Markerless Multiview Motion Capture with 3D Shape Model Adaptation

P. Fechteler, A. Hilsmann and P. Eisert
Fraunhofer HHI & HU Berlin, Berlin, Germany
{philipp.fechteler, anna.hilsmann, peter.eisert}@hhi.fraunhofer.de

Abstract
In this paper, we address simultaneous markerless motion and shape capture from 3D input meshes of partial views onto a moving subject. We exploit a computer graphics model based on kinematic skinning as template tracking model. This template model consists of vertices, joints and skinning weights learned a priori from registered full-body scans, representing true human shape and kinematics-based shape deformations. Two data-driven priors are used together with a set of constraints and cues for setting up sufficient correspondences. A Gaussian mixture model-based pose prior of successive joint configurations is learned to soft-constrain the attainable pose space to plausible human poses. To make the shape adaptation robust to outliers and non-visible surface regions and to guide the shape adaptation towards realistically appearing human shapes, we use a mesh-Laplacian-based shape prior. Both priors are learned/extracted from the training set of the template model learning phase. The output is a model adapted to the captured subject with respect to shape and kinematic skeleton as well as the animation parameters to resemble the observed movements. With example applications, we demonstrate the benefit of such footage. Experimental evaluations on publicly available datasets show the achieved natural appearance and accuracy.

Keywords: computer vision - tracking, methods and applications, geometric modelling, modelling, motion capture, animation
ACM CCS: Computing methodologies → Motion capture; Shape analysis

1. Introduction
Motion and performance capture as well as human modelling techniques have revolutionized modern movie and game productions as well as telecommunication scenarios. There is a clear demand for 3D capturing of performing humans and the ability to manipulate such footage manually/automatically, e.g. by exchanging the appearance model while keeping the motions (retargeting), assembling certain movements into a desired sequence, fine-tuning certain actions (e.g. gaze correction), etc.

Inspired by this demand, the work presented here provides an approach to automatically model subject-specific human geometry (shape vertices as well as kinematic skeleton) while tracking the subjects motion (Figure 1). Adaptation of the template model towards the observed subject is required to accurately resemble the captured input sequence. The surface of the human tracking model is based on skinning techniques and thereby provides a typical, easy-to-manipulate kinematic skeleton control structure. A fully rigged model of the subject is learned and can be animated in various ways, e.g. slight corrections of the input sequence as well as completely new movements. Our motion capture and shape adaptation framework processes an input sequence of noisy meshes, which represent a partial view onto a moving subject. As a result, an animation of a topologically consistent template model is generated which closely resembles the input sequence. Alternatively, the proposed optimization framework is well suited to create a kinematically animatable human model of a subject from a set of 3D scans of static poses. The main challenges are the inconsistency of vertices respectively, triangle topology, the noise level obscuring the details in the captured data as well as that the visibility is of only partial views onto the subject so that considerable surface areas of the subject are not always observed and need to be inferred in a plausible manner.

2. Related Work
Markerless motion capture is an ill-posed problem because of occlusions, non-visible surface areas, high degree of freedom (DoF) and the fact that multiple different model configurations produce
similar observations. Typically, a generic model (or even just a broad approximation, e.g. set of cylinders) of a human is used, which simplifies the estimation of motion data. However, such a rough model usually lacks the personal characteristics and consequently cannot explain certain observable peculiarities. On the other side, modelling very realistic human models has become relatively simple with nowadays modelling software, but modelling an existing real person in detail is still a very cumbersome task.

This has been considered in the last years by several researchers and has led to approaches combining modelling with tracking. Quite thorough overviews of the literature can be found in [JDKL14, Bla16, CCM*18]. In the following, we point out some approaches closely related to our motion capture.

**Animation and modelling**

Animation is the process of transforming the vertices of a model in rest pose to a desired target pose. One approach is to specify certain conditions the result should comply to and transform vertices by optimizing the error function with the use of optimization or similar techniques. A frequently used method is to control the orientation of triangles [ASK*05, HSS*09, PMRMB15, BBLR15, LMR*15, KPMP*17] or to employ a probabilistic approach [ZB15]. This flexible formulation supports very realistic animations. However, solving for all vertices at once produces usually considerable computational costs.

In contrast, skinning-based methods provide the capability of generating or manipulating animation sequences. Linear blend skinning (LBS) is dominantly used throughout academic research as well as industrial applications [MG03, LD12, LD14, CLC*13] because of its simplicity and wide spread support in graphics hardware. However, direct vertex animations based on linear skinning typically suffer from severe artefacts, e.g. candy-wrapper, as shown in Figure 2. Various non-linear approaches have been investigated to mitigate such negative effects [KCV08, KS12, FHE16].

In order to model human characters effortlessly in a realistic and detailed manner, data-driven methods have been developed. Such statistical models are generated with the use of Principal Component Analysis (PCA) to extract a parametric model from a set of training data consisting of registered example scans. By varying the eigenvalues, the shape of the resulting model is adjusted according to the statistical dominant variations of the training data. Deformations can either be represented by vertex displacements [JTST10, PWH*17, LMR*15, RTB17, JSYS18] or triangle deformations [ASK*05, ZFL*10, HTRS10, WHB11, HLRB12, ZFYY14, ZYZD16].

**Motion capture and shape adaptation** Motion capture aims at recovering the frame-by-frame pose configuration of a human captured in a temporally sequential manner. Marker-based methods require the subject to accurately wear placed markers (active or passive) or clothes with certain patterns. In contrast, markerless motion capture directly exploits the captured image data. The omission of any subject preparation allows for example to also extract texture maps. Several of newer approaches in motion capture exploit machine learning techniques to infer pose information from sensor
input [BS10, YY14, YSD*16]. The recent advancements in deep learning motivated the use of artificial neural networks for motion capture purposes [BKL*16, PIT*16, RLGK16, CSWS17, MSS*17, EdAJ*17, KBJM18].

A very recent trend in motion capture is to simultaneously adapt a geometric template model with individual geometric details towards the captured subject during tracking. In [BBLR15], a low-resolution model is fitted with respect to shape and pose to the depth channel of a monocular RGB-D sequence in order to initialize a registered high-resolution model. Then, the initialized high-resolution model is refined to match the input frames by exploiting photo-consistency and thereby simultaneously calculating a texture map. Finally, a displacement map is calculated encoding the remaining geometric details. An approach to process sequences of full-body 3D scans is presented in [ZPBPM17]. A statistical model, enhanced with a term to compensate for cloth deviations, is initially fitted to an input sequence. Based on the per input frame adjusted statistical model, a free-form surface is optimized to better resemble the input sequence. This free-form surface is coupled to the statistical model, in order to prevent too strong deviations from it.

In [YZG*18], a human model representation is presented consisting of two layers that support model adaptation and tracking in real-time on a single depth input stream. The inner layer is a parametric shape model encoding the subject identity and pose without cloth or other disturbances. An outer layer, which is progressively refined and extended with the input frames, is used to model everything observable, like cloth, hair, etc., to support free-form deformations. Theses layers are coupled through a node graph, so that the kinematic skeleton structure of the inner layer constrains the deformation of the outer layer. In [AMX*18b, AMX*18a], the authors enhance a statistical model with an auxiliary term to support individual geometric peculiarities. Using this enhanced statistical model, a geometrically individualized and textured model of a subject is generated from a monocular RGB image sequence by using robust silhouette matching in a canonical reference frame. Further fine geometric details are included into the individualized statistical model by means of additional vertices, which are initially placed onto the already existing triangles. Using shape-from-shading techniques, these additional vertices are moved along the triangle normal direction in order to optimally resemble the captured input.

### Contribution of proposed approach

The proposed method falls in the latter category of algorithms: detailed model adaptation together with pose optimization. The contribution of this paper is the following: A motion capture framework with shape adaptation is presented, focusing on processing sequences of noisy 3D reconstructions captured with a multi-view setup covering only partial views of a moving subject. Alternatively, the method supports the generation of a subject-adapted model from sets of 3D reconstructions representing independent poses.

In contrast to indirect animation approaches, the underlying tracking model used here is based on a recent compact, vertex-based, kinematic skinning function [FHE16] and thus can be directly animated. Since the model is optimized with respect to surface vertices as well as skeleton joints, the result closely resembles the captured world data. The output is a kinematically animatable model resembling the shape of the captured subject as well as the animation parameters extracted from the input sequence.

This is achieved by applying several heuristics to tackle the challenge of processing such imperfect real-world input data, such as outlier filtering, constraining the visibility of vertices, using a robust norm, etc. Further, prior knowledge extracted a priori from preprocessed examples is used to guide the optimization towards realistic results. By using a data-efficient Gaussian mixture model (GMM) formulation as prior on valid joint configurations, we reduce the risk of unlikely poses.

In contrast to statistical models with their limited parametric space of attainable shapes, we support free-form deformations to adapt the template model’s surface to the details of the captured subject. Soft-constraints, learned a priori from example data, are used to ensure realistic human shapes without compromising subject-specific peculiarities, e.g. like a pony tail hair style.

### 3. Model Adaptive Pose Estimation and Tracking

The proposed method calculates a kinematically animatable human model from 3D reconstructions of a continuously moving subject, although a set of independent poses is also suitable. The input data are given as one noisy, non-registered 3D mesh per time instant, each representing a partial view onto the subject.

The template tracking model is adapted with respect to shape vertices as well as skeletal joints to closely resemble the observed subject in all captured pose configurations. Since optimizing shape and pose independently resembles a chicken-and-egg problem [HLRB12], we jointly optimize for all parameters simultaneously: pose variables of each input frame (global translation and rotation and kinematic rotations for each skeletal joint) and subject-specific model/shape parameters for the complete sequence (template vertices and skeletal joints). For accelerated convergence, the optimization is divided into two steps: given a generic template model, the pose parameters for each input frame are optimized first, followed by jointly adapting the model towards the captured input data in conjunction with refining the pose parameters. Further, in order to prevent convergence to undesired model configurations (unnatural poses and/or models), prior knowledge, learned a priori from training data, is used to guide the optimization towards plausible poses and humanoid models.

#### 3.1. Template model for tracking

Our framework is based on a skinned humanoid template model with fixed-topology used for tracking partial 3D reconstructions with inconsistent topology. This, on a kinematic skinning function-based template model, is set up only once. In order to capture true human characteristics, the template model is generated with a data-driven optimization using registered 3D scans. This template model is adapted during tracking of the captured subject with respect to kinematic pose as well as shape including skeletal joints.

We chose a computer graphics model based on a kinematic skinning function as template tracking model [FHE16] – in contrast to complex physiology simulation-based models like [LST09]. As
shown in Figure 2, typical skinning functions are prone to undesired artefacts, e.g. LBS suffers from candy-wraper for twist motions and dual quaternion skinning (DQS) suffers from bulging artefacts in swing motions.

We use a combination of skinning functions that drastically reduces these artefacts, as shown in [KS12]: LBS for swing motions and DQS for twist motions. As a consequence, each joint rotation needs to be decomposed into its 2-DoF-swing- and 1-DoF-twist-components [PB12]. This implies that each joint has its uniquely oriented coordinate system in which the joint rotation takes place. In order to increase the expressivity flexibility of the skinning function, we use dual quaternion linear blending (DLB) [KCvO08] instead of DQS, which supports the blending of multiple joint rotations. This allows vertices to be influenced by more than one joint and thereby better approximates the true behaviour (Figure 2).

We denote the complete pose aligning transformation by \( T() \) and DLB/LBS-based skinning interpolation by \( S() \). Rigid transformations representing the transformations of all previous joints in the kinematic chains, denoted by \( K() \), and global similarity alignment by \( \Phi() \). Hence, the full vertex transformation is given by

\[
v' = T(v) = \Phi(K(S(v))),
\]

with \( v \in \mathbb{R}^3 \) representing a vertex of the template model in rest pose and \( v' \) being the resulting vertex of the animated template model.

In our blending function \( S() \) of joint transformations, a model vertex \( v \) in rest pose is first transformed by the vertex-specific DLB-combination of twist rotations of kinematic joints. The twist transformation is followed by the swing transformation using LBS

\[
S(v) = \text{LBS}(\text{DLB}(v)).
\]

Using the \( x \)-axis of the joint’s local coordinate system as axis to twist around, the unit quaternion \( q = [q_w, q_x, q_z, q_y] \) representing a joint rotation is decomposed into its swing and twist components:

\[
q^\text{twist} = \frac{[q_w, q_x, 0, 0]}{\sqrt{q_w^2 + q_x^2}} \quad \text{and} \quad q^\text{swing} = q \cdot q^\text{twist}^*,
\]

with \( q = q^\text{swing} \cdot q^\text{twist} \) and \( q^* \) denoting quaternion conjugation.

Representing a joint’s local orientation with quaternion \( o \), the joint rotations combined within DLB are given as \( o \cdot q^\text{twist} \cdot o^* \). These quantities are augmented with a dual quaternion part representing the joint rotation centre to form \( q^\text{twist} \) [DKL98]. According to DLB, a vertex-specific twist transformation is calculated by summing up the twist dual quaternions, each weighted with its vertex-specific DLB skinning weight \( w^\text{twist}_j \), with subsequent normalization:

\[
q^\text{DLB} = \text{DLB} \left( w^\text{twist}_1 \cdot q^\text{twist}_1, \ldots, w^\text{twist}_j \cdot q^\text{twist}_j \right)
= \frac{\sum_j w^\text{twist}_j \cdot q^\text{twist}_j}{||\sum_j w^\text{twist}_j \cdot q^\text{twist}_j||}.
\]

The resulting DLB blended unit dual quaternion \( q^\text{DLB} \) represents the combined twist rotation of all involved joints and is used to transform a rest pose vertex \( v \) into \( v^\text{DLB} \).

The following swing transformation acts on the twisted transformed vertex \( v^\text{DLB} \). For this purpose, the oriented swing components \( o^*q^\text{swing} \) of all involved joints are converted into rotation matrices \( R^\text{swing}_j \) and blended together with the joint rotation centres \( t \) according to standard LBS:

\[
S(v) = \text{LBS}(v^\text{DLB}) = \sum_j w^\text{swing}_j \left( R^\text{swing}_j(v^\text{DLB} - t_j) + t_j \right).
\]

Having applied the interpolated twist/swing transformation, joint rotations of the remaining kinematic chain are applied \( K() \), followed by global rigid transformation \( \Phi() \) to locate/orient the model.

The template model based on this skinning function consists of the mesh itself, kinematic joints as well as per vertex and joint skinning weights for LBS and DLB. Similar to our previous work [FHE16], we optimize all these three constituents to closely approximate an example real-world dataset of registered scans of a single person in various poses to fit the shape and pose of the captured person (we use SCAPE [ASK*05] with \( M = 71 \) meshes, shown in Figure 3). Since this optimization scheme is fully data driven, we achieve a very natural appearing human model. In comparison to our previous approach [FHE16], we added a heavily weighted symmetry soft-constraint to enforce that corresponding bones on the left- and right-hand-side have the same length. Additionally, we enhance the deformation characteristics of resulting animations by using a block coordinate descent algorithm [JDCL14] for the skinning weight optimization, instead of simple coordinate descent. Finally, the flexibility of the model is improved by increasing the number of joints from \( J = 15 \) to \( J = 18 \) (Figure 3).

### 3.2. Shape adaptive motion capture

Our optimization framework processes meshes, 3D reconstructed from multi-view setups covering a partial view of the working volume where a subject is captured (one mesh for each captured multi-view frame). The key idea is to track the subjects pose for each frame by kinematically fitting the template model in an iterative closest point manner in combination with adapting the template models shape and skeleton to the 3D reconstructed meshes. To achieve this successfully for noisy and incomplete real-world data, prior knowledge on plausible shape and poses is integrated as soft-constraints.

#### 3.2.1. Heuristics to set up meaningful vertex correspondences

For motion capture, an objective function \( Q \) (Equation (6)) is minimized based on vertex correspondences between the template tracking model and each input frame. Setting up meaningful correspondences for a sequence of independent 3D reconstructions is challenging [vKZHCO11], because the number of captured vertices and their connecting mesh topology varies (in contrast to the input data used to learn the base template model in Section 3.1). Additionally, real-world 3D reconstructions may contain strong noise as well as 3D reconstruction artefacts and may cover only a pose dependent partial area of the subjects surface.

Finding correspondences using simple nearest neighbour search, e.g. exploiting kd-trees, is not sufficient. In order to find sufficiently good correspondences, the following additional heuristics have been integrated:

© 2019 The Authors. Computer Graphics Forum published by Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd
Visibility constraint Adopting the idea from [HBB*13], we use the external calibration data to limit the set of potential correspondences among the vertices of the template model per frame to the ones which are visible by any of the capturing sensors, ignoring all template vertices that are facing backwards or are occluded with respect to all cameras.

Normal constraint The nearest neighbour search ignores matching candidates if the angle between their normals exceeds the threshold of 90°. This drastically reduces the risk that vertices of close but opposite surface areas are matched.

2-Way matching Correspondences are set up by finding for each vertex in the input mesh a correspondence within the visible template mesh and vice versa. This principle enforces that the visible template mesh and input mesh should cover each other, in contrast to only one being covered by the other one.

Uniform outlier removal In order to reduce the negative influence of strong noise, the correspondences with the largest distances are skipped and not used for optimization.

Every iteration of pose optimization or model parameter adaptation is based on an up-to-date set of vertex correspondences. Consequently, the first step of any optimization is the registration of the input frame(s) with the template model. Based on these vertex correspondences, one iteration of optimizing pose and/or model parameters takes place. Then, vertex correspondences are recalculated, prior to using them again for the next iteration.

3.2.2. Setting up a suitable objective function \( Q \)

The proposed model adaptive pose estimation is treated as an optimization problem. The objective \( Q \) to be minimized in order to find optimal parameters for pose, shape, and skeleton is the sum of several objective terms, each introduced to enforce compliance to a certain sub-criteria:

\[
Q = Q_{\text{data}} + w_{\text{pose}} Q_{\text{pose}} + w_{\text{temp}} Q_{\text{temp}} + w_{\text{shape}} Q_{\text{shape}} + w_{\text{sym}} Q_{\text{sym}},
\]

The scalar-valued, non-negative weights \( w_{\text{pose}} \), \( w_{\text{temp}} \), \( w_{\text{shape}} \), \( w_{\text{sym}} \) are used to weight the influence of their corresponding objective terms \( Q_{\text{pose}}, Q_{\text{temp}}, Q_{\text{shape}}, Q_{\text{sym}} \) in relation to the data term \( Q_{\text{data}} \).

For each input frame, one set of pose parameters is optimized, consisting of global similarity and kinematic joint rotations. The subject-specific rest pose shape vertices and skeletal joints are optimized commonly for all input frames.

Data term \( Q_{\text{data}} \) The objective term \( Q_{\text{data}} \) enforces that corresponding vertices of template model and input meshes are close to each other for all input frames \( N \). It is formulated as a normalized sum over all \( C_n \) vertex correspondences independently for each input frame \( n \):

\[
Q_{\text{data}} = \frac{1}{N} \sum_{n=1}^{N} Q_{\text{data}}^n \quad \text{with}
\]

\[
Q_{\text{data}}^n = \frac{1}{C_n} \sum_{c=1}^{C_n} \left| v_{\text{rec}}^{c,n} - T_n^{\text{templ}}(v_{\text{rec}}^{c,n}) \right|^r,
\]

with \( v_{\text{rec}}^{c,n} \) being the \( c \)-th correspondence vertex of the 3D reconstruction in input frame \( n \), \( v_{\text{rec}}^{c,n} \) being the template vertex corresponding to \( v_{\text{rec}}^{c,n} \), \( T_n^{\text{templ}} \) being the pose alignment transformation (Equation (1)), parameterized to map the template model from rest pose to the pose of input frame \( n \).

In order to make \( Q_{\text{data}}^n \) more robust towards outliers, a robust norm is used, i.e. the Charbonnier norm [SRB10]:

\[
|x|^r = \sqrt{x^2 + \varepsilon^2},
\]

with \( \varepsilon \) being a small constant, e.g. \( 10^{-6} \). This Charbonnier norm can be interpreted as a continuous approximation of the L1 norm. Minimizing this robust norm can be implemented by using an iteratively reweighted least squares scheme [Zha95].

This data term \( Q_{\text{data}}^n \) is influenced by any parameter change and needs to be considered in any optimization case (in contrast to, e.g. \( Q_{\text{sym}}^n \), which needs to be considered only if the skeletal joint locations are optimized).
Objective term $Q^\text{pose}$ to enforce plausible poses

Estimating the pose of a human from depth information imposes the challenge, that ambiguities need to be resolved. Twist rotations of cylinder-like shaped limbs cannot be determined from depth information because of their almost symmetric nature. Unnatural poses might be the minimum of objective $Q$, e.g. self-intersection and unnatural joint angles. Further, missing views onto surface areas (due to occlusion) can result in an under-determined equation system to optimize (singular matrix).

Consequently, a prior term $Q^\text{pose}$ for plausible pose configurations is introduced. We set up a likelihood function of joint angle configurations, which is maximized within the objective $Q$. We exploit the assumption, that rotations of neighbouring joints are stronger correlated than rotations of joints further apart in the kinematic chain. Hence, we set up a data-efficient likelihood function by using the GMM [Jae16] for each pair of successive joint rotations. Thus, $P$ pairs of consecutive kinematic joints result in $P$ GMMs. Since we use the 3D exponential map [Ude98] as representation for rotations, the employed GMM probability function $\mathcal{P}_p(o)$ for each pair of successive joints is defined on the six-dimensional space.

According to the GMM framework, the probability density function $\mathcal{P}_p(o)$, consisting of $|\mathcal{P}_p|$ weighted Gaussian distributions $\mathcal{N}(\cdot)$, is formulated as follows:

$$\mathcal{P}_p(r_{\text{pre}}, r_{\text{suc}}) = \sum_{c=1}^{|\mathcal{P}_p|} \pi_c \cdot \mathcal{N}\left(\begin{bmatrix} r_{\text{pre}} \\ r_{\text{suc}} \end{bmatrix} | \mu_c, \Sigma_c \right).$$  

(8)

with $r_{\text{pre}}$ and $r_{\text{suc}}$ representing the rotations of two successive joints, $\pi_c$ being positive mixing weights summing up to one and $\mu_c$ (6D mean vector) and $\Sigma_c$ ($6 \times 6$ covariance matrix) representing the model parameters of mixture component $c$ of the $p$’s GMM.

Within the GMM learning phase, we set the off-diagonal values in the precision matrix (=inverse covariance matrix) to zero if their correlation turns out to be too small. Because the precision matrix needs to be positive semi-definite [RW06], zeroing is not always possible. In such cases, we lower the respective values in magnitude as long as the matrix remains positive semi-definite, which can be checked by inspecting the eigenvalues. In order to reduce the risk of over-fitting, the optimally configured GMM for each pair of successive joints is determined by using the leave-one-out cross-validation principle, with respect to the number of mixture models as well as the location of zeros for off-diagonal precision matrix elements. In order to train the GMMs, the pose parameters resulting from the initial template model learning phase are used (shown in Figure 3).

The resulting prior term $Q^\text{pose}$ for plausible pose configurations consists of a normalized sum to enforce independently a natural pose for each frame:

$$Q^\text{pose} = \frac{1}{N} \sum_{n=1}^N Q_n^\text{pose} \quad \text{with} \quad Q_n^\text{pose} = 1 - \frac{1}{p} \sum_{p=1}^P \mathcal{P}_{p,n},$$  

(9)

with $Q_n^\text{pose}$ being the objective term penalizing for frame $n$ the deviation of the pose parameters from the $a$ priori learned GMM of likely poses, and $\mathcal{P}_{p,n}$ being the GMM (Equation (8)) for a pair of successive kinematic joints $p$.

This pose prior $Q^\text{pose}$ is influenced solely by the kinematic pose parameters. It does not need to be considered during optimization, if the pose parameters are kept constant.

Objective term $Q^\text{temp}$ to enforce temporal consistency

Real-world data of successfully recorded input frames typically contain noise. Consequently, extracted poses are noisy as well, if the input frames are processed independently and the perception of a smooth flow of movement is distorted. On the other side, smoothing the pose variables independently from the captured sequence can successfully remove the jitter, but at the cost of no longer resembling the true motion of the observed subject.

The objective term $Q^\text{temp}$ is a soft constraint to enforce temporal consistency within the recovered movement across a complete input sequence. Since motion corresponds to the difference of pose parameters in successive frames, undesired changes in motion are reduced by penalizing the difference of the difference of pose parameters in two successive pairs of input frames. This objective term $Q^\text{temp}$ is formulated as a normalized sum over all input frames (the first $n = 1$ and last $n = N$ frame are considered implicitly, see Equation (11)):

$$Q^\text{temp} = \frac{1}{N-2} \sum_{n=2}^{N-1} Q_n^\text{temp},$$  

(10)

with $Q_n^\text{temp}$ being the objective term penalizing the pose parameter difference in frame $n$ with respect to the linear interpolation of its neighbour frames $n-1$ and $n+1$. Using the exponential map representation [Ude98] for the rotations of the $J$ joints, encoded as 3D vector $r = [r_1, r_2, r_3]$, the temporal consistency soft constraint is formulated as:

$$Q_n^\text{temp} = \frac{1}{J} \sum_{j=1}^J \sum_{i=1}^3 \left| r_i^{j,n} - \frac{1}{2} (r_i^{j,n-1} + r_i^{j,n+1}) \right|^2.$$  

(11)

This constraint $Q^\text{temp}$ is influenced solely by the frame-by-frame pose parameters and reduces the likelihood of drastic changes in motion. It is optional in the sense, that it should be discarded from the objective function $Q$, if the input frames contain independently captured poses of a subject, a set of laser scans for example.

Shape prior term $Q^\text{shape}$

The shape of the template model is successively adapted to the captured data in order to approximate the subject and thereby increasing the pose estimation accuracy. We use prior knowledge extracted from the training phase of the initial template to guide this shape adaptation towards plausible human shapes. We calculate the uniform Laplacian vector [Sor06] for each vertex $v$ of each of the $M$ registered training scan meshes, where the pose parameters are known through the template model learning phase (Figure 3):

$$\mathcal{L}(v) = v - \frac{1}{|\mathcal{N}(v)|} \sum_{n \in \mathcal{N}(v)} v_n,$$  

(12)

with $\mathcal{N}(v)$ denoting the set of indices neighbouring vertex $v$ according to the mesh topology. As a result, for each template model vertex, a set of $M$ Laplacian vectors, one for each of the different known pose configurations in the training set, is available.

© 2019 The Authors. Computer Graphics Forum published by Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd
The prior term $Q^{\text{shape}}$ is introduced into the objective function $Q$ to guide the shape adaptation robustly towards realistic shape results. This prior enforces that the local neighbourhood of each template model’s vertex is similar to the corresponding vertex found in the training set. It is formulated as normalized sum over all $M$ registered example scan meshes used for training the initial template model:

$$Q^{\text{shape}} = \frac{1}{M} \sum_{m=1}^{M} Q_{m}^{\text{shape}}.$$  \hspace{1cm} (13)

For each of the training set mesh $m$, the shape prior is formulated as normalized sum over all $V$ vertices:

$$Q_{m}^{\text{shape}} = \frac{1}{V} \sum_{i=1}^{V} \left| \mathcal{L}(v_{i,m}^{\text{train}}) - \mathcal{L}(\mathcal{T}_{m}(v_{i}^{\text{temp}})) \right|^2,$$  \hspace{1cm} (14)

with $v_{i,m}^{\text{train}}$ being the $i^{\text{th}}$ vertex of registered scan $m$ used for learning the template model and $\mathcal{T}_{m}(\cdot)$ being the skinning transformation parameterized to map template vertex $v_{i}^{\text{temp}}$ from rest pose to pose $m$ of the registered scans.

Soft-constraining the proposed objective function $Q$ with this shape prior $Q^{\text{shape}}$ allows to adapt the shape vertices of the template model directly to the captured input frames without suffering from imperfect input data. A suitable value for weight $w^{\text{shape}}$ depends on the nature of the input data. As illustrated in Figure 4, too low values result in noisy shapes, whereas too big values result in loosing the details intended to be learned from the input data.

The shape prior $Q^{\text{shape}}$ is influenced by the rest pose vertices as well as the kinematic skeleton used within the skinning function $\mathcal{T}_{m}(\cdot)$ to transform the rest pose vertices to the registered training scan $m$. It does not need to be considered during optimization, if only pose parameters are optimized.

**Symmetry term $Q^{\text{sym}}$ within the kinematic skeleton** Since the ability of the template model to adopt certain poses depends on the proportion of the embedded kinematic skeleton, the location of the joints is adapted towards the captured data as well. While optimizing the locations of the joints, symmetry within the kinematic skeleton is enforced using a heavily weighted soft-constraint. The prior term $Q^{\text{sym}}$ to penalize differences in length of corresponding bones on the left- and right-hand-side of the kinematic skeleton is formulated for the set $J_{s}$ of pairs of joint indices of bones intended to be of equal length as follows:

$$Q^{\text{sym}} = \sum_{l,r \in J_{s}} \left| (t_{l} - t_{\text{sucl}[l]}) - R_{l,r}(t_{r} - t_{\text{sucl}[r]}) \right|^2,$$  \hspace{1cm} (15)

with $t_{l}$ and $t_{r}$ being the location of corresponding joints on the left and right of the skeleton control structure and $\text{sucl}[-]$ the successor function to yield the index of the next joint within the kinematic chain. The $3 \times 3$ rotation matrix $R_{l,r}$ is used to bring the orientation of both ‘bones’ into alignment, so that their length is directly comparable. This $R_{l,r}$ matrix is calculated from the ‘bones’ $t_{l} - t_{\text{sucl}[l]}$ and $t_{r} - t_{\text{sucl}[r]}$ using the dot- and cross-product.

### 3.2.3. Optimization of objective $Q$

The proposed model adaptive pose estimation is treated as an optimization, which is based on minimizing the objective $Q$. Because of unknown shape and pose parameters, an accurate registration of the template model with the input scans is difficult, particularly because of the high DoF the pose variables and model parameters impose. If solved naively, the likelihood is high to end up in a local minimum of the objective function.

Our optimization can be used either for a sequence of frames representing a continuous movement, or alternatively, for a set of input frames representing independent poses. In the latter case, a rough pose initialization of all input frames is required, whereas in the former case such an initialization is required only for the first frame. Such a pose initialization can be achieved manually, through pre-defined poses (A-pose, T-pose, etc.), or automatic methods, like shape similarity trees [BHH11].

**First step: optimization of pose parameters** The first step of the proposed method is to optimize for the pose parameters of all input frames using the initial non-adapted template model, without any temporal consistency enforcement ($w^{\text{temp}} = 0$). Since neither the skeleton control structure nor the template vertices are modified in this step, no shape prior or symmetry constraint is used ($w^{\text{shape}} = 0$ and $w^{\text{sym}} = 0$). Consequently, the objective function reduces to $Q = Q^{\text{data}} + w^{\text{pose}} Q^{\text{pose}}$.

In the case of independent poses, already roughly pose initialized, the pose parameters are independently optimized in parallel for each
input frame. For sequences of continuous movements, one frame after another is optimized, using the result of the already optimized predecessor frame as initialization of the next pose.

The specific variables to be optimized during pose fitting for each of the input frames are the global translation (3DoF) and rotation (3DoF) as well as the rotation parameters for each kinematic joint ($J \times 3\text{DoF}$). This results in a highly non-linear equation system, which is minimized using a combination of Levenberg–Marquardt and line search. The Levenberg–Marquardt scheme is used to determine the optimal direction of the optimization update step and the line search algorithm is used to determine the optimal step width.

Additionally, for robustness purposes, the optimization is restricted to a reduced solution space of poses. A hard min/max box constraint is imposed independently for each DoF of the joint rotation variables. These lower/upper bounds correspond to the min/max values present in the training set of the initial template model (Figure 3).

Using all pose-adapted frames simultaneously as input (or alternatively a representative subset of them), the subject-specific uniform scale parameter (1DoF) can be jointly optimized together with the pose parameters. Solving this optimization including scale is accelerated by exploiting the block structure found in the Jacobian matrix due to the independence of the pose parameters among the different input frames.

Second step: adaptation of template model jointly together with pose parameters Our model adaptation optimizes the model parameters, specifically skeletal joints ($J \times 3\text{DoF}$) and shape vertices ($V \times 3\text{DoF}$), in conjunction with refining the pose variables mentioned above in one common joint optimization for all input frames (or alternatively a representative subset). Besides optimizing the objective terms $Q_{\text{data}}$ and $Q_{\text{pose}}$, also the shape prior $Q_{\text{shape}}$ and skeletal symmetry constraint $Q_{\text{sym}}$ are employed at this step to keep the skeleton control structure symmetric as well as the shape plausibly human.

Inspecting the complete pose aligning transformation $T()$ (including kinematics-based skinning as well as global similarity, Equation (1)), it can be seen, that an animated template vertex $v'$ linearly depends on its rest-pose vertex $v$. Thus, the transformation can be rewritten in matrix form:

$$v' = T(v) = T_F v + t_F,$$

(16)

with $T_F$ being a $3 \times 3$-matrix and $t_F$ a 3-vector. Note, that as a consequence of LBS, $T_F$ is not a rotation. Both quantities, $T_F$ and $t_F$, are specific for each vertex because they are influenced by their corresponding skinning weights.

Employing a first-order Taylor approximation scheme to discard higher order terms, an equation system is set up resulting in a huge sparse matrix. Following the Gauss–Newton optimization scheme, this equation system is solved efficiently in an (iteratively reweighted) least squares sense using a sparse solver, such as the sparse PARDISO solver from the Intel MKL library.

Optional third step: enforcement of temporal consistency Jittering motions are avoided by enforcing temporal consistency in an additional optimization pass with coupled pose variables of successive frames using the objective term $Q_{\text{temp}}$. The resulting equation system for a whole sequence requires the complete input data to be processed at once and kept in memory. This high memory requirement is reduced, by employing an overlapping sliding window algorithm: Process the first $n$ frames (e.g. $n = 10$) from index 1 to $n$, then process the frames from index $\frac{n}{2}$ to $\frac{n}{2} + n$ with keeping the pose parameters of frame $\frac{n}{2}$ constant, then $n$ to $2n$ and so forth. Because the pose parameters of the first frame in the sliding window are already optimized by the previous sliding window iteration and kept constant for the current one, smooth continuous movements along the complete sequence are enforced, while still correctly following the observed movements of the captured subject.

Practical considerations Typically, an insufficiently good adapted template model is not well suited to resemble a captured input sequence, especially for complex motions and/or a subject too different to the initial template model. Mismatching proportions of limbs as well as non-conforming shape lead to tracking loss. In order to increase the tracking robustness, the model parameters can be adapted multiple times, e.g. for the first pose optimized input frame, after 10, 50, 100, etc. frames again.

4. Evaluation, Discussion and Applications

The proposed model adaptive pose estimation supports tracking of the kinematic pose of a moving human from noisy, partial 3D observations while adapting the template model’s skeleton joints and shape vertices to better resemble the captured subject. This optimization is guided towards realism with integrated prior knowledge, which is controlled through prior weights. As demonstrated experimentally in Section 4.1, this approach is capable to track motion and learn a realistic articulated model from a 3D reconstructed multi-view sequence or a set of different poses. The output is a model, adapted to the captured subject with respect to shape vertices and kinematic skeleton joints, as well as animation parameters to resemble the observed movements. The proposed approach is focused to model kinematically caused motions. Handling of other kinds of non-rigid deformations (e.g. facial expressions and loose cloth) can be integrated, as we will discuss in Section 4.2. The separation of animation parameters and 3D shape perfectly renders the model suited for various manual/automatic manipulations. Further, exploiting the vertex correspondences, this approach can be used to directly animate captured 3D scans.

4.1. Experimental evaluation

The proposed optimization provides pose estimation and tracking while simultaneously adapting the template tracking model towards the observed subject. The model used for all experimental evaluations is shown in Figure 3. It has been learned from the SCAPE dataset [ASK*05] using the algorithms described in Section 3.1.

Several investigations have been carried out to experimentally evaluate the performance of the complete algorithmic framework as well as of its two main components: model adaptation and pose estimation. All results are shown in the accompanied video (Video S1) to provide a better impression of the performance.
To judge the accuracy of the proposed algorithms, the Euclidean distance between fitted template model and input data is analysed. The distance between every input 3D scan vertex and its closest triangle of the fitted template model is computed. The widely used mean Euclidean distance and standard deviation are calculated. Since the input data are imperfect 3D measurements with noise as well as severe outliers, these non-robust statistical measures are negatively impacted and give a skewed impression. Consequently, the robust, outlier tolerant statistical measure median (MED) is computed as well. In order to robustly quantify the dispersion, we use the median absolute deviation (MAD) and interquartile range (IQR), which are commonly used in robust statistics [IH93, UC97].

4.1.1. Model adaptation

The capability to adapt the template model with respect to shape vertices and skeletal joints towards captured 3D data of other individuals is validated using the publicly available dataset MPI FAUST [BRLB14]. This dataset consists of non-watertight full-body 3D scans of 10 subjects, five males and five females, in an age range between 18 and 70 with strong variation in physical properties (height, weight and fitness). Captured meshes of 20 different static pre-defined poses are provided for each subject, resulting in a dataset of 200 3D reconstructions in total.

Initially, the template model is manually brought into coarse kinematic alignment. Since each subject adopted the same set of poses, this task is done for the first subject only, providing a suitable initial pose configuration for all subjects. The pose parameters for each input frame are optimized independently for each frame using the initial non-adapted template model. Using the resulting optimized pose parameters, the template model is adapted with respect to shape vertices and skeletal joints, jointly together with refining the already optimized pose parameters. The prior weights are empirically set to \( w_{\text{pose}} = 0.001, w_{\text{shape}} = 1000 \) and \( w_{\text{temp}} = 0 \).

As shown qualitatively in Figure 5, the adapted models resemble the geometry of the captured subjects very well in all cases. This wide range of spectrum, men as well as women and skinny as well as corpulent physical properties, is recovered in detail from the input data, leading to an appropriately adapted template model. The fists of the initial template model are transformed and approximate closely the finger postures of the different input scans.
Table 1: Quantitative analysis of model adaptation capability on MPI FAUST dataset, given in mm-distances between 3D scan vertices and closest template model triangle.

| id | Mean  | std dev | MED   | MAD   | IQR   |
|----|-------|---------|-------|-------|-------|
| 0  | 2.96/6.45 | 3.95/6.29 | 1.76/4.39 | 1.23/3.28 | 2.98/7.91 |
| 1  | 2.73/6.27 | 3.28/7.16 | 1.74/3.53 | 1.18/2.60 | 2.74/7.03 |
| 2  | 3.31/7.22 | 4.38/8.47 | 1.98/4.37 | 1.38/2.93 | 3.27/6.88 |
| 3  | 3.36/7.34 | 4.43/8.02 | 2.03/4.75 | 1.40/3.46 | 3.30/8.18 |
| 4  | 3.01/6.20 | 3.60/6.81 | 1.92/3.81 | 1.32/2.66 | 3.09/6.35 |
| 5  | 2.50/5.00 | 3.10/5.22 | 1.56/3.31 | 1.08/2.35 | 2.53/5.60 |
| 6  | 2.52/5.37 | 3.12/5.86 | 1.61/3.40 | 1.08/2.33 | 2.47/5.54 |
| 7  | 3.63/7.47 | 4.13/7.73 | 2.39/4.68 | 1.62/3.30 | 3.72/8.48 |
| 8  | 2.96/7.22 | 3.89/8.35 | 1.79/4.12 | 1.24/3.11 | 2.92/8.04 |
| 9  | 3.16/6.24 | 3.84/6.59 | 1.98/4.12 | 1.36/2.94 | 3.24/6.91 |
| avg | 3.01/6.48 | 3.77/7.05 | 1.88/4.05 | 1.29/2.90 | 3.03/7.09 |

The left value in each column denotes the measurement across the whole sequence, on the right the maximum value is provided.

A quantitative evaluation is summarized in Table 1. Despite the 3D scan artefacts (noise, missing data and outliers) shown in Figure 5, the broad range of physical properties present in this dataset is well approximated. On average, the median Euclidean distance of input scan vertices to the closest template model triangle is well below 2 mm with a median absolute deviation of 1.3 mm.

4.1.2. Pose estimation and tracking

The pose estimation and tracking accuracy of our optimization is evaluated using the publicly available dataset MPI BUFF [ZPBPM17]. This dataset, captured with 30 FPS, consists of unregistered full-body 3D reconstruction sequences of five different subjects (three males and two females). Each subject performed the same set of three different movements with two different types of cloth, resulting in six sequences per subject. In total, the complete dataset of all subjects contains ≈10 k non-watertight meshes with an average per-scan size of ≈125 k vertices and ≈250 k triangles.

A crude pose alignment is manually generated for the first frame of the first sequence. This pose is used as tracking initialization for all sequences, because all start with a nearly identical pose. The proposed optimization processes each input sequence according to Section 3.2.3. In order to keep the computational load and memory requirement at a moderate level, every 25th frame is used for the model adaptation, resulting on average in using 43 frames. The temporal consistent pose estimation uses the overlapping sliding window mechanism with a window size of 11 frames and 5 frames increments. The prior weights are empirically set to $w_{\text{pose}} = 5$, $w_{\text{temp}} = 10^3$, $w_{\text{shape}} = 10^5$ and $w_{\text{sym}} = 10^6$. The intermediate results of the different processing steps are shown in Figure 6.
The described processing scheme successfully recovers the pose parameters of all frames of all MPI BUFF subject’s sequences, without any divergence or tracking loss.

A qualitative result is shown in Figure 7 (also Figure 8) with input/output comparison for every subject’s frame. It can be clearly seen, that even difficult motions are well recovered, e.g. rotated upper body part. The hands, represented in the template model as closed fists and in the dataset as forced apart fingers, are well approximated without any orientation error. Individual geometric peculiarities are recognizably reproduced, e.g. facial appearance, hair cut and to a certain degree the loose cloth.

Quantitatively, as summarized in Table 2, the median Euclidean distance of input frame vertices to the closest template model triangle is on average well below 2 mm with a median absolute deviation from this value of 1.4 mm. With this result, the proposed model adaptive motion capture qualifies as being as accurate as current state of the art approaches. The method presented in [ZPBPM17] evaluated an average mean accuracy of 2.5 mm on the same MPI BUFF dataset, whereas the proposed method achieves 3.3 mm. Qualitatively, [ZPBPM17] is able to better recover the details of the hand and face, but does not extract details of cloth, like folds and wrinkles of the trousers and shirt, as the proposed method does. The method presented in [AMX+18b] evaluated an average mean accuracy of 5.37 mm on this dataset. But it needs to be stated, that this comparison is informative only, because they solve a different problem: model adaptive motion capture using a single monocular 2D RGB sequence of a moving person as input (generated artificially for this dataset via projection onto the 2D plane).

4.1.3. Influence of model adaptation on pose estimation

The influence of model adaptation on pose estimation accuracy is experimentally evaluated using again the MPI BUFF dataset. Therefore, the interim results of the different processing steps presented in the previous section are analysed. Additionally, each sequence has been processed both with and without model adaptation, but without temporal consistency enforcement.

A qualitative result is shown in Figure 8. The non-adapted template model is not able to accurately adopt the pose of the input frame. Only a very rough approximation is achieved, because the kinematic bone structure and surface properties of the template model do not match the subject’s characteristics. The single frame adaptation of the template model to the first frame of the input sequence roughly individualizes the model, but already sufficiently to allow accurate tracking of the complete sequence. The multi-frame adaptation of the template model finally resembles the captured subject in finer detail, e.g. the face becomes recognizable.
Figure 8: Influence of model adaptation on tracking accuracy with MPI BUFF subject 00032: 3D scan data (red), template model (cyan) and heat map visualizing the point-to-plane distance between MPI BUFF vertices and corresponding closest point on model surface; from left to right: no model adaptation at all, model adapted to first frame only, model adapted to complete sequence.

The quantitative analysis, summarized in Table 3, measures the distance between the vertices of input 3D scans and the closest triangle of the adapted template model. The median distance for tracking with a non-adapted template model is on average 6.5 mm. After adapting to the first frame, the median tracking accuracy already reduces to 2.8 mm. Finally, adapting the model to the complete sequence results in a median tracking accuracy of less than 2 mm. Similarly, the median absolute deviation from the median distance reduces from 4.4 mm for the non-adapted template model, over 1.9 mm for the first-frame-adapted template model, to 1.4 mm for the fully adapted template model.

Temporal consistency The enforcement of temporal consistency using a weight $w_{temp} > 0$ reduces the jittery effect to a non-visible minimum. The resulting sequence of poses is smooth, but still follows accurately the captured subject, thereby enhancing the qualitative perception. Quantitatively, the statistics of enforcement and non-enforcement of temporal consistency are nearly identical, indicating that there is no loss in accuracy.

4.1.4. Noisy, incomplete input data

Our method provides pose estimation with simultaneous model adaptation on incomplete 3D reconstructions, where not every direction has been captured (missing views). Such footage results, for example, from hardware setups with cameras on only one frontal plane, or if the cameras are placed along a half circle. To experimentally evaluate the achievable quality and accuracy, three different hardware setups with increasing hardware performance have been used to capture three different subjects (one male and two females). Two of the datasets have been made publicly available [Fec17].

In all captured sequences, the subjects wore everyday, relatively tight fitting cloth. The subjects were instructed to perform movements, where every major limb was moved at least once (in order to properly extract the characteristics of the kinematic joints) and to turn around $360^\circ$ during this movement (in order to capture every surface patch at least once).

The characteristics of the three multi-view hardware setups are summarized in Table 4. All sequences were recorded with 25 FPS (hardware synchronized). The $1920 \times 1080$ resolution cameras

| Cloth             | id   | Mean  | std dev | MED   | MAD  | IQR  |
|-------------------|------|-------|---------|-------|------|------|
| T-shirt, long pants | 0005 | 4.17/6.50 | 9.22/34.72 | 2.61/4.10 | 1.82/2.76 | 4.29/6.28 |
|                   | 0032 | 3.73/6.23 | 5.10/17.76 | 2.11/4.06 | 1.51/2.96 | 3.78/7.07 |
|                   | 0096 | 4.07/6.19 | 6.45/31.65 | 2.42/3.57 | 1.69/2.55 | 4.09/6.13 |
|                   | 0114 | 2.76/4.64 | 4.38/14.87 | 1.60/3.13 | 1.11/2.09 | 2.71/4.65 |
|                   | 3223 | 2.45/5.30 | 6.78/31.51 | 1.47/2.79 | 1.02/1.91 | 2.42/4.50 |
| Soccer outfit     | 0005 | 3.85/6.31 | 8.27/33.55 | 2.29/3.52 | 1.59/2.32 | 3.81/5.24 |
|                   | 0032 | 3.55/5.39 | 5.36/9.01  | 1.83/3.72 | 1.38/2.72 | 3.67/6.30 |
|                   | 0114 | 2.58/4.76 | 3.72/9.53  | 1.50/3.20 | 1.07/2.19 | 2.67/5.10 |
|                   | 3223 | 2.63/4.81 | 3.80/17.15 | 1.44/2.92 | 1.05/2.05 | 2.71/4.84 |
| avg               | 3.31/5.57 | 5.90/22.19 | 1.92/3.45  | 1.36/2.39 | 3.35/5.57 |

The left value in each column denotes the measurement across the whole sequence, on the right the maximum value is provided.
were Basler ace acA2000-50gc and the 5120×3840 resolution cameras were Ximea CB200CG-CM. In all capture scenarios, the cameras were placed in pairs to allow stereo 3D reconstruction algorithms to be used, which are merged to one 3D mesh per captured frame [WFS*16].

As can be seen in Figure 9 of the qualitative results, the three captured sequences are of different quality. The resulting animations of the template model of all three sequences are accurately aligned in pose and shape with the captured 3D reconstructions. Complex motions, like turning around, are recovered without difficulty, which, for example, is not possible with the Kinect device. The model adaptations closely resemble the subject including detailed individual peculiarities, e.g. a ponytail hair style.

The quantitative measurements are summarized in Table 4. The achieved accuracy varies with the quality of the input data: a high noise level within the input data produces higher measurement errors (median: >4.3 mm, MAD: >2.7 mm), whereas a low level of noise results in higher accuracy (median: <2.2 mm, MAD: <1.5 mm).

### 4.2. Validation and discussion

The specific run time of the proposed method depends on the number of input meshes and vertices. On an Intel E5-2687W CPU with 3.40 GHz, one iteration of pose parameter optimization for an MPI FAUST dataset mesh takes less than 300 ms. One iteration of model adaptation for one MPI FAUST subject (20 example meshes) takes ≈100 s. The complete computation of one MPI FAUST adapted template model including pose refinement requires up to 6 h. Processing the MPI BUFF sequences per subject requires ≈8 h for one sequence (≈400 frames), to calculate the adapted model together with temporally consistent pose parameters.

The implemented algorithms are only moderately optimized with strong potential for further enhancements. For instance, computational load can be reduced by using a hierarchical approach similar to [dAST*08, BBLR15]: a registered low-poly and high-poly model of the template model is employed, using the low-poly version for the majority of computations and the corresponding high-poly model only for the final high-detail steps. Similarly, in cases of high-poly input 3D scans, a mesh simplification [KH13] or sampling approach [YSD*16] might be beneficial. Last but not least, inherent parallelism could be exploited to a much stronger degree on CPU as well as GPU and thereby again reducing computational time. Using such acceleration mechanisms, real-time tracking could be achieved by shifting the model adaptation to a prior offline process. Alternatively, a progressive model adaptation scheme could be realized by employing the recent Levenberg–Marquardt Kalman filter [SHA15, TTR*17].

The proposed optimization approach focuses on modelling and handling surface deformations evoked by kinematic dynamics. It does not model pose-dependent shape deformations (e.g. muscle bulging) or facial expressions. Such capabilities could be integrated using blend shape approaches in a similar way as [LMR*15, AMX*18b]. Moreover, small-scale deformations, as required for finger motions, are out of scope of this work. Principally, it is straightforward to directly extend the template model with kinematically animatable fingers. But this significantly increases the demand for high-resolution 3D scanning [JSYS18] or alternatively requires other means to robustly extract the hand poses from captured input data, e.g. [RTB17]. Explicit handling of complex non-rigid non-kinematic deformations, e.g. from loose cloth or hair, has not been considered in this work. Identification of cloth vertices, automatic estimation of cloth properties as well as real-time animation of cloth deformations could be integrated straightforward by blending...
Figure 9: Evaluation of proposed method on multi-view datasets, from top to bottom: atvp, hbrl, c4am; from left to right: one RGB-camera input view, 3D reconstruction (red), adapted template model (cyan) and the heat map visualizing the point-to-plane distance between captured vertices and corresponding closest points on model surface.
in a vertex-specific manner between the non-rigid cloth deformation model and the kinematic body deformation model [ZPBPM17, MHHR07, SGdA*10].

However, despite the restrictions on kinematically evoked deformations, the proposed optimization experimentally proved to perform well. Further, its versatile applicability is demonstrated in the next section with relevant example applications.

4.3. Applications

The output of our framework is a model adapted in shape and skeleton to a captured subject as well as animation parameters to resemble the observed movements. This footage is beneficial for several kinds of applications.

Since the model animates the whole surface instead of only a partial view, it can be used for free viewpoint video. The separation of animation parameters and shape renders it well suited for manipulations. Combining sequences of animation parameters is straightforward. Since animation parameters correspond to joint angles, interpolations between configurations provide valid interim poses.

Textured animations can be achieved in real-time [FPHE16, FPE14]. The model is animated with Kinect using retargeting to map the skeleton parameters. A static full-body texture is extracted from the multi-view sequence of the subject-adapted model generation. Realistic facial expressions are provided by dynamically updating the textures facial area from camera input.

4.3.1. Manipulation and editing of motion

Since the animation model is based on skinning using a kinematic skeleton, they are directly accessible and can be manipulated artistically by hand as well as in an automatic manner. Generating new/edited animations is compliant to traditional character modelling, thereby making it amenable to human modelling artists/animators. Importing it into standard human modelling software can be achieved using methods like [KCO09]. Automatic adaptations can be achieved by exploiting the knowledge of the underlying mesh topology of the original template model, providing semantic annotations.
An example is to automatically control the gaze direction in real-time. A virtual line going through the area between the eyes and the back of the head is used to infer the required animation parameters to let the model look into a certain direction. To provide a natural appearance of the head movement, the animation parameters are blended, temporally and spatially: Temporally — the magnitude of gaze direction adaptation is faded in and out using sigmoid functions to start and stop seamlessly the control over the viewing direction. Spatially — the gaze correction is distributed among the breasts joint (10%), neck joint (40%) and head joint (50%). As shown in Figure 10, this automatic gaze direction control seamlessly adapts the viewing direction without disturbing the natural appearance.

4.3.2. Direct animation of 3D reconstructed input data

To further increase the realism, 3D reconstructed input data can be directly animated by using the model only internally without visualizing it. This is achieved by starting from a frame, where shape and pose is already fitted to a captured subject. For each of the scan vertices, the location relative to the closest triangle of the model is calculated and held fixed, virtually gluing the scan vertex to the template triangle with constant distance and orientation. Therefore, the scan vertex is projected onto the plane of the closest triangle and the barycentric coordinates are calculated as well as the orthogonal distance vector between the scan vertex and the triangles plane. Modifying an animation parameter alters the location of model vertices. The updated model triangles are then used to calculate new positions of the scan vertices, thereby animating the scan itself.

In Figure 11, a full-body scan of a human is shown, consisting of 227 k vertices and 453 k triangles, 3D reconstructed using [NLM*17]. Using the pose and shape optimized model to deform a captured human produces realistic animations of 3D scans. Since no model is visualized, but the original captured data, no loss of detail happens, thereby reproducing as accurately as possible every captured detail.

This method can also be applied for the animation of dynamic body scans (e.g. created by [EFR*17]) in order to allow for pose modification while keeping the original details of the scans.

5. Conclusions

In this paper, we have presented a novel markerless model adaptive motion capture method. A rigged shape model, learned a priori from real-world data, is used as kinematic tracking model. In order to properly resemble an input 3D mesh sequence, the template model is adapted to the observed input data with respect to shape vertices and skeleton joint positions. Prior knowledge extracted from the training set of the initial template model is used to guide the optimization towards plausible poses and shape: (1) A data-efficient pose prior based on GMMs defined on every pair of successive joints in the kinematic chain and (2) a shape prior based on mesh Laplacians defined on the vertex neighbourhood structure of the template model’s triangle topology. Since a semantic mapping of a template to the input 3D sequence is achieved, it is straightforward to manipulate the resulting animation sequence, e.g. looping over sub-sequences, exchanging the model, adapting movements, etc. Due to the underlying skinning-based animation style, manipulations are easy accessible to human modelling artists/animators.

The high accuracy of pose and model fitting of this approach has been experimentally demonstrated with publicly available datasets and captured multi-view sequences.

Acknowledgements

This research has received funding from the EUs Horizon 2020 research and innovation program under grant agreement numbers 687757 (REPLICATE) and 762021 (CONTENT4ALL).

References

[AMX*18a] Allieck T., Magnor M., Xu W., Theobalt C., Pons-Moll G.: Detailed human avatars from monocular video. In International Conference on 3D Vision (September 2018).

[AMX*18b] Allieck T., Magnor M., Xu W., Theobalt C., Pons-Moll G.: Video based reconstruction of 3D people models. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Salt Lake City, Utah, USA, June 2018). CVPR Spotlight Paper.

[ASK*05] Anguelov D., Srinivasan P., Koller D., Thrun S., Rodgers J., Davis J.: SCAPE: Shape completion and animation of people. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH’05) 24 (August 2005).

[BBLR15] Bogo F., Black M. J., Loper M., Romero J.: Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015).

[BHH11] Budd C., Huang P., Hilton A.: Hierarchical shape matching for temporally consistent 3D video. In Proceedings of Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT) (May 2011).

[BKL*16] Bogo F., Kanazawa A., Lassner C., Geheimer P., Romero J., Black M. J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In Computer Vision – ECCV 2016 (October 2016), Lecture Notes in Computer Science, Springer International Publishing.

[Bla16] Black M.: The future and generative models: A case study of human bodies in motion. International Computer Vision Summer School, July 2016.

[BRLB14] Bogo F., Romero J., Loper M., Black M. J.: FAUST: Dataset and evaluation for 3D mesh registration. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Columbus, Ohio, USA, June 2014).

[BS10] Bo L., Smichiesescu C.: Twin Gaussian processes for structured prediction. International Journal of Computer Vision 87, 1-2 (2010), 28–52.

[CCM*18] Cheng Z.-Q., Chen Y., Martin R. R., Wu T., Song Z.: Parametric modeling of 3D human body shape – A survey. Computers & Graphics 71 (2018), 88–100.

© 2019 The Authors. Computer Graphics Forum published by Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd
[LD14] Le B. H., Deng Z.: Robust and accurate skeletal rigging from mesh sequences. *Proceedings of ACM SIGGRAPH* (2014).

[LMR*15] Loper M., Mahmood N., Romero J., Pons-Moll G., Black M. J.: SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics* 34, 6 (October 2015), 248:1–248:16.

[LSST09] Lee S.-H., Sifakis E., Terzopoulos D.: Comprehensive biomechanical modeling and simulation of the upper body. *ACM Transactions on Graphics* 28 (August 2009), 99:1–99:17.

[MG03] Mohr A., Gleicher M.: Building efficient, accurate character skins from examples. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)* 22 (July 2003), 562–568.

[MHHR07] Müller M., Heidelberger B., Hess M., Ratcliffe J.: Position-based dynamics. *Journal of Visual Communication and Image Representation* 18, 2 (April 2007), 109–118.

[MSS*17] Mehta D., Sridhar S., Sonynchenko O., Rhodin H., Shahri M., Seidel H.-P., Xu W., Casas D., Theobalt C.: VNect: Real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics* 36, 4 (July 2017), 44:1–44:14.

[NLM*17] Nocerino E., Lago F., Morabito D., Remondino F., Porzi L., Foi F., Rota Bulo S., Cippendale P., Locher A., Havlena M., Van Goor L., Eder M., Fütschel A., Hilsmann A., Kaush L., Eisert P.: A smartphone-based 3D pipeline for the creative industry - The REPLICATE EU project. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W3 (2017), 535–541.

[PB12] Piyan G., Bullo F.: On coordinate-free rotation decomposition: Euler angles about arbitrary axes. *IEEE Transactions on Robotics* 28, 3 (2012), 728–733.

[PI*T*16] Pishchulin L., Insafutdinov E., Tang S., Andres B., Andriluka M., Geheimer P., Schiele B.: DeepCut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

[PMRMB15] Pons-Moll G., Romero J., Mahmood N., Black M. J.: Dyna: A model of dynamic human shape in motion. *Proceedings of ACM SIGGRAPH* (July 2015).

[PWH*17] Pishchulin L., Wuerger S., Helten T., Theobalt C., Schiele B.: Building statistical shape spaces for 3D human modeling. *Pattern Recognition* 67 (2017), 276–286.

[RLGK16] Rafi U., Leibe B., Gall J., Kostrikov I.: An efficient convolutional network for human pose estimation. In *Proceedings of the British Machine Vision Conference (BMVC)* (September 2016), R. C. Wilson E. R. H., Smith W. A. P. (Eds.). BMVA Press, pp. 109.1–109.11.

[RTB17] Romero J., Tzionas D., Black M. J.: Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 36, 6 (November 2017), 245:1–245:17.

[RW06] Rasmussen C. E., Williams C. K. I.: Gaussian Processes for Machine Learning. MIT Press, 2006.

[SGdA*10] Stoll C., Gall J., Aguilar E., Thrun S., Theobalt C.: Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics* 29, 6 (December 2010), 139:1–139:10.

[SHA15] Skoglund M., Hendey G., Axehill D.: Extended Kalman filter modifications based on an optimization viewpoint. In *Proceedings of International Conference on Information Fusion* (July 2015).

[Sor06] Sorkine O.: Differential representations for mesh processing. *Computer Graphics Forum* 25 (December 2006), 789–807.

[SRB10] Sun D., Roth S., Black M. J.: Secrets of optical flow estimation and their principles. In *Proceedings of CVPR’10* (San Francisco, CA, USA, June 2010).

[TR*T*17] Tkach A., Tagliasacchi A., Remelli E., Pauly M., Fitzgibbon A.: Online generative model personalization for hand tracking. *ACM Transactions on Graphics* 36, 6 (November 2017), 243:1–243:11.

[UC97] Upton G., Cook I.: *Understanding Statistics*. Oxford University Press, January 1997.

[Ude98] Ude A.: Nonlinear least squares optimisation of unit quaternion functions for pose estimation from corresponding features. In *Proceedings of ICPR’98* (Brisbane, Australia, August 1998).

[vKZHCO11] v. Kaick O., Zhang H., Hamarneh G., Cohen-Or D.: A survey on shape correspondence. *Computer Graphics Forum* 30, 6 (2011), 1681–1707.

[WFS*16] Wazenegg W., Feldmann L., Schreer O., Kauff P., Eisert P.: Real-time 3D body reconstruction for immersive TV. In *Proceedings of ICIP’16* (Phoenix, Arizona, USA, September 2016).

[WLB11] Weiss A., Herschberg D., Black M. J.: Home 3D body scans from noisy image and range data. In *Proceedings of the International Conference on Computer Vision* (Washington, DC, USA, 2011), ICCV’11, IEEE Computer Society, pp. 1951–1958.

[YSD*16] Ye M., Shen Y., Du C., Pan Z., Yang R.: Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (August 2016), 1517–1532.

[YY14] Ye M., Yang R.: Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2014).
[YZG*18] YU T., ZHENG Z., GUO K., ZHAO J., DAI Q., LI H., PONS-MOLL G., LIU Y.: DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Salt Lake City, Utah, USA, June 2018).

[ZB15] ZUFFI S., BLACK M. J.: The stitched puppet: A graphical model of 3D human shape and pose. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015).

[ZFL*10] ZHOU S., FU H., LIU L., COHEN-OR D., HAN X.: Parametric reshaping of human bodies in images. ACM Transactions on Graphics 29, 4 (July 2010), 126:1–126:10.

[ZFYY14] ZHANG Q., FU B., YE M., YANG R.: Quality dynamic human body modeling using a single low-cost depth camera. In The IEEE Conference on Computer Vision and Pattern Recognition (June 2014), CVPR’14.

[Zha95] ZHANG Z.: Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting. Research Report RR-2676, INRIA, October 1995.

[ZPBPM17] ZHANG C., PUJADES S., BLACK M. J., PONS-MOLL G.: Detailed, accurate, human shape estimation from clothed 3D scan sequences. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Piscataway, NJ, USA, July 2017), IEEE, CVPR2017.

[ZYKD16] ZHU H., YU Y., ZHOU Y., DU S.: Dynamic human body modeling using a single RGB camera. MDPI Sensors 16, 3 (2016), 402.

Supporting Information
Additional supporting information may be found online in the Supporting Information section at the end of the article.

Video S1