AMASS: Archive of Motion Capture as Surface Shapes

Naureen Mahmood[1] Nima Ghorbani[2] Nikolaus F. Troje[3]
Gerard Pons-Moll[4] Michael J. Black[2]
[1]Meshcapade [2]MPI for Intelligent Systems [3]York University [4]MPI for Informatics
nmahmood@meshcapade.com, {nghorbani, black}@tue.mpg.de
troje@yorku.ca, gpons@mpi-inf.mpg.de

Abstract

Large datasets are the cornerstone of recent advances in computer vision using deep learning. In contrast, existing human motion capture (mocap) datasets are small and the motions limited, hampering progress on learning models of human motion. While there are many different datasets available, they each use a different parameterization of the body, making it difficult to integrate them into a single meta dataset. To address this, we introduce AMASS, a large and varied database of human motion that unifies 15 *different optical marker-based mocap datasets by representing them within a common framework and parameterization. We achieve this using a new method, MoSh++, that converts mocap data into realistic 3D human meshes represented by a rigged body model. Here we use SMPL [26], which is widely used and provides a standard skeletal representation as well as a fully rigged surface mesh. The method works for arbitrary markersets, while recovering soft-tissue dynamics and realistic hand motion. We evaluate MoSh++ and tune its hyperparameters using a new dataset of 4D body scans that are jointly recorded with marker-based mocap. The consistent representation of AMASS makes it readily useful for animation, visualization, and generating training data for deep learning. Our dataset is significantly richer than previous human motion collections, having more than* 40 *hours of motion data, spanning over* 300 *subjects, more than* 11000 *motions, and is available for research at* https://amass.is.tue.mpg.de/.

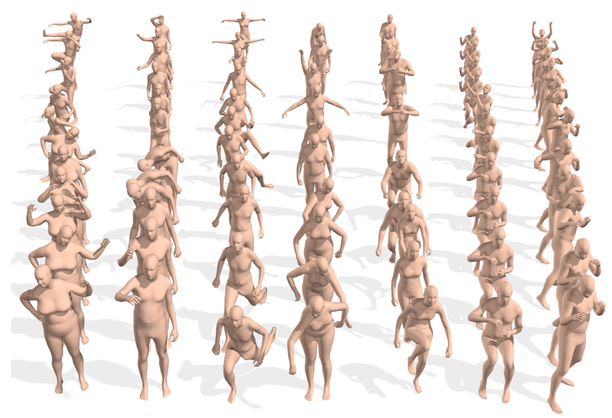

Figure 1: We unify a large corpus of archival marker-based optical human mocap datasets by representing them within a common framework and parameterization. A sampling of shapes and poses from a few datasets in AMASS is shown, from left to right: CMU [9], MPI-HDM05 [30, 31], MPI-Pose Limits [3], KIT [27], BMLrub [42], TCD [21] and ACCAD [34] datasets. The input is sparse markers and the output is SMPL body models.

1. Introduction

This paper addresses two interrelated goals. First, we develop a method to accurately recover the shape and pose of a person in motion from standard motion capture (mocap) marker data. This enables the second goal, which is to create the largest publicly available database of human motions that can enable machine learning for applications in animation and computer vision. While there have been attempts in both these directions, existing mocap databases are insufficient in terms of size and complexity to exploit the full power of existing deep learning tools. There are many different mocap datasets available, but pulling them together into a coherent formulation is challenging due to the use of widely varying markersets and laboratory-specific procedures [16]. We achieve this by extending MoSh [25] in several important ways, enabling us to collect a large and varied dataset of human motions in a consistent format (Fig. 1).

MoSh employs a generative model of the body, learned from a large number of 3D body scans, to compute the full 3D body shape and pose from a sparse set of motion capture markers. The results are realistic, but the method has several important limitations, which make it inappropriate for our task. First, MoSh relies on a formulation of the *SCAPE body model* [8], which is not compatible with

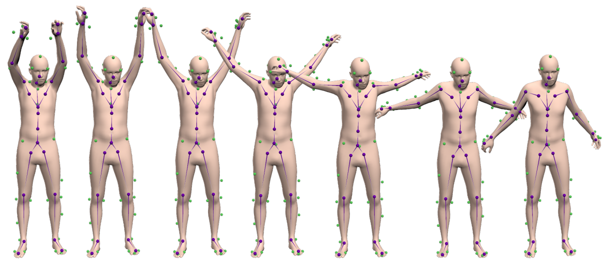

Figure 2: MoSh++ captures body shape, pose, and soft-tissue dynamics by fitting the surface of the SMPL/DMPL body model to observed mocap markers (green), while also providing a rigged skeleton (purple) that can be used in standard animation programs. Conventional mocap methods estimate only the skeleton, filtering out surface motion as noise and losing body shape information.

existing body representations and graphics software, making it a poor choice for distributing a dataset. We replace SCAPE with the SMPL body model [26], which uses a kinematic tree, has joints, and is based on blend skinning. SMPL comes with a UV map, which allows researchers to generate their own textures for rendering images and video sequences. SMPL is readily available, widely used, and compatible with most game engines and graphics packages. Second, while MoSh captures some soft-tissue motions, these are approximate and represented by changing the *identity* of a subject over time; that is, they are not true soft-tissue deformations. Here we take the dynamic shape space from DMPL, which models these soft-tissue deformations for SMPL [26] using a shape space learned from 4D scans of various subjects in motion. We show that we can recover the soft-tissue motions realistically from a sparse set of markers. The resulting body shapes and motions look natural and we show that they are metrically accurate. Third, MoSh does not solve for the pose and motion of the hands. Here we add the recent MANO hand model [37], which is compatible with SMPL, and solve for body and hand pose when hand markers are present. This provides richer and more natural animations. Fourth, to fine-tune and evaluate our proposed method, we collect a novel dataset, *SSM (Synchronized Scans and Markers)*, that consists of dense 3D meshes in motion, captured with a 4D scanner, together with traditional marker-based mocap. We separate the sequences into training and testing sets, and train the hyperparameters of MoSh++ to minimize the distance between the ground truth 3D scans and the estimated 3D body meshes. We then evaluate the performance of MoSh++ on the test set, demonstrating the accuracy of the method and allowing a quantitative comparison to MoSh.

MoSh++ enables our key goal of creating a large database of human motions. While there are many motion capture datasets available online for research purposes [3, 9, 10, 21, 25, 31, 39, 34, 42, 43], even the largest ones are too limited in size and variety to support serious deep learning models. Additionally, datasets vary in the format of the data and the kinematic structure of the body, making it hard for researchers to combine them. There have been several efforts to create data supersets [20, 27, 29], but the process of unifying the datasets typically means standardizing to fixed body proportions, which fundamentally alters the data. A good dataset should capture the articulated structure of the body in a way that is consistent with standard body models so that it can easily be adapted to new problems. Additionally, richness of the source marker data should be retained as much as possible. It should also be possible to produce high-quality animations that are realistic enough to train computer vision algorithms; that is, the dataset should include full 3D human meshes.

SMPL provides the unifying representation that is independent of the markerset, yet maintains the richness of the original marker data, including the 3D body shape. We know of no other attempt that provides access to full body shape and soft-tissue from mocap data, while also providing accurate body and hand pose. Here we combine 15 existing motion capture datasets into one large dataset: the *Archive of Mocap as Surface Shapes (AMASS)*. AMASS has 40 hours of mocap, 344 subjects, and 11265 motions. The source datasets all contain varying markersets ranging in size from 37 to 91 markers; AMASS unifies these into a single format. Each frame in AMASS includes the SMPL 3D shape parameters (16 dimensions), the DMPL soft-tissue coefficients (8 dimensions), and the full SMPL pose parameters (159 dimensions), including hand articulations, and body global translation. Users who only care about pose can ignore body shape and soft-tissue deformations if they wish. Similarly, the SMPL shape space makes it trivial to normalize all bodies to the same shape if users want joint locations normalized to a single shape. Figure 1 shows a selection of poses and body shapes in the dataset while Fig. 2 illustrates the difference between MoSh++ and traditional mocap. Traditional datasets contain skeletons and/or markers, while the AMASS dataset also provides fully rigged 3D meshes. With MoSh++ it is easy to add more data and we will continue to expand the dataset. We make AMASS available to the research community at `https://amass.is.tue.mpg.de/`, and will support the community in adding new captures as long as they can be similarly shared.

In summary, we provide the largest unified mocap dataset (AMASS) to the community, enabling new applications that require large amounts of training data.

2. Related Work

There is a vast literature on estimating skeletal parameters from mocap markers as well as several commercial solutions that solve this problem. As shown by Gorton et al. [16], different solutions use different skeletal models and pre-specified markersets, which makes it hard to unify the existing corpora of marker-based human recordings. Furthermore, all the methods that fit skeletons to data effectively lose rich surface information in the process. We review the most related work: fitting surface models to markers, capturing hands and soft-tissue motion from markers, and previous motion capture datasets.

Surface Models from Markers. To reconstruct bodies from markers, most methods first build a statistical model of body shape [5] or body shape and pose [6, 8, 26]. Allen et al. [5] reconstruct body shape using 74 landmarks. They do this only for a fixed body pose, assuming that the correspondences between the model and the markers are known. The approach cannot deal with arbitrary poses because the model cannot be posed. Anguelov et al. [8] go further by learning a model (SCAPE) of shape and non-rigid pose deformations. Their method requires a dense 3D scan of each subject. This restricts its application to archival mocap.

Loper et al. [25] address some of these limitations with MoSh, and remove the requirement for individual 3D dense scans. However, MoSh uses a BlendSCAPE body model formulation [18], which is not compatible with standard graphics packages making it sub-optimal for distribution. Furthermore, MoSh does not capture real soft-tissue dynamics, and does not capture hands.

Hands. There is a large body of work on fitting hand models to RGB-D data [40, 41] but here we focus on methods that capture hand motion from sparse markers. Maycock et al. [28] combine an optimal assignment method with model fitting but can capture only hands in isolation from the body and require a calibration pose. Schroder et al. [38] propose an optimization method to find a reduced sparse markerset and, like us, they use a kinematic subspace of hand poses. Alexanderson et al. [4] capture hand motion using sparse markers (3-10). They generate multiple hypotheses per frame and then connect them using the Viterbi algorithm [13]. They can track hands that exit and re-enter the scene and the method runs in real-time. However, a new model needs to be trained for every markerset. Han et al. [17] address the problem of automatically labeling hand markers using a deep network. The above methods, either do not estimate hands and bodies together or do not provide a 3D hand shape.

Soft-tissue motion. Most of the work in the mocap community focuses on *minimizing* the effect of skin deformations on the marker motions [7, 23]. In some biomechanical studies, the markers have even been fixed to the bones via percutaneous pins [22]. Our work is very different in spirit. We argue that such soft-tissue and skin deformation makes captured subjects look alive. In [25] they capture soft-tissue by fitting the parameters of a space of static body shapes to a sparse set of markers. This corresponds to modeling soft-tissue deformation by changing the *identity* of a person. Instead, using the dynamic shape space of DMPL [26] results in more realistic soft-tissue motions with minimal increase in model complexity.

Motion Capture Datasets. There are many motion capture datasets [3, 9, 10, 21, 25, 31, 30, 39, 34, 42, 43, 45], as well as several attempts to aggregate such datasets into larger collections [20, 27, 29]. Previous attempts to merge datasets [20, 27] adopt a common body representation in which the size variation among subjects is normalized. This enables methods that focus on modeling pose and motion in terms of joint locations. On the other hand, such an approach throws away information about how body shape and motion are correlated and can introduce artifacts in retargeting all data to a common skeleton. For example, Holden et al. [20] retarget several datasets to a common skeleton to enable deep learning using *joint positions*. This retargeting involves an inverse kinematics optimization that fundamentally changes the original data.

Our philosophy is different. We work directly with the markers and not the skeleton, recovering the full 3D surface of the body. There is no loss of generality with this approach as it is possible to derive any desired skeleton representation or generate any desired markerset from the 3D body model. Moreover, having a body model makes it possible to texture and render virtual bodies in different scenes. This is useful for many tasks, including generating synthetic training for computer vision tasks [44].

3. Technical Approach

To create the AMASS dataset, we generalize MoSh in several important ways: 1) we replace BlendSCAPE by SMPL to democratize its use (Sec. 3.1); 2) we capture hands and soft-tissue motions (Sec. 3.2); 3) we fine-tune the weights of the objective function using cross-validation on a novel dataset, SSM (Sec. 4).

3.1. The Body Model

AMASS is distributed in the form of SMPL body model parameters. SMPL uses a learned rigged template $\mathbf{T}$ with $N = 6890$ vertices. The vertex positions of SMPL are adapted according to identity-dependent shape parameters, $\boldsymbol{\beta}$, the pose parameters, $\boldsymbol{\theta}$, and translation of the root in the world coordinate system, γ. The skeletal structure of the human body is modeled with a kinematic chain consisting of rigid bone segments linked by joints. Each body joint has 3 rotational Degrees of Freedom (DoF), parametrized with exponential coordinates. We use a variant of SMPL, called SMPL-H [37], which adds hand articulation to the model using a total of $n = 52$ joints, where 22 joints are for the body and the remaining 30 joints belong to the hands. For simplicity of notation, we include the 3D translation vector

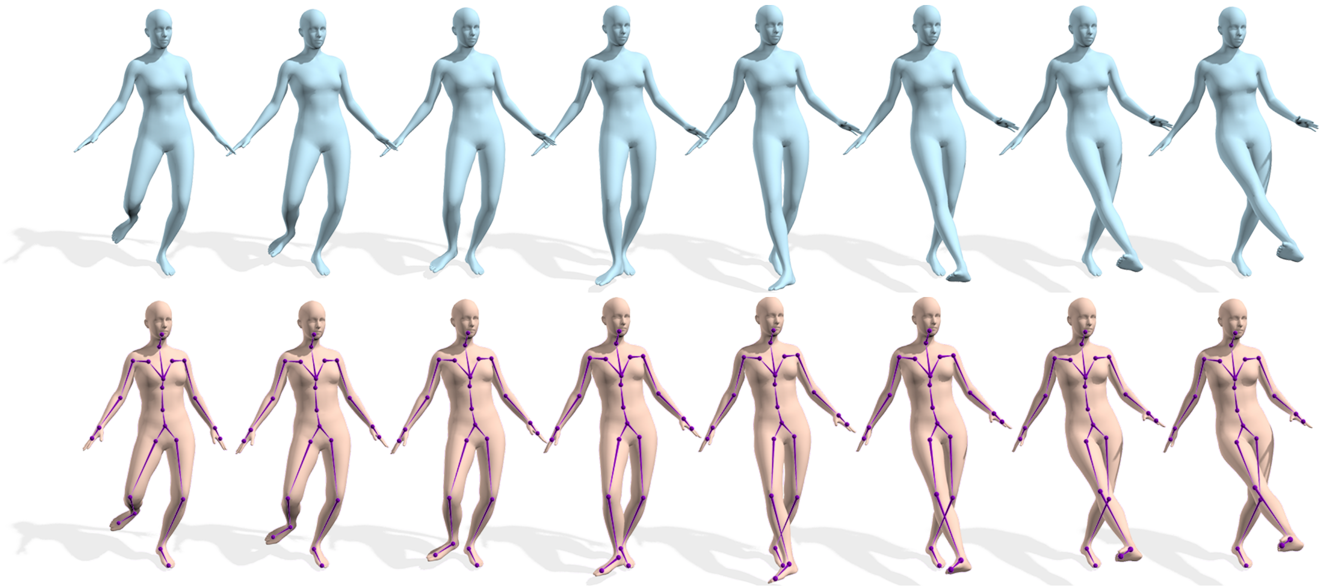

Figure 3: MoSh with BlendSCAPE (blue) vs. MoSh++ with SMPL (orange); visually similar, but MoSh++ is more accurate and SMPL provides a standard rigged mesh with a skeleton.

γ in the pose vector. The pose $\boldsymbol{\theta}$ is determined by a pose vector of $3 \times 52 + 3 = 159$ parameters. The remaining attributes of the SMPL-H model are the same as SMPL.

We combine SMPL-H with DMPL to obtain a model that captures both hand pose and soft-tissue deformations. For brevity we refer to the combined SMPL-H + DMPL model as SMPL throughout this paper, although this goes beyond any previously published model.

SMPL modifies the template in an additive way. It applies additive shape, pose, and dynamic blendshapes to a template in a canonical pose and predicts joint locations from the deformed surfaces. The model is

$$\begin{aligned} S(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}) &= G(T(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \quad (1) \\ T(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}) &= \mathbf{T}_\mu + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) + B_d(\boldsymbol{\phi}) \quad (2) \end{aligned}$$

where $G(\mathbf{T}, \mathbf{J}, \boldsymbol{\theta}, \mathbf{W}) : \mathbb{R}^{3N} \times \mathbb{R}^{|\theta|} \times \mathbb{R}^{3K} \times \mathbb{R}^{4 \times 3N} \mapsto \mathbb{R}^{3N}$ is a linear blend skinning function that takes vertices of the model in the rest pose $\mathbf{T}$, K joint locations stacked in $\mathbf{J}$, a pose $\boldsymbol{\theta}$, and the blend weights $\mathbf{W}$, and returns the posed vertices. The blendshape functions $B_s(\boldsymbol{\beta})$, $B_p(\boldsymbol{\theta})$, and $B_d(\boldsymbol{\phi})$ output vectors of vertex offsets relative to the mean template, $\mathbf{T}_\mu$ (see [26, 36] for a detailed explanation of the functions). We call these shape, pose, and dynamic blend shapes respectively. Note that the pose blendshapes are a function of the pose $\boldsymbol{\theta}$, while $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ correspond to linear coefficients that determine the shape and soft-tissue deformation.

SMPL captures the dimensionality of body space more compactly than BlendSCAPE. With only 16 shape, and 8 dynamics components, MoSh++ achieves better accuracy than MoSh using 100 shape components. The number of shape and dynamics coefficients is chosen using the SSM dataset such that MoSh++ does not over-fit to mocap markers (see Supplementary Material).

3.2. Model Fitting

Similar to MoSh [25], MoSh++ uses two stages to fit a body model to a sparse markerset. We summarize these stages, review the necessary details, and hightlight the differences relative to MoSh. We use a similar notation to the original MoSh paper.

Stage I: Following MoSh, we use a marker parametrization $m(\tilde{\mathbf{m}}_i, \boldsymbol{\beta}, \boldsymbol{\theta}_t)$ that maps a latent, pose invariant representation of the markers, $\tilde{\mathbf{m}}_i$, to estimate their position in a posed frame, $\boldsymbol{\theta}_t$. In the first stage, for $F = 12$ randomly chosen frames from the subject-specific mocap sequences, given an initial guess for marker-body correspondences, we optimize poses $\Theta = \boldsymbol{\theta}_{1\ldots F}$, a single shape $\boldsymbol{\beta}$, and latent marker positions $\tilde{\mathcal{M}} = \{\tilde{\mathbf{m}}_i\}$ to fit the observed marker locations $\mathcal{M} = \{\mathbf{m}_{i,t} \in \mathcal{M}_t\}_{1\ldots F}$, where i indexes the markers in a frame; at this stage we exclude soft-tissue deformations. More specifically, similar to MoSh, we optimize the following objective function:

$$\begin{aligned} E(\tilde{\mathcal{M}}, \boldsymbol{\beta}, \Theta_B, \Theta_H) &= \lambda_D E_D(\tilde{\mathcal{M}}, \boldsymbol{\beta}, \Theta_B, \Theta_H) \\ &+ \lambda_{\boldsymbol{\beta}} E_{\boldsymbol{\beta}}(\boldsymbol{\beta}) + \lambda_{\boldsymbol{\theta} B} E_{\boldsymbol{\theta}_B}(\boldsymbol{\theta}_B) + \lambda_{\boldsymbol{\theta} H} E_{\boldsymbol{\theta}_H}(\boldsymbol{\theta}_H) \\ &+ \lambda_R E_R(\tilde{\mathcal{M}}, \boldsymbol{\beta}) + \lambda_I E_I(\tilde{\mathcal{M}}, \boldsymbol{\beta}). \end{aligned} \quad (3)$$

The data term E_D measures distance between *simulated markers* $m(\tilde{\mathbf{m}}_i, \boldsymbol{\beta}, \boldsymbol{\theta}_t)$ and the observed ones $\mathbf{m}_{i,t}$; $E_{\boldsymbol{\beta}}$ is a Mahalanobis distance shape prior on the SMPL shape components; $E_{\boldsymbol{\theta}_B}$ regularizes the body pose parameters; E_R encourages the latent markers to remain a prescribed distance

d from the body surface (here we use an average value of $d = 9.5mm$); and E_I penalizes deviations of latent markers from their initialized locations defined by the markerset (see [25] for further details).

In addition to the original terms of MoSh in Eq. 3, we add $E_{\boldsymbol{\theta}_H}$, which regularizes the hand pose parameters. We project the full hand pose (i.e. 90 hand parameters) into the 24-D MANO pose space for both hands and compute the Mahalanobis distance in this space

$$E_{\boldsymbol{\theta}_H}(\boldsymbol{\theta}_H) = \hat{\boldsymbol{\theta}}_H^T \Sigma_{\boldsymbol{\theta}_H}^{-1} \hat{\boldsymbol{\theta}}_H, \quad (4)$$

where $\hat{\boldsymbol{\theta}}$ represents the projection of the pose and $\Sigma_{\boldsymbol{\theta}_H}$ is the diagonal covariance matrix of the 24-dimensional low-D PCA space [37].

In contrast to MoSh, the λ hyper-parameters are determined by line search on the training set of SSM (Sec. 4.2). The data term, E_D, in Eq. 3 uses a sum of squared distances, which is affected by the number of observed markers in the mocap data. This is noteworthy since a standard 46-markerset was used to determine the λ weights during the hyper-parameter search. To deal marker variation due to occlusion or using different markersets, we automatically adjust the weight of this term, scaling it by a factor, $b = 46/n$, where n is the number of observed markers in a frame.

To help avoid local optima while minimizing Eq. 3, we use the Threshold Acceptance method [11] as a fast annealing strategy. Over 4 annealing stages of graduated optimization, we increase λ_D by multiplying it by a constant factor $s = 2$ while dividing the regularizer weights by the same factor. The weights at the final iteration are as follows:

$$\lambda_D = 600 \times b, \lambda_{\boldsymbol{\beta}} = 1.25, \lambda_{\boldsymbol{\theta} B} = 0.375, \\ \lambda_{\boldsymbol{\theta} H} = 0.125, \lambda_I = 37.5, \lambda_R = 1e4. \quad (5)$$

The surface distance regularization weight, λ_R, remains constant throughout the optimization. The 24 hand pose components are added into the optimization only during the final two iterations.

Stage II: In this stage, the latent marker locations and body shape parameters $\boldsymbol{\beta}$ of the model are assumed constant over time and the objective at this stage optimizes pose for each frame of mocap in the sequence.

Like MoSh, we add a temporal smoothness term for pose changes, E_u, to help reduce the effect of jitter in the mocap marker data. Yet in contrast to MoSh, we optimize for the soft-tissue deformation coefficients, $\boldsymbol{\phi}$. We add a prior and a temporal smoothness terms, $E_{\boldsymbol{\phi}}(\boldsymbol{\phi})$ and $E_v(\boldsymbol{\phi})$ respectively, to regularize the soft-tissue deformations. Then the final objective function for this stage becomes

$$\begin{aligned} E(\boldsymbol{\theta}_B, \boldsymbol{\theta}_H, \boldsymbol{\phi}) = \lambda_D E_D(\boldsymbol{\theta}_B, \boldsymbol{\theta}_H, \boldsymbol{\phi}) \\ + \lambda_{\boldsymbol{\theta} B} E_{\boldsymbol{\theta}_B}(\boldsymbol{\theta}_B) + \lambda_{\boldsymbol{\theta} H} E_{\boldsymbol{\theta}_H}(\boldsymbol{\theta}_H) \\ + \lambda_u E_u(\boldsymbol{\theta}_B, \boldsymbol{\theta}_H) \\ + \lambda_{\boldsymbol{\phi}} E_{\boldsymbol{\phi}}(\boldsymbol{\phi}) + \lambda_v E_v(\boldsymbol{\phi}). \end{aligned} \quad (6)$$

The data, body, and hands pose prior terms, E_D, $E_{\boldsymbol{\theta}_B}$, and $E_{\boldsymbol{\theta}_H}$, are the same as described in the first stage. To regularize the soft-tissue coefficients, we add a Mahalonobis distance prior on the 8 DMPL coefficients.

$$E_{\boldsymbol{\phi}}(\boldsymbol{\phi}) = \boldsymbol{\phi}_t^T \Sigma_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}_t, \quad (7)$$

where the covariance $\Sigma_{\boldsymbol{\phi}}$ is the diagonal covariance matrix computed from the DYNA dataset [36].

When hand markers are present, MoSh++ optimizes the hand pose parameters in the same way as all the other pose parameters except that we use 24 dimensions of MANO's [37] low-dimensional representation of the pose for both hands. In cases where there are no markers present on the hands of the recorded subjects, the hand poses are set to the average pose of the MANO model.

The initialization and fitting for the first frame of a sequence, undergoes a couple of extra steps compared to the rest of the motion. For the first frame, we initialize the model by performing a rigid transformation between the estimated and observed markers to repose the model from its rest pose $\mathbf{T}$ to roughly fit the observed pose. Then we use a graduated optimization for Eq. 6 with only the data and body pose prior terms, while $\lambda_{\boldsymbol{\theta} B}$ is varied from $[10, 5, 1]$ times the final weight. Later, for each of the subsequent frames, we initialize with the solution of the previous frame to estimate the pose and soft-tissue parameters.

The per-frame estimates of dynamics and pose after the first frame are carried out in two steps. During the first step, we remove the dynamics and dynamics smoothness terms, and optimize only the pose. This prevents the dynamics components from explaining translation or large pose changes between consecutive frames. Then, we add the dynamics, $\boldsymbol{\phi}$, and the dynamics smoothness terms into the optimization for the final optimization of pose and dynamics.

We explain details of tuning the weights λ in Sec. 4.2. The velocity constancy weights λ_u and λ_v depend on the mocap system calibration and optical tracking quality, data frame rate, and the types of motions. Therefore, these values could not be optimized using just one source of data, so we empirically determined them through experiments on different datasets of varying frame rates and motions. The final weights determined for this stage are:

$$\lambda_D = 400 \times b, \lambda_{\boldsymbol{\theta} B} = 1.6 \times q, \lambda_{\boldsymbol{\theta} H} = 1.0 \times q, \\ \lambda_u = 2.5, \lambda_{\boldsymbol{\phi}} = 1.0, \lambda_v = 6.0. \quad (8)$$

Similar to b, which adjusts the weight of the data term to varying markersets, q is a weight-balancing factor for the pose prior $\lambda_{\boldsymbol{\theta}}$. During a mocap session, markers may get occluded by the body due to pose. If multiple markers of a particular body part are occluded simultaneously, the optimization may result in unreliable and implausible poses, such as the estimated pose shown in Fig. 4 (left). To address this, we introduce a coefficient $q = 1 + \left(\frac{x}{|\mathcal{M}|} * 2.5\right)$,

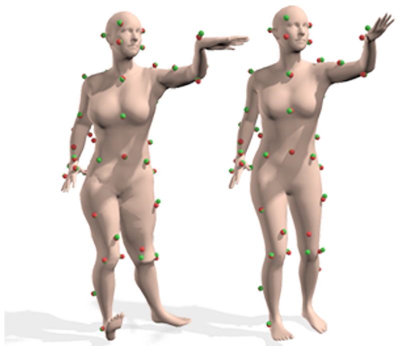

Figure 4: Pose estimation with heavy marker occlusion. Pose optimization with constant pose prior weight $\lambda_{\boldsymbol{\theta}}$ (left), variable pose prior weight $\lambda_{\boldsymbol{\theta}}$ (right). $\lambda_{\boldsymbol{\theta}}$ is allowed to vary as a factor of fraction of visible markers resulting in more plausible poses even when toe markers (right foot) and all foot markers (left foot) are missing. Estimated and observed markers are shown in red and green, respectively.

where x is the number of missing markers in a given frame, $|\mathcal{M}|$ are the total number of markers. This updates the pose prior weight as a factor of the number of missing markers. The more markers that are missing, the higher this weights the pose prior. This term can increase the prior weight by up to a factor of $q = 3.5$, in the worse case scenario where $x = |\mathcal{M}|$, and goes down to having no effect, $q = 1.0$ when all session markers are visible $x = 0$. An example of the effect of this factor is shown in Fig. 4 (right).

3.3. Optimization and Runtime

Similar to MoSh we use Powells gradient based dogleg minimization [33] implemented in the Chumpy [24] auto-differentiation package. Details on the runtime are presented in the Supplementary Material.

4. Evaluation

In order to set the hyperparameters and evaluate the time-varying surface reconstruction results of MoSh++, we need reference ground truth 3D data with variations in shape, pose and soft-tissue deformation. To that end, we introduce the SSM dataset (Sec. 4.1) and optimize the weights of the objective functions (Eqs. 3 and 6) using cross-validation on SSM (Sec. 4.2). After optimizing the hyper-parameters, we evaluate the accuracy of MoSh++, e.g. shape reconstruction accuracy (Sec. 4.3), pose, and soft-tissue motion reconstruction (Sec. 4.4) on the test set.

4.1. Synchronized Scans and Markers (SSM)

We use an OptiTrack mocap system [32] to capture subjects with 67 markers; i.e. using the *optimized marker-set* proposed by MoSh. The system was synchronized to record the mocap data together with a 4D scanning system [1].

Figure 5: SSM dataset. 3D scans with mocap markers (gray) and fitted bodies (orange). The average scan to model distance between them is $7.4mm$.

(See Fig. 5; details are provided in the Supplementary Material). The dataset consists of three subjects with varying body shapes, performing a total of 30 different motions. Two of the three subjects were professional models who signed modeling contracts; this allows us to release their 4D scan data, along with the synchronized mocap data for the research community.

We evaluate the accuracy of MoSh++ using the 67 markers, as well as a more standard 46 marker subset of the 67 markers. For both testing and evaluation, we use scan-to-model distances between the 3D scans (our ground truth mesh) of the SSM dataset and the corresponding estimated meshes for each trial of the hyper-parameter search and evaluation. For each reconstructed mocap frame, we take a uniform sampling of 10,000 points of the corresponding synchronized 3D scan and compute the distance from each of these to the closest surface point on our reconstructed mesh. We measure the average of these distances (in mm).

4.2. Hyper-parameter Search using SSM

The goal is to set the λ weights in Eq. 3 and Eq. 6 to minimize the reconstruction error for the validation data. Grid search complexity grows exponentially with the number of parameters (i.e. 5 parameters in the case of shape estimation, 4 in the case of pose estimation). Therefore, we perform line search on each parameter keeping the others fixed.

For the shape estimation stage, the optimization uses 12 randomly chosen mocap frames from each training subject to estimate shape and marker location for that subject. Instead of choosing a single, unseen pose to evaluate shape accuracy as in [25], we report the average error over the 12 randomly selected frames from the first stage of Mosh (see Sec. 3.2). Here the duration of the mocap sessions does not matter, but variation of body shape among the testing and training subjects is important. Therefore, we use only mocap data from two out of the three SSM subjects as training set while keeping the data from the third subject for testing and evaluation. We repeat the process 4 times for the training subjects, using a different random set of 12 frames for each trial. Validation is performed by running the optimization a fifth time, and initializing with a new randomization seed. We use a line search strategy to determine objective λ weights of Eq. 3 by finding a combination of these weights that provide the lowest reconstruction error for the estimated body mesh in the 12 frames picked during each trial. The final weights are described in Sec. 3.2.

For pose estimation, we separated 20% of the total captured mocap files from the three subjects as a held-out set for testing and evaluation. The first 200 frames of the rest of the motion files are used for training, leaving the remaining frames (roughly 60% of the training set) for validation. We perform a line search on the objective weights $[\lambda_D, \lambda_{\boldsymbol{\theta}}, \lambda_{\boldsymbol{\phi}}]$ of Eq. 6 and the missing-marker coefficient q, obtaining the final weights described in Sec. 3.2.

4.3. Shape Estimation Evaluation

Compared to MoSh, we obtain more accurate results on SSM. Fig. 6 (left) shows that the shape estimation accuracy on SSM is $12.1mm$ and $7.4mm$ for MoSh and MoSh++ respectively, when using a standard 46-markerset. Note that we use SSM to determine the optimal number of shape and dynamic coefficients (16 and 8 respectively). Adding more decreases marker error but this over-fits to the markers, causing higher error compared with the ground truth shape. Details are in the Supplementary Material.

4.4. Pose and Soft-tissue Estimation Evaluation

We also evaluate the per frame accuracy of pose and soft-tissue motion estimation of MoSh++. Fig. 6 (middle) shows that the pose estimation accuracy on SSM without soft-tissue motion estimation is $10.5mm$ and $8.1mm$ for MoSh and MoSh++ respectively, when using a standard 46-markerset. Similarly, with dynamics terms turned-on, MoSh++ achieves more accurate results than MoSh ($7.3mm$ vs $10.24mm$), Fig. 6 (right). The importance of soft-tissue estimation can be observed in Fig. 7. This result is expected since MoSh [25] models soft-tissue motion in the form of changes in the identity shape space of the BlendSCAPE model, whereas MoSh++ fits the DMPL space of soft-tissue motions learned from data [26].

4.5. Hand Articulation

We do not have ground-truth data for evaluating accuracy of hand articulation. Qualitative results of our joint body and hand captures can be seen in Fig. 8. Notice how MoSh++ with hand capture leads to more realistic hand poses. This illustrates that MoSh++ is not limited to the main body but can be extended to capture other parts if a model is available.

	Markers	Subjects	Motions	Minutes
ACCAD [34]	82	20	252	26.74
BMLrub [42]	41	111	3061	522.69
CMU [9]	41	96	1983	543.49
EKUT [27]	46	4	349	30.74
Eyes Japan [12]	37	12	750	363.64
HumanEva [39]	39	3	28	8.48
KIT [27]	50	55	4232	661.84
MPI HDM05 [31]	41	4	215	144.54
MPI Limits [3]	53	3	35	20.82
MPI MoSh [25]	87	19	77	16.53
SFU [15]	53	7	44	15.23
SSM (us)	86	3	30	1.87
TCD Hands [21]	91	1	62	8.05
TotalCapture [43]	53	5	37	41.1
Transitions (us)	53	1	110	15.1
Total		**344**	**11265**	**2420.86**

Table 1: Datasets contained in AMASS. We use MoSh++ to map more than 40 hours of marker data into SMPL parameters, giving a unified format.

5. AMASS Dataset

We *amassed* in total 15 mocap datasets, summarized in Table 1. Each dataset was recorded using a different number of markers placed at different locations on the body; even within a dataset, the number of markers varies. The publicly available datasets were downloaded from the internet. We obtained several other datasets privately or recorded them ourselves (Dancers, Transitions, BMLrub and SSM). We used MoSh++ to map this large amount of marker data into our common SMPL pose, shape, and soft-tissue parameters. Problems inherent with mocap, such as swapped or mislabeled markers, were fixed by manually inspecting the results and either correcting or holding out problems. Fig. 1 shows a few representative examples from different datasets. The result is AMASS, the largest public dataset of human shape and pose, including 344 subjects, 11265 motions and 40 hours of recordings and is available to the research community at https://amass.is.tue.mpg.de/. See the website for video clips that illustrate the diversity and quality of the dataset.

6. Future Work and Conclusions

Future work will extend the SSM dataset to include captures with articulated hands. We also intend to extend MoSh++ to work with facial mocap markers. This should be possible using the recently published SMPL-X model [35], which represents the face, body, and hands together. Current runtime for MoSh++ is not real-time (see Supplementary Material). However, in principle it should be possible to improve the runtime of MoSh++ significantly by using a parallel implementation of SMPL using frameworks such as TensorFlow [2]. Finally, we see an opportunity to push our

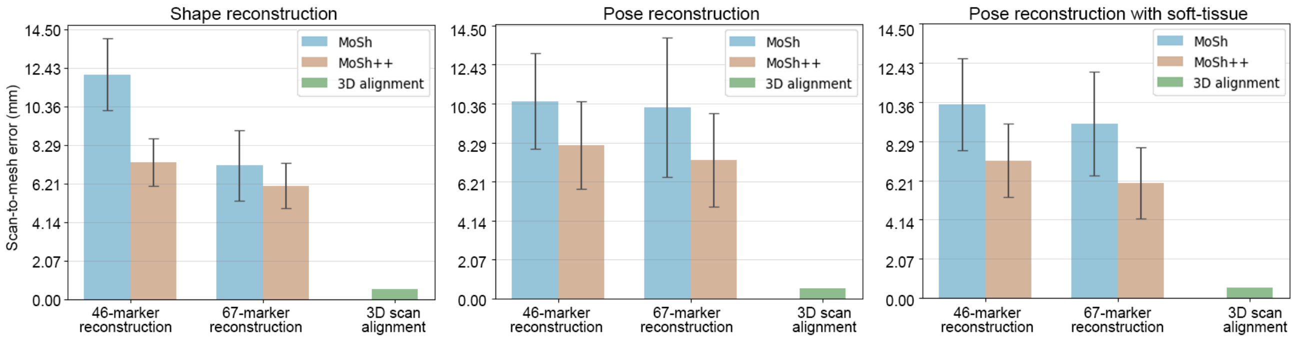

Figure 6: MoSh vs MoSh++ shape and pose reconstruction: Mean absolute distance of body shapes reconstructed, using MoSh with the BlendSCAPE model (blue bars) and MoSh++ with SMPL and optimized hyper-parameters (orange bars), to ground-truth 3D scans. Error in 1) Shape estimation, 2) Pose estimation, 3) Pose estimation with DMPL. Error bars indicate standard deviations. We compare a standard 46 markerset with the 67 markerset of MoSh [25]. MoSh++ with only 46 markers is nearly as good as MoSh with 67 markers. Average scan-to-mesh surface distance between 3D scan alignments and the original scans are shown in green as a baseline for comparison, e.g. an average value of $0.5mm$.

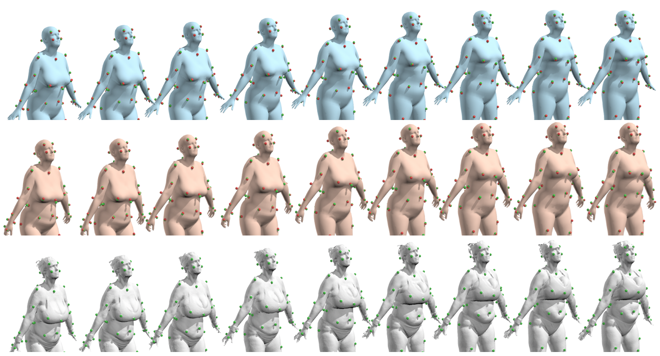

Figure 7: Soft-tissue Dynamics. MoSh [25] (blue), MoSh++ with dynamics from DMPL (orange), and ground truth scans synced with Mocap (gray). MoSh++ captures motion of the chest and stomach more accurately. Estimated markers (red) and observed markers (green) are also displayed for both MoSh and MoSh++.

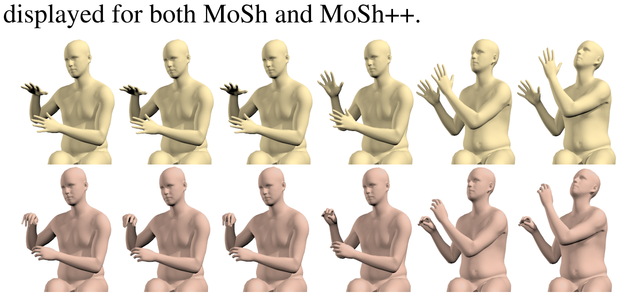

Figure 8: Articulated hands: If hand markers are present MoSh++ fits hand poses using SMPL-H [37]. Model fitting without hands (yellow) vs. MoSh++(orange).

approach further to address the problems of missing markers and to exploit the body for fully automatic marker labeling. AMASS itself can be leveraged for this task and used to train models that denoise mocap data [14] (cf. [19]).

In conclusion, we have introduced MoSh++, which extends MoSh and enables us to unify marker-based motion capture recordings, while being more accurate than simple skeletons or the previous BlendSCAPE version. This allowed us to collect the AMASS dataset containing more than 40 hours of mocap data in a unified format consisting of SMPL pose (with articulated hands), shape and soft-tissue motion. We will incorporate more mocap data into AMASS as it becomes available.

7. Acknowledgments

Foremost, we are grateful to the authors who made their mocap datasets available and allowed us to include them in AMASS. We also thank I. Abbasnejad and H. Feng for their support during early development phase and data collection, S. Polikovsky and A. Keller for help in project coordination, motion capture and 4D scanning, and T. Zaman, J. Romero, M. Loper and D. Tzionas for their invaluable advice, assistance, guidance and discussions. We thank M. Al Borno, J. Romero and A. Keller who contributed to the design and capture of Transitions dataset. NFT held an NSERC Discovery Grant in support of his research and was partially supported by the Humboldt Research Award from the Alexander-von-Humboldt Foundation. GPM has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 409792180 (Emmy Noether Programme, project: Real Virtual Humans), and by the MPI. Online datasets used in AMASS were created with funding from NSF EIA-0196217 (CMU [9]), NUS AcRF R-252-000-429-133 and SFU Presidents Research Start-up Grant (SFU [15]).

Conflict of Interest Disclosure: NM is a founder and shareholder of Meshcapade GmbH, which is commercializing body shape and motion technology; this work was performed primarily at the MPI. MJB has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at MPI. MJB is also an investor in Meshcapde.

References

[1] 3dMD LLC. 4D Scan. http://www.3dmd.com/. 6

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467 [cs]*, 2016. 7

[3] Ijaz Akhter and Michael J. Black. Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, 2015. 1, 2, 3, 7

[4] Simon Alexanderson, Carol O'Sullivan, and Jonas Beskow. Robust online motion capture labeling of finger markers. In *Proceedings of the 9th International Conference on Motion in Games*. ACM, 2016. 3

[5] B. Allen, B. Curless, and Z. Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Transactions on Graphics (TOG)*, 2003. 3

[6] Brett Allen, Brian Curless, Zoran Popović, and Aaron Hertzmann. Learning a Correlated Model of Identity and Pose-dependent Body Shape Variation for Real-time Synthesis. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '06, 2006. 3

[7] Thomas P Andriacchi and Eugene J Alexander. Studies of human locomotion: Past, present and future. *Journal of Biomechanics*, 2000. 3

[8] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape Completion and Animation of PEople. *ACM Transactions on Graphics*, 2005. 1, 3

[9] CMU Graphics Lab. CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/, 2000. 1, 2, 3, 7, 8

[10] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. *Robotics Institute*, 2008. 2, 3

[11] Gunter Dueck and Tobias Scheuer. Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing. *Journal of computational physics*, 1990. 5

[12] Eyes, JAPAN Co. Ltd. Eyes, Japan. http://mocapdata.com, 2018. 7

[13] G David Forney. The viterbi algorithm. *Proceedings of the IEEE*, 1973. 3

[14] Saeed Ghorbani, Ali Etemad, and Nikolaus F Troje. Auto-labelling of markers in optical motion capture by permutation learning. In *Computer Graphics International Conference*, pages 167–178. Springer, 2019. 8

[15] KangKang Yin Goh Jing Ying. SFU Motion Capture Database. http://mocap.cs.sfu.ca/. 7, 8

[16] George E Gorton, David A Hebert, and Mary E Gannotti. Assessment of the kinematic variability among 12 motion analysis laboratories. *Gait & posture*, 2009. 1, 3

[17] Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher D. Twigg, and Kenrick Kin. Online optical marker-based hand tracking with deep labels. *ACM Trans. Graph.*, 37(4):166:1–166:10, July 2018. 3

[18] David A Hirshberg, Matthew Loper, Eric Rachlin, and Michael J Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *European Conference on Computer Vision*, Lecture Notes in Computer Science. Springer, Springer Berlin Heidelberg, 2012. 3

[19] Daniel Holden. Robust Solving of Optical Motion Capture Data by Denoising. *ACM Transactions on Graphics*, 2018. 8

[20] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 2016. 2, 3

[21] Ludovic Hoyet, Kenneth Ryall, Rachel McDonnell, and Carol O'Sullivan. Sleight of Hand: Perception of Finger Motion from Reduced Marker Sets. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D '12. ACM, 2012. 1, 2, 3, 7

[22] MA Lafortune, PR Cavanagh, HJD Sommer, and A Kalenak. Three-dimensional kinematics of the human knee during walking. *Journal of biomechanics*, 1992. 3

[23] Alberto Leardini, Lorenzo Chiari, Ugo Della Croce, and Aurelio Cappozzo. Human movement analysis using stereophotogrammetry: Part 3. Soft tissue artifact assessment and compensation. *Gait &/ Posture*, 2005. 3

[24] Matthew Loper. Chumpy. https://github.com/mattloper/chumpy, 2013. 6

[25] Matthew Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and Shape Capture from Sparse Markers. *ACM Trans. Graph.*, 33(6):220:1–220:13, Nov. 2014. 1, 2, 3, 4, 5, 6, 7, 8

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015. 1, 2, 3, 4, 7

[27] Christian Mandery, Oemer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The KIT Whole-Body Human Motion Database. In *International Conference on Advanced Robotics (ICAR)*, 2015. 1, 2, 3, 7

[28] Jonathan Maycock, Tobias Rohlig, Matthias Schroder, Mario Botsch, and Helge Ritter. Fully automatic optical motion tracking using an inverse kinematics approach. In *Humanoid Robots (Humanoids), 2015 IEEE-RAS 15th International Conference On*. IEEE, 2015. 3

[29] Motion Capture Club. MocapClub. http://www.mocapclub.com/, 2009. 2, 3

[30] Meinard Müller, Andreas Baak, and Hans-Peter Seidel. Efficient and Robust Annotation of Motion Capture Data. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, 2009. 1, 3

[31] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation Mocap Database HDM05. Technical report, Universität Bonn, 2007. 1, 2, 3, 7

[32] NaturalPoint, Inc. Motion Capture Systems. https://optitrack.com/. 6

[33] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006. 6

[34] OSU ACCAD. ACCAD. https://accad.osu.edu/research/motion-lab/system-data. 1, 2, 3, 7

[35] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. Osman, D. Tzionas, and M.J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *CVPR*, 2019. 7

[36] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A Model of Dynamic Human Shape in Motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 2015. 4, 5

[37] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017. () Two first authors contributed equally. 2, 3, 5, 8

[38] Matthias Schröder, Jonathan Maycock, and Mario Botsch. Reduced marker layouts for optical motion capture of hands. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*. ACM, 2015. 3

[39] L. Sigal, A. 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*, 2010. 2, 3, 7

[40] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics (TOG)*, 2016. 3

[41] Anastasia Tkach, Andrea Tagliasacchi, Edoardo Remelli, Mark Pauly, and Andrew Fitzgibbon. Online generative model personalization for hand tracking. *ACM Transactions on Graphics (TOG)*, 2017. 3

[42] Nikolaus F Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2002. 1, 2, 3, 7

[43] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In *BMVC17*, 2017. 2, 3, 7

[44] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from Synthetic Humans. In *CVPR*, 2017. 3

[45] Christoph von Laßberg, Walter Rapp, Betty Mohler, and Jürgen Krug. Neuromuscular onset succession of high level gymnasts during dynamic leg acceleration phases on high bar. *Journal of Electromyography and Kinesiology*, 2013. 3