuracy in 3D is evaluated by mean per joint position error (MPJPE) in mm, percentage of correct keypoints (PCK) and the area under the receiver operating characteristic (ROC) curve abbreviated as AUC. The reprojection or mean pixel error $e_{2D}$ is obtained by projecting the estimated 3D joints onto the input images and taking the average per frame distance to the ground truth 2D joint positions. We report $e_{2D}$ and its standard deviation denoted by $\sigma_{2D}$ with the images of size 1024 x 1024 pixels.

As explained earlier, these metrics only evaluate limited aspects of captured 3D poses and do not account for essential aspects of temporal stability, smoothness and physical plausibility in reconstructions such as jitter, foot sliding, foot-floor penetration and unnaturally balanced postures. As we show in the supplemental video, top-performing methods on MPJPE and 3D PCK can fare poorly with respect to these criteria. Moreover, MPJPE and PCK are often reported after rescaling of the result in 3D or Procrustes alignment, which further makes these metrics agnostic to the aforementioned artefacts. Thus, we introduce four additional metrics which allow to evaluate the physical plausibility of the results, i.e., reprojection error to unseen views $e_{2D}$, motion jitter error $e_{\text{smooth}}$ and two floor penetration errors – Mean Penetration Error (MPE) and Percentage of Non-Penetration (PNP).

When choosing a reference side view for $e_{2D}$, we make sure that the viewing angle between the input and side views has to be sufficiently large, i.e., more than $\pi/2$. Otherwise, if a side view is close to the input view, such effects as unnatural leaning forward can still remain undetected by $e_{2D}$ in some cases. After reprojection of a 3D structure to an image plane of a side view, all further steps for calculating $e_{2D}$ are similar to the steps for the standard reprojection error. We also report $\sigma_{2D}$, i.e., the standard deviation of $e_{2D}$.

To quantitatively compare the motion jitter, we report the deviation of the temporal consistency from the ground truth 3D pose. Our smoothness error $e_{\text{smooth}}$ is computed as follows:

$$
\begin{align*}
J_{\text{X}} &= ||p^t_X - p^{t-1}_X||, \\
J_{\text{GT}} &= ||p^t_{\text{GT}} - p^{t-1}_{\text{GT}}||, \\
\epsilon_{\text{smooth}} &= \frac{1}{TM} \sum_{t=1}^{T} \sum_{m=1}^{M} |J_{\text{GT}} - J_{\text{X}}|, \\
\end{align*}
$$

where $p^t_X$ represents the 3D position of joint $s$ in the time frame $t$. $T$ and $m$ denote the total numbers of frames in the video sequence and target 3D joints, respectively. The subscripts $X$ and $GT$ stand for the predicted output and ground truth, respectively. A lower $\epsilon_{\text{smooth}}$ indicates lower motion jitter in the predicted motion sequence.

MPE and PNP measure the degree of non-physical foot penetration into the ground. MPE is the mean distance between the floor and 3D foot position, and it is computed only when the foot is in contact with the floor. We use the ground truth foot contact labels (Sec. 4.2) to judge the presence of the actual foot contacts. The complementary PNP metric shows the ratio of frames where the feet are not below the floor plane over the entire sequence.

5.3.2 Quantitative Evaluation Results. Table 2 summarises MPJPE, PCK and AUC for root-relative joint positions with (first row) and
Table 3. 2D projection error of a frontal view (input) and side view (non-input) on DeepCap dataset [Habermann et al. 2020]. PhysCap performs similarly to VNect on the frontal view, and significantly better on the side view. For further details, see Sec. 5.3 and Fig. 7.

<table>
<thead>
<tr>
<th></th>
<th>Front View</th>
<th>Side View</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>input $\epsilon_{2D}$ [pixel]</td>
<td>input $\epsilon_{2D}$ [pixel]</td>
</tr>
<tr>
<td>Ours</td>
<td>21.1</td>
<td>14.3</td>
</tr>
<tr>
<td>VNect [Mehta et al. 2017b]</td>
<td>6.7</td>
<td>2.7</td>
</tr>
</tbody>
</table>

In terms of MPJPE, PCK and AUC, our method does not outperform the other approaches consistently but achieves an accuracy that is comparable and often close to the highest on Human3.6M, DeepCap and MPI-INF-3DHP. In the third row, we additionally evaluate the global 3D base position accuracy, which is critical for character animation from the captured data. Here, PhysCap consistently outperforms the other methods on all the datasets.

As noted earlier, the above metrics only paint an incomplete picture. Therefore, we also measure the 2D projection errors to the input and side views on the DeepCap dataset, since this dataset includes multiple synchronised views of dynamic scenes with a wide baseline. Table 3 summarises the mean pixel errors $\epsilon_{2D}$ and $\epsilon_{side}$ together with their standard deviations. In the frontal view, i.e., on $\epsilon_{2D}$, VNect has higher accuracy than PhysCap. However, this comes at the price of frequently violating physics constraints (floor penetration) and producing unnaturally leaning and jarring 3D poses (see also the supplemental video). In contrast, since PhysCap explicitly models physical pose plausibility, it excels VNect in the side view, which reveals VNect’s implausibly leaning postures and root position instability in depth, also see Figs. 6 and 7.

To assess motion smoothness, we report $\epsilon_{smooth}$ and its standard deviation $\sigma_{smooth}$ in Table 4. Our approach outperforms VNect and HMR by a big margin on both datasets. Our method is better than HMMR on DeepCap dataset and marginally worse on Human3.6M. HMMR is one of the current state-of-the-art algorithms that has an explicit temporal component in the architecture.

Table 5 summarises the MPE and PNP for VNect and PhysCap on DeepCap dataset. Our method shows significantly better results compared to VNect, i.e., about a 30% lower MPE and a by 100% better result in PNP, see Fig. 8 for qualitative examples. Fig. 9 shows plots of contact forces as the functions of time calculated by our approach on the walking sequence from our newly recorded dataset.
The notion of physical plausibility can be understood and perceived subjectively from person to person. Therefore, in addition to the quantitative evaluation with existing and new metrics, we perform an online user study which allows to subjectively assess and compare the perceived degree of different effects in the reconstructions by a broad audience of people with different backgrounds in computer graphics and vision. In total, we prepared 34 questions with videos, in which we always showed one or two reconstructions at a time (our result, a result by a competing method, or both at the same time). In total, 27 respondents have participated.

There were different types of questions. In 16 questions (category I), the respondents were asked to decide which 3D reconstruction out of two looks more physically plausible to them (the first, the second or undecided). In 12 questions (category II), the respondents were asked to rate how natural the 3D reconstructed motions are or evaluate the degree of an indicated effect (foot sliding, body leaning, etc.) on a predefined scale. In five questions (category III), the respondents were also asked to decide which visualisation has a more pronounced indicated artefact. For two questions out of five, 2D projections onto the input 2D image sequence were shown, whereas the remaining questions in this category featured 3D reconstructions. Finally (category IV), the participants were encouraged to list which artefacts in the reconstructions seem to be most apparent and most frequent.

In category I, our reconstructions were preferred in 89.2% of the cases, whereas a competing method was preferred in 1.6% of the cases. Note that at the same time, the decision between the methods has not been made in 8.9% cases. In category II, the respondents have also found the results of our approach to be significantly more physically plausible than the results of competing methods. The latter were also found to have consistently more jitter, foot sliding and unnatural body leaning. In category III, noteworthy is also that the participants have indicated a higher average perceived accuracy of our reprojections, i.e., 32.7% voted that our results reproject better, whereas the choice felt on the competing methods in 22.6% of the cases. Note that the smoothness and jitter in the results are also reflected in the reprojections, and, thus, both influence how natural the reprojected skeletons look like. At the same time, a high uncertainty of 44.2% indicates that the difference between the re-projections of PhysCap and other methods is volatile. For the 3D motions in this category, 82.7% voted that our results show fewer indicated artefacts compared to other approaches, whereas 13.5% of the respondents preferred the competing methods. The decision has not been made in 3.7% of the cases. In category IV, 59% of the participants named jitter as the most frequent and apparent disturbing effect of the competing methods, followed by unnatural body leaning (22%), foot-floor penetration (15%) and foot sliding (15%).

The user study confirms a high level of physical plausibility and naturalness of PhysCap results. We see that also subjectively, a broad audience coherently finds our results of high visual quality, and the gap to the competing methods is substantial. This strengthens our belief about the suitability of PhysCap for computer graphics and primarily virtual character animation in real time.

6 DISCUSSION

Our physics-based monocular 3D human motion capture algorithm significantly reduces the common artefacts of other monocular 3D pose estimation methods such as motion jitter, penetration into the floor, foot sliding and unnatural body leaning. The experiments have shown that our state prediction network generalises well across scenes with different backgrounds (see Fig. 11). However, in the case of foot occlusion, our state prediction network can sometimes mispredict the foot contact states, resulting in the erroneous hard zero velocity constraint for feet. Additionally, our approach requires the calibrated floor plane to apply the foot contact constraint effectively; standard calibration techniques can be used for this.

Swift motions can be challenging for stage I of our pipeline, which can cause inaccuracies in the estimates of the subsequent stages, as well as in the final estimate. In future, other monocular kinematic pose estimators than [Mehta et al. 2017b] could be tested in stage I, in case they are trained to handle occlusions and very fast motions better. Moreover, note that – although we use a single parameter set for PhysCap in all our experiments (see Sec. 5) – users can adjust the quality of the reconstructed motions by tuning the gain parameters of PD controller depending on the scenario. By increasing the derivative gain value, the reconstructed poses are smoother, which, however, can cause motion delay compared to the input video, especially when the observed motions are very fast. By reducing the derivative gain value, our optimisation with a virtual character can track image sequence with less motion delay, at the cost of less temporally coherent motion. We demonstrate this trade-off in the supplemental video.
Further, while our method works in front of general backgrounds, we assume there is a ground plane in the scene, which is the case for most man-made environments, but not irregular outdoor terrains. Finally, our method currently only considers a subset of potential body-to-environment contacts in a physics-based way. As part of future work, we will investigate explicit modelling of self-collisions, as well as hand-scene interactions or contacts of legs and body in sitting and lying poses.

7 CONCLUSIONS
We have presented PhysCap – the first physics-based approach for a global 3D human motion capture from a single RGB camera that runs in real time at 25 fps. Thanks to the pose optimisation framework using PD joint control, the results of PhysCap evince improved physical plausibility, temporal consistency and significantly fewer artefacts such as jitter, foot sliding, unnatural body leaning and foot-floor penetration, compared to other existing approaches (some of them include temporal constraints). We also introduced new error metrics to evaluate these improved properties which are not easily captured by metrics used in the established pose estimation benchmarks. Moreover, our user study further confirmed these improvements. In future work, our algorithm can be extended for various contact positions (not only the feet).

REFERENCES

Fig. 11. Representative 2D reprojections and the corresponding 3D poses of our PhysCap approach. Note that, even with the challenging motions, our global poses in 3D have high quality and 2D reprojections to the input images are accurate as well. See our supplementary video for more results on these sequences. The backflip video in the first row is taken from [Peng et al. 2018]. Other sequences are from our own recordings.


Yuxiang Zhang, Liang An, Tao Yu, xiu Li, Kun Li, and Yebin Liu. 2020. 4D Association Graph for Realtime Multi-Person Motion Capture Using Multiple Video Cameras. In *International Conference on Computer Vision (ICCV)*.


Yuhang Zou, Jimei Yang, Duygu Ceylan, Jianming Zhang, Federico Perazzi, and Jia-Bin Huang. 2020. Reducing Footskate in Human Motion Reconstruction with Ground Contact Constraints. In *Winter Conference on Applications of Computer Vision (WACV)*.